SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

1222-2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

STATA

sas

UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Contents

Preface

This book includes the contributions presented at the 51st Scientific Meeting of the Italian Statistical Society (SIS) held in Caserta at the Università della Campania "Luigi Vanvitelli", from the 22nd to 24th of June, 2022.

The conference has registered more than 300 presentations, including 4 keynotes in plenary invited sessions and 9 presentations in 3 guest sessions, 48 presentations collected in 16 specialized sessions and 68 presentations in 17 solicited sessions, all dealing with specific themes in methodological and/or applied statistics and demography. Furthermore, more than 200 contributions, with one or more authors, have been spontaneously submitted to the Program Committee and arranged in 43 contributed sessions.

The high number of contributions and the large participation at the conference show that researchers have met the challenge of pursuing working even in the face of the pandemic period from which we are only now emerging. The research activity in our field therefore has never stopped, and the desire to participate in scientific events, as a place for exchange and discussion on new developments in our field, remains a living characteristic of our scientific community.

With the publication of this book, we wish to offer to all members of the Italian Statistical Society, all international academics, researchers, Ph.D. students, and all interested practitioners, a good snapshot of the on-going research in the statistical and demographic fields. We deeply thank all contributors for having submitted their works to the conference and all the researchers for their remarkable job in acting as referees accurately and timely. We also would like to thank the International Biometric Society (IBS) – Italian region, the European Network for Business and Industrial Statistics (ENBIS) and the Italian Society of Statistical Physics (SIFS) we had the pleasure of hosting. A special thanks is addressed to the Scientific and Organizational Committees for their great efforts devoted to all the organizational aspects, to the Università della Campania "Luigi Vanvitelli" and to the Department of Mathematics and Physics who made this event possible, as well as to the Municipality of the Town of Caserta who has patronized the event and to all the funders for their supports.

Finally, we wish to express our gratitude to the publisher Pearson Italia for all the support received.

1 Plenary Sessions

Causal inference in air pollution epidemiology

Francesca Dominici

Causal inference in air pollution epidemiology [1]

Inferenza causale negli studi di epidemiologia ambientale

Francesca Dominici

Abstract Many studies link long-term fine particle ($PM_{2.5}$) exposure to mortality, even at levels below current U.S. air quality standards (12 micrograms per cubic meter). These findings have been disputed with claims that the use of traditional statistical approaches does not guarantee causality. In this paper we review five statistical methods for estimating causal link between exposure to air pollution and health outcomes. Leveraging 16 years of data—68.5 million Medicare enrollees—we provide strong evidence of the causal link between long-term $PM_{2.5}$ exposure and mortality under a set of causal inference assumptions. We found that a decrease in $PM_{2.5}$ (by 10 micrograms per cubic meter) leads to a statistically significant 6 to 7% decrease in mortality risk. Our study provides the most comprehensive evidence to date of the link between long-term $PM_{2.5}$ exposure and mortality, even at levels below current standards.

Abstract in Italian

Molti studi collegano l'esposizione a lungo termine alle particelle fini ($PM_{2.5}$) alla mortalità, anche a livelli inferiori agli attuali standard di qualità dell'aria statunitensi (12 microgrammi per metro cubo). Questi risultati sono stati contestati con affermazioni secondo cui l'uso di approcci statistici tradizionali non garantisce la causalità. In questo articolo esaminiamo cinque metodi statistici per stimare il nesso causale tra l'esposizione all'inquinamento atmosferico e gli esiti sulla salute. Sfruttando 16 anni di dati, 68,5 milioni di iscritti a Medicare, forniamo prove evidenti del nesso causale tra l'esposizione a lungo termine al $PM_{2.5}$ e la mortalità in base a una serie di ipotesi di inferenza causale. Abbiamo scoperto che una diminuzione del $PM_{2.5}$ (di 10 microgrammi per metro cubo) porta a una diminuzione statisticamente significativa dal 6 al 7% del rischio di mortalità. Il nostro studio fornisce la prova più completa fino ad oggi del legame tra l'esposizione a lungo termine al $PM_{2.5}$ e la mortalità, anche a livelli inferiori agli standard attuali.

Key words: causal inference, air pollution, mortality

[1] Results of this brief paper are also reported here and here.

1 Background and Motivation

As air pollution levels continue to decrease and regulatory actions become more costly, the quantification of the public health benefits of cleaner air are subject to an increased level of scrutiny. Epidemiological analyses of claims data have provided strong evidence of air pollution adverse health effects, mostly using data from urban areas. Yet, significant gaps in knowledge persisted, particularly regarding the health effects of long-term exposure to lower levels of air pollution. The estimation of health effects associated with long-term exposure to low levels of air pollution presents key methodological challenges, including the estimation of an exposure response (ER) within a traditional regression framework does not have a causal interpretation and can be highly sensitive to model choice for both the shape of the ER and the adjustment for confounding. In this paper we present an overview of several statistical methods for estimating the ER in air pollution epidemiology. More specifically, we consider five approaches: 1) a survival model (Andersen and Gill 1982) used in Di et al. (Di et al. 2017b); (2) a more computationally efficient Poisson formulation that is equivalent to the Andersen-Gill model under certain assumptions; and (3) three methods for causal inference based on the Generalized Propensity Score (GPS). Two of these methods have been previously published and one is a new method developed by our group (Wu et al. 2018b). We apply these methods to the largest data platform on air pollution and health outcomes assembled to date and estimate the ER for long term exposure to fine particulate matter and all-cause mortality.

1.1 Why do we need causal inference methods in addition to standard regression approaches?

Causal inference methods have advantages and disadvantages compared to traditional regression methods. The strengths are:

- They separate the design stage from the outcome analysis, thus increasing the objectiveness of causal analysis, and mimic a randomized experiment under a set of explicit identification assumptions.
- They guide researchers to state explicitly all the identification assumptions needed for statistical analysis and equip them with a body of sensitivity analysis tools to understand how likely the identification assumptions are held (e.g., covariate balance.).
- They are more robust to model misspecification compared to traditional regression approaches.

But they also have limitations:

- Causal inference methods often require increased computational resources due to the complexity of algorithms.
- Some causal inference methods require steeper learning curves for new researchers due to the logic complexity and are often less familiar to many researchers.
- Methods based on generalized propensity scores are still affected by unmeasured confounding bias.
- Propagation of exposure error in health effects analyses under a causal inference framework are very challenging because error in the exposure also affects the propensity score. See Wu et al. (Wu et al. 2019) for a broad description of the challenges and a proposed solution.

Under a causal inference framework, we articulate our research question using a potential outcome framework, that is, philosophically, we state a hypothetical causal question explicitly by mathematical formulas, for example, "If the pollution level is reduced from 12 units to 10 units, how many premature deaths can be saved?"

Study Population

Table 1 provides the characteristics of the study cohort. Our study population was comprised of more than 68.5 million Medicare enrollees (≥65 years of age) between 2000 and 2016. Medicare claims data, obtained from the Centers for Medicare & Medicaid Services (CMS) (Centers for Medicare and Medicaid Services), is an open cohort, including demographic information such as age, sex, race/ethnicity, date of death, and residential zip code. A unique patient ID was assigned to each person to allow for tracking over time. After enrollment, each subject was followed annually until the year of their death or the end of our study period (31 December 2016).

Table 1. Characteristics for the Study Cohorts

Variables	**Entire Medicare Enrollees**	**Medicare Enrollees Exposed to $PM_{2.5}$ ≤12 μg/m^3**
Number of individuals	68,503,979	38,366,800
Number of deaths	27,106,639	10,124,409
Total person-years	573,370,257	259,469,768
Median years of follow-up	8.0	8.0
Individual-level characteristics		
Age at entry (years)		
65-74 (%)	80.6	88.1
75-84 (%)	14.9	9.0
85-94 (%)	4.1	2.6
95 or above (%)	0.4	0.2
Mean (SD)	69.2 (6.7)	67.6 (5.6)
Sex		
Female (%)	55.5	53.8

Male (%)	44.5	46.2
Race		
White (%)	83.9	84.7
Black (%)	9.1	7.3
Asian (%)	1.8	1.8
Hispanic (%)	2.0	2.2
North American Native (%)	0.3	0.4
Medicaid eligibility		
Eligible (%)	11.7	10.9
Area-level risk factors characteristics		
Ever smoked (%)	47.3	47.3
Below poverty level (%)	10.5	10.1
Below high school education (%)	28.5	25.6
Owner-occupied housing (%)	72.0	72.9
Hispanic (%)	8.9	7.5
Black (%)	8.9	9.2
Population density (persons/km^2)	600.0 (1953.9)	489.1 (1634.0)
Mean BMI (kg/m^2)	27.6 (1.1)	27.6 (1.1)
Median household income (1000 $)	48.9 (21.7)	50.3 (22.0)
Median home value (1000 $)	162.5 (140.9)	170.9 (146.2)
Meteorological variables		
Summer temperature (°C)	29.5 (3.7)	29.5 (3.9)
Winter temperature (°C)	7.6 (7.2)	7.4 (7.6)
Summer relative humidity (%)	88.0 (11.7)	86.7 (12.7)
Winter relative humidity (%)	86.2 (7.3)	86.4 (7.6)
$PM_{2.5}$ concentrations (μg/m^3)	9.8 (3.2)	8.4 (2.3)

Note: Mean (SD) is presented for continuous variables. BMI, body mass index.

2. Statistical Analysis

In this section, we provide mathematical details on our statistical analyses. The R code for the implementation of all five statistical approaches is published and available at https://github.com/NSAPH/National-Casual-Analysis. We implemented five statistical approaches to estimate the effect of $PM_{2.5}$ exposure on mortality, accounting for potential confounders 1) Cox Proportional Hazard Approach; 2) Poisson Regression Approach; 3) GPS matching; 4) GPS weighting; and 5) GPS adjustment.

GPS Estimation: The three proposed causal inference approaches required the estimation of GPS as the first step. In our study, we modeled the conditional density of exposure (i.e., zip code-level annual average $PM_{2.5}$) on the 14 zip code- or county-level time-varying covariates, as well as a dummy region variable and dummy calendar year variable, by using gradient boosting machine with normal residuals (Chen and Guestrin 2016). The gradient boosting machine model is

specified as: $PM_{2.5}$~ area-level risk factors + meteorological variables + dummy year + dummy region + ε, where ε ~N(0, σ^2).

GPS Matching Approach: The GPS matching approach is an approach newly developed by our team, and is described in detail in (Wu et al. 2018b) (under review for publication). The ultimate objective for matching is to construct matched datasets that approximate a randomized experiment as closely as possible by achieving good covariate balance. In the continuous exposure setting, the challenge is that it is unlikely that two units will have the exact same level of exposure; thus, it is infeasible to create a finite sample representing a quasi-experimental arm with the same exposure level by solely matching on GPS. Therefore, we proposed a nearest neighbor caliper-matching procedure with replacement, which jointly matches on both the estimated GPS and exposure values. The closeness of exposure level guarantees that the matched unit is a valid representation of observations for a particular exposure level, whereas the closeness of GPS ensures that we are properly adjusting for confounding. Importantly we assessed covariate balance in the matched population, and if covariate balance was achieved, we fit a univariate Poisson regression model specified as: log(E[death counts]) ~ $PM_{2.5}$+ strata(age, race, gender, Medicaid eligibility, follow-up year) + offset(log[person year]), on the matched pseudo-population.

GPS Weighting Approach: Following Robins et al.(Robins et al. 2000), the weighting approach involves using the inverse of the GPS to weigh the observations.

GPS Adjustment Approach: Following Hirano and Imbens (Hirano and Imbens 2004), a covariate adjustment approach includes the estimated GPS as a covariate in the outcome model.

3. Results

The causal inference framework lends itself to the evaluation of covariate balance for measured confounders. The covariate balance indicates the quality of the causal inference approach at recovering randomized experiments and informs the degree to which we can make a valid causal assessment. Covariate balance was evaluated using mean AC, with values <0.1 indicating high quality in recovering randomized experiments. Results are summarized in Figures 1 and 2.

Figure 1 summarizes the effect estimates for the period 2000–2016. The effect estimates are presented as HRs per 10 $\mu g/m^3$ increase in annual $PM_{2.5}$. 95% CIs for all models were evaluated by m-out-n blocked bootstrap to account for spatial correlation.

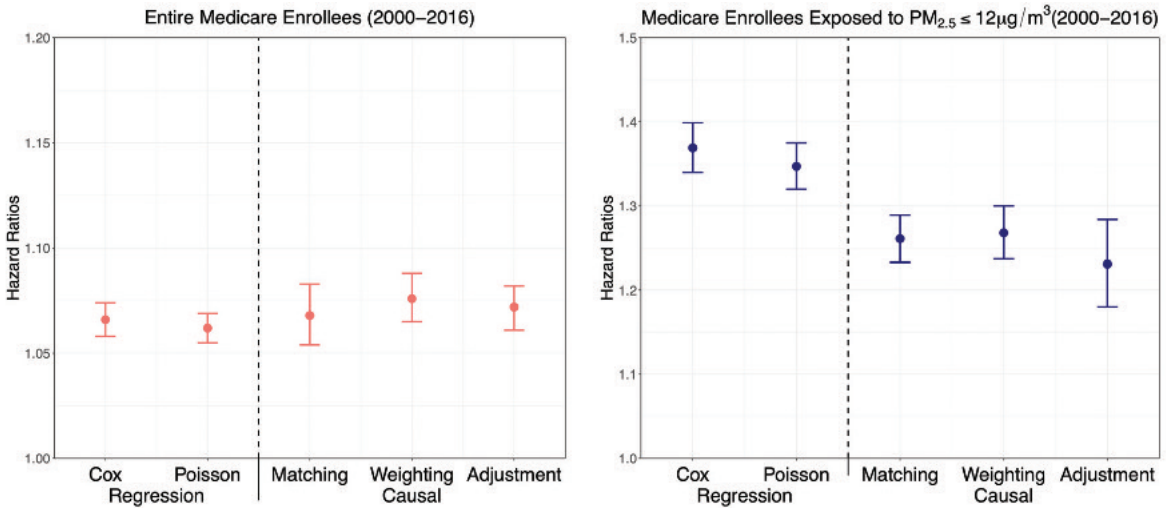


Figure 2: HR and 95% CIs.
The estimated HRs were obtained under five different statistical approaches (two traditional approaches and three causal inference approaches). HRs were adjusted by 10 potential confounders, four meteorological variables, geographic region, and year.

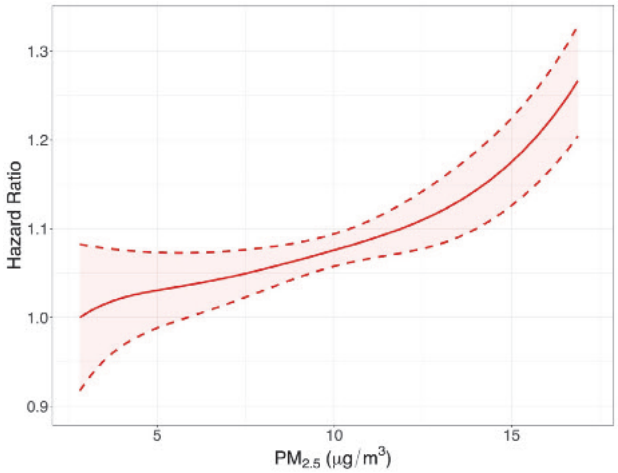


Figure 2. Estimated causal exposure-response curve relating $PM_{2.5}$ to all-cause mortality among Medicare enrollees (2000–2016) with associated 95% confidence bands obtained via bootstrap, only adjusting for one pollutant NO_2 as a potential confounder. We define the baseline rate as the estimated hazard rate corresponding to an exposure level set at the 1st percentile of the distribution of each pollutant. To avoid extrapolation at the support boundaries, we exclude the highest 1% and lowest 1% of pollutants exposures.

4. Conclusions

We report results on the causal link between long-term exposure to $PM_{2.5}$ and mortality, even at $PM_{2.5}$ levels below 12 $\mu g/m^3$, and mortality among Medicare enrollees (65 years of age or older) (Wu et al. 2020). This work relies on newly developed causal inference methods for continuous exposures (Wu et al. 2018b).

Our studies are based on publicly available data sources, and we have made all code developed for our analyses publicly available. Our approach maximizes reproducibility and transparency. We provide robust evidence that the current US standards for $PM_{2.5}$ concentrations are not protective enough and should be lowered to ensure that vulnerable populations, such as the elderly, are safe. Our results raise awareness of the continued importance of assessing the impact of air pollution exposure on mortality.

5. Citations and References

Andersen PK, Gill RD. 1982. Cox's regression model for counting processes: A large sample study. Ann Statist 10:1100-1120.

Chen T, Guestrin C. 2016. Xgboost: A scalable tree boosting system. In kdd '16: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. Available: *https://doi.org/10.1145/2939672.2939785* [accessed June 27 2020].

Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, et al. 2017b. Air pollution and mortality in the medicare population. N Engl J Med 376:2513-2522.

Hirano K, Imbens GW. 2004. The propensity score with continuous treatments. In: Applied bayesian modeling and causal inference from incomplete-data perspectives, (Gelman A, Meng X-L, eds). Hoboken, NJ:John Wiley & Sons, Ltd.

Robins JM, Hernan MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. Epidemiology 11:550-560.

Wu X, Braun D, Kioumourtzoglou MA, Choirat C, Di Q, Dominici F. 2018a. Causal inference in the context of an error prone exposure: Air pollution and mortality. Available: https://arxiv.org/abs/1712.00642 [accessed September 12 2020].

Wu X, Mealli F, Kioumourtzoglou MA, Dominici F, Braun D. 2018b. Matching on generalized propensity scores with continuous exposures. Available: https://arxiv.org/abs/1812.06575 [accessed September 12 2020].

Wu X, Braun D, Kioumourtzoglou MA, Choirat C, Di Q, Dominici F. 2019. Causal inference in the context of an error prone exposure: Air pollution and mortality. Ann Appl Stat 13:520-547.

Wu X, Braun D, Schwartz J, Kioumourtzoglou MA, Dominici F. 2020. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. Sci Adv 6:eaba5692.

Clustering of Attribute Data and Network

Anuška Ferligoj

Clustering of Attribute Data and Network Data

Anuška Ferligoj

Abstract A large class of clustering problems is formulated as an optimization problem in which the best clustering is searched among all feasible clusterings according to a selected criterion function. Clustering problem including attribute data and network data can also be formulated as an optimization problem. Here, the attribute data are considered in the definition of the criterion function and network data by an appropriate definition of the feasibility of clusterings. The agglomerative hierarchical algorithm is adapted for solving this clustering problem. The proposed approach is illustrated by clustering of US counties into regions such that counties inside a region are as similar as possible according to the selected variables (attributes) and form contiguous part of the territory (network constraint). Some open problems are given.

Abstract Un'ampia classe di problemi di clustering è formulata come un problema di ottimizzazione in cui viene ricercato il miglior raggruppamento tra tutti quelli possibili secondo una funzione di criterio selezionata. Anche il problema di individuare cluster in dati di rete con inclusi gli attributi sui nodi, può essere formulato come un problema di ottimizzazione. Qui, gli attributi dei nodi sono inclusi nella definizione della funzione criterio e i dati di rete da un'opportuna definizione di fattibilità dei raggruppamenti. L'algoritmo gerarchico agglomerativo è adattato per risolvere questo problema di raggruppamento. L'approccio proposto è illustrato raggruppando le contee degli Stati Uniti in regioni, tali che le contee all'interno di una regione siano il più simili possibile in base alle variabili selezionate (attributi) e siano contigue per territorio (vincolo di rete). Nel contributo vengono presentati alcuni problemi aperti.

Key words: cluster analysis, network analysis, relation, constraints, US counties, regionalization, contiguity

1 Introduction

[1] Anuška Ferligoj, University of Ljubljana, Slovenia and NRU HSE, Moscow, Russia; email: anuska.ferligoj@fdv.uni-lj.si

Social network analysis has attracted considerable interest from social, behavioural and other science communities in recent decades. Much of this interest can be attributed to the focus of social network analysis on relationship among units, and on the patterns of these relationships. Social network analysis is a rapidly expanding and changing field with broad range of approaches and substantive applications. Among them is network clustering, covering blockmodeling, community detection and many other clustering approaches (Doreian et al., 2020). Clustering attribute data and network data belongs to this class of methods.

A large class of clustering problems can be formulated as an optimization problem in which the best clustering is searched among all feasible clusterings according to a selected criterion function. This clustering approach can be applied to a variety of very interesting clustering problems, as it is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and/or by the definition of the set of feasible clusterings. The clustering problem for clustering of attribute and network data treats attribute data by an appropriately defined criterion function and network data by an appropriately defined set of feasible clusterings.

In the clustering optimization problem, the set of units if finite, therefore, also the set of feasible clusterings is finite and a solution of the clustering problem always exists. Since this set is usually very large, it is not easy to find an optimal solution. In general, most of the clustering problems are NP-hard. For this reason, different efficient heuristic algorithms are used which give "good" results, but not necessarily the best, in a reasonable time. The most important heuristic algorithms are local optimization (e.g., relocation procedure), hierarchical (agglomerative, divisive and adding), k-means (known also as leaders method or dynamic clusters method) and graph theory methods.

2 Clustering of Attribute and Network Data

Suppose that units are described by attribute data (variables) and related by a binary relation that determines the network data. In general, the relation is non-symmetric. The problem is to cluster similar units according to the selected variables, but also considering the network data. The network imposes constraints on the set of feasible clusterings, usually in the following form: each cluster from a feasible clustering induces a subnetwork in the network of the required type of connectedness. Such types of connectedness can be:

- weakly connected units,
- weakly connected units that contain at most one center (with strongly connected units),
- strongly connected units,

- clique, or
- the existence of a trail (all arcs are distinct) containing all the units of the cluster.

Ferligoj and Batagelj (1983, 2000) proved that the types of connectedness (clustering types) determine the minimum number of clusters in a feasible clustering.

Ferligoj and Batagelj (1983) adapted agglomerative hierarchical algorithm for solving some types of clustering problems of attribute and network data. The algorithm begins with two data matrices: the dissimilarity matrix calculated from the attribute data and the network matrix. In each iteration when fusing two clusters dissimilarities between fused cluster and the other clusters has to be updated. This defines different clustering methods (e.g., minimum, maximum, average, Gower, Ward). However, in each iteration also relationships between the new cluster and the other clusters has to be updated using the rules compatible with the types of connectedness in the feasible clusterings.

In the adapted agglomerative hierarchical algorithm (Ferligoj, Batagelj, 1983), a complete dissimilarity matrix is needed. To obtain fast algorithm for large sparse networks we proposed to consider only the dissimilarities between linked units (Batagelj et al., 2014, Ch. 9). This fast algorithm includes the minimum, maximum and average hierarchical methods. It is available in the program Pajek (de Nooy et al., 2018).

Batagelj (2020) presented an interesting application of the adapted agglomerative hierarchical algorithm for non-symmetrical networks following the leader strategy when analysing the weighted citation network between the authors. He considered the links' weights (measuring the intensity of the citation) as similarities between two authors. With the adapted algorithm, he searched for weakly connected units in the citation network that contain a single center.

Many other approaches for clustering attribute data and network data were proposed, but only for the symmetric networks. An example is clustering spatial units where (symmetric) contiguity network is considered. Different clustering approaches were used and adapted, e.g.

- k-means algorithm (e.g., Constanzo, 2001; Młodak, 2021),
- mathematical programming (e.g., Lari, 1998; Duque et al., 2012);
- cluster analysis using spatial autocorrelation (e.g., Hussain, Fuchs, 1996),
- mixed models (Weibel, Walsh, 2008),
- fuzzy clustering model (Coppi et al., 2010),
- supervised machine learning (e.g., Simbahan et al., 2006; Govorov et al., 2019; Kim et al., 2021).

Such clustering approaches have not been proposed for clustering attribute and non-symmetric networks data yet.

3 An Application

The proposed approach to clustering attribute and network data is illustrated by clustering of US counties into regions such that counties inside a region are as similar as possible according to the selected variables (attributes) and form contiguous parts of the territory (network constraint). In this case, the network is symmetric. The research question is the following one: Can we empirically reproduce the map proposed by Garreau (1983) in his book on *The Nine Nations of North America* using our clustering of attribute and network data approach? This is a regionalization problem where smaller territorial units have to be clustered into large ones -- regions such that units inside the region will be similar according to selected attributes (variables) and form contiguous parts of the territory (neighbouring relation). We search for connected clusters.

We analyse 3110 US counties. Variables that are congruent with the Gareau monograph and are available for the year 2000 (with some exceptions) are selected for the following topics: demography, age, education, poverty, race, income, labour force, employment, housing, crime, land, water, and political topic. The variables are standardized and the Euclidean distances between linked counties, are computed. The tolerant strategy (each cluster induces a connected subnetwork) with the Maximum hierarchical method is applied.

From the obtained dendrogram there is an obvious split into two regions (one with 1996 counties and the other with 1080 counties). We preserved the largest 15 clusters with at least 20 counties and all others are considered as unclassified counties. There are 34 counties as outliers (e.g., New York City and adjacent counties, for example those on Long Island). Counties that contain large cities and university cities are quite different from their neighbourhoods and it is possible to assign these counties by a post-processing to the corresponding neighbourhoods.

The obtained regionalization is similar to the Gareau regionalization. Some differences between our obtained clusters and the clusters from Garreau's book can be the result of changes in the last decades and the fact that some variables mentioned by Garreau are not available.

4 Open Problems

For clustering attribute and symmetric network data, many different approaches were developed as discussed in Section 2. There is a lack of the proposed approaches for clustering attribute and non-symmetric binary network data. Here, the adapted hierarchical agglomerative algorithm was described for non-symmetric networks. The local optimization algorithms (e.g., relocation algorithm) could be also adapted, in a similar way to that proposed by Ferligoj and Batagelj (1982) for

the symmetric networks. Here, it is necessary to provide procedures for the generation of the random initial feasible clusterings and for testing of the selected type of the connectedness. In a similar way, also other optimization clustering approaches could be used for clustering attribute and non-symmetric network data.

Another open problem is to develop approaches for clustering attribute and valued networks (symmetric or non-symmetric). One possible direction how to solve such problem is to define such a problem as a two-criteria clustering problem (Ferligoj, Batagelj, 1992).

References

1. Batagelj V.: Clustering approaches to networks. In: P. Doreian, V. Batagelj, A. Ferligoj (eds.), Advances in Network Clustering and Blockmodeling, Wiley, Chichester, 65-104 (2020)
2. Batagelj V., Ferligoj A.: Clustering relational data. In: W. Gaul, O. Opitz, M. Schader (eds.), Data Analysis, Springer, Berlin, 3-15 (2000)
3. Batagelj, V., Doreian, P., Ferligoj, A. Kejžar, N.: Understanding Large Temporal Networks and Spatial Networks. Wiley, Chichester (2014)
4. Coppi, R., D'Urso, P., Giordani, P.: A fuzzy clustering model for spatial multivariate time series. Journal of Classification, **27**, 54–88 (2010)
5. Costanzo, G.D.: A constrained k-means clustering algorithm for classifying spatial units. Statistical Methods and Applications, **10**, 237–256 (2001)
6. Doreian, P., Batagelj, V., and Ferligoj, A. (eds.): Advances in Network Clustering and Blockmodeling, Wiley, Chichester (2020)
7. Duque, J.C., Anselin, L. Rey, S.J.: The max-p-regions problem. Journal of Regional Science, **52**, 397-419 (2012)
8. Ferligoj A., Batagelj V.: Clustering with relational constraint. Psychometrika, **47**, 413-426 (1982)
9. Ferligoj A., Batagelj V.: Some types of clustering with relational constraints. Psychometrika, **48**, 541-552 (1983)
10. Ferligoj A., Batagelj V.: Direct multicriteria clustering algorithms. Journal of Classification, **9**, 43-61 (1992)
11. Garreau J.: The Nine Nations of North America. Houghton Mifflin (1981)
12. Govorov, M., Beconytė, G., Gienko, G., Putrenko, V.: Spatially constrained regionalization with multilayer perceptron. Transactions in GIS, **23**, 1048-1077 (2019)
13. Hussain M., Fuchs K.: Cluster analysis using spatial autocorrelation. In: Bock H.H., Polasek W. (eds) Data Analysis and Information Systems, pp 52-63, Springer, Berlin (1996)
14. Kim, D., Jung, S., Jeong, Y.: Theft prediction model based on spatial clustering to reflect spatial characteristics of adjacent Lands. Sustainability, **13**, 7715 (2021) doi.org/10.3390/su13147715
15. Lari I., Maravalle M., Simeone B.: A linear programming based heuristic for a hard clustering problem on trees. In: Rizzi A., Vichi M., Bock HH. (eds.) Advances in Data Science and Classification, pp 161-170, Springer, Berlin (1998)
16. Młodak, A.: k-means, Ward and probabilistic distance-based clustering methods with contiguity constraint. Journal of Classification, **38**, 313-352 (2021)
17. Simbahan, G.C., Dobermann, A.: An algorithm for spatially constrained classification of categorical and continuous soil properties. Geoderma, **136**, 504-523 (2006)
18. Weibel, E. J., Walsh, J.P.: Territory analysis with mixed models and clustering. In: 2008 CAS Discussion Paper Program: Applying Multivariate Statistical Models, pp. 91–169, Casualty Actuarial Society, Arlington, VA (2008)

Bayesian approaches for capturing the heterogeneity of neuroimaging experiments

Francesco Denti, Laura D'Angelo and Michele Guindani

Bayesian approaches for capturing the heterogeneity of neuroimaging experiments

Metodi Bayesiani per descrivere la variabilità degli esperimenti di neuroscience

Francesco Denti and Laura D'Angelo and Michele Guindani

Abstract In the neurosciences, it is now widely established that brain processes are characterized by heterogeneity at several levels. For example, neuronal processes differ by external stimuli, and patterns of brain activations vary across subjects. In this paper, we will discuss a few Bayesian strategies for characterizing heterogeneity in the neurosciences, where time-series data are assumed to be organized in different, but related, units (e.g., neurons and/or regions of interest) and some sharing of information is required to learn distinctive features of the units. First, we will discuss models for multi-subject analysis that will identify population subgroups characterized by similar brain activity patterns, also by integrating available subject information. Then, we will look at how novel techniques in intracellular calcium signals may be used to analyze neuronal responses to external stimuli in awake animals. Finally, we will discuss a mixture framework for identifying differentially activated brain regions that can classify the brain regions into several tiers with varying degrees of relevance. The performance of the models will be demonstrated by applications to data from human fMRI and animal fluorescence microscopy experiments.

Abstract *Nelle neuroscienze è ormai ampiamente stabilito che i processi cerebrali sono caratterizzati da eterogeneità a più livelli. Ad esempio, i processi neuronali differiscono in base agli stimoli esterni e i tipi di attivazione cerebrale variano tra i soggetti. In questo manoscritto, discuteremo alcune strategie bayesiane per caratterizzare l'eterogeneità nelle neuroscienze, in cui si presume che i dati delle serie temporali siano organizzati in unità diverse, ma correlate (ad esempio, neuroni e/o regioni di interesse) e una certa condivisione di informazioni sia necessaria per apprendere le caratteristiche distintive delle unità neuronali. In primo luogo, discuteremo modelli per l'analisi multi-soggetto che identificheranno sottogruppi di popo-*

Francesco Denti
Dipartimento di Scienze Statistiche, Universitá Cattolica del Sacro Cuore, Milano, Italia e-mail: francesco.denti@unicatt.it

Laura D'Angelo
Dipartimento di Economia, Metodi Quantitativi e Strategie di Impresa, Universitá degli Studi di Milano - Bicocca, Milano, Italia e-mail: laura.dangelo@unimib.it

Michele Guindani
Department of Biostatistics, University of California, Los Angeles, USA e-mail: micheleguindani@gmail.com

lazione caratterizzati da modelli di attività cerebrale simili, anche integrando le informazioni sui soggetti disponibili. Quindi, esamineremo come nuove tecniche nei segnali intracellulari di calcio possono essere utilizzate per analizzare le risposte neuronali agli stimoli esterni negli animali svegli. Infine, discuteremo un modello mistura per identificare regioni cerebrali attivate in modo differenziale che possono classificare le regioni cerebrali in diversi livelli con vari gradi di importanza. Le prestazioni dei modelli saranno dimostrate mediante applicazioni ai dati provenienti da esperimenti di fMRI in umani e di microscopia a fluorescenza in animali.

Key words: Bayesian methods, Heterogeneity, Clustering, Neurosciences

1 Introduction

Standard methods in brain research have long assumed that it is possible to group together the brain maps of all subjects in a study. Indeed, average maps have been typically used to investigate different aspects of brain functioning. Some commonly employed preprocessing steps also encode dimension reduction steps [e.g., GICA in fMRI studies, 7] and implicitly assume the existence of common patterns across subjects (e.g. by encouraging to match ICA components across subjects). However, assuming spatial homogeneity of brain patterns may lead to a reduced ability to capture inter-subject variability [24].
There is an increasing recognition that brain functioning is heterogeneous and varies greatly both within and between individuals, either because of differences in activation patterns shown in response to a series of stimuli, or due to differences in brain connectivity (i.e. how regions of the brain interact with each other). Finally, the different brain activity patterns may be differently associated to a clinical outcome or to different behaviors [e.g. large brain responses to food-related cues predict cue-induced eating, 34].

In this manuscript, we will discuss in detail three frameworks where Bayesian methods have been successfully used for describing the heterogeneity of brain patterns. More specifically, we will briefly discuss the use of hierarchical mixture models to conduct multi-subject inference (Section 2), methods to capture activity spikes in *in-vivo* experiments in animals (Section 3) and a hierarchical mixture model approach to capture the heterogeneity of the signal in brain regions activated during a neuroimaging experiment (Section 4). We will then provide some concluding remarks.

2 Mixtures for capturing between-subjects heterogeneity.

Functional magnetic resonance imaging (fMRI) is a noninvasive neuroimaging technique which measures the blood oxygenation level dependent (BOLD) contrast, i.e.

the difference in magnetization between oxygenated and deoxygenated blood arising from changes in regional cerebral blood flow. Common modeling approaches for the analysis of task-related fMRI data rely on the linear model formulation that was first proposed by Friston et al. [17]. In multi-subject studies two-stage "group analysis" approaches are often adopted as computationally attractive methods where summary estimates of model parameters are obtained at the individual level and then used in a second stage model at the group/population level [4, 31, 21].
In Zhang et al. [38], a unified, single stage, and probabilistically coherent Bayesian framework is proposed for the analysis of task-related brain activity in multi-subject fMRI experiments. More specifically, let $Y_{iv} = (Y_{iv1}, \ldots, Y_{ivT})^T$ be the $T \times 1$ vector of the BOLD response data at the v-th region in the ith subject, with $i = 1, \ldots, N, v = 1, \ldots V$, and with the symbol $(\cdot)^T$ indicating the transpose operation. Then, the BOLD time-series response can be modeled with a general linear model

$$Y_{iv} = X_{iv}\beta_{iv} + \varepsilon_{iv},\ \varepsilon_{iv} \sim N_T(0, \Sigma_{iv}), \tag{1}$$

where X_{iv} is a $T \times p$ covariate matrix encoding the hemodynamics of the experimental stimulus [22], $\beta_{iv} = (\beta_{iv1}, \ldots, \beta_{ivp})^T$ is a $p \times 1$ vector of regression coefficients and $\varepsilon_{iv} = (\varepsilon_{iv1}, \ldots, \varepsilon_{ivT})^T$ is a $T \times 1$ vector of errors. The error terms in (1) capture temporal correlation in the fMRI data and are typically assumed autocorrelated, accounting for both hardware and subject-related noise.

The identification of brain areas activated in response to a stimulus reduces to a problem of variable selection, i.e., the identification of the nonzero β_{iv}. In the Bayesian framework, the selection can be achieved imposing a mixture prior, often called *spike-and-slab* prior, on the regression coefficients [37, 19, 20]. Zhang et al. [38] embed the selection into a clustering framework and effectively define a multi-subject nonparametric variable selection prior with spatially informed selection within each subject. More specifically, let γ_{iv} be the binary indicator of whether voxel v in subject i is active or not, i.e., $\gamma_{iv} = 0$ if $\beta_{iv} = 0$ and $\gamma_{iv} = 1$ otherwise. Zhang et al. [38] impose a spiked hierarchical Dirichlet Process [HDP, 32] prior on β_{iv}, i.e. a spike-and-slab prior where the slab distribution is modeled by a HDP prior,

$$\begin{aligned} \beta_{iv}|\gamma_{iv}, G_i &\sim \gamma_{iv} G_i + (1-\gamma_{iv})\delta_0 \\ G_i|\eta_1, G_0 &\sim DP(\eta_1, G_0) \\ G_0|\eta_2, P_0 &\sim DP(\eta_2, P_0) \\ P_0 &= N(0, \tau), \end{aligned} \tag{2}$$

with δ_0 a point mass at zero, with τ fixed, η_1, η_2 the mass parameters and P_0 the base measure. With this prior formulation, the subject-specific distribution G_i varies around a population-based distribution G_0, which is centered around a known parametric model P_0. The mass parameters η_1 and η_2 control the variability of the distribution of the coefficients at the subject and population level, respectively. Both G_i and G_0 can be written as a mixture of point masses as $G_i = \sum_{k=1}^{\infty} \pi_{ik}\delta_{\phi_k}$ and $G_0 = \sum_{k=1}^{\infty} \xi_k \delta_{\phi_k}$, where δ_x indicates a point mass at x and the mixture weights are given, respectively, by $\pi_{ik} = \pi'_{ik}\prod_{l=1}^{k-1}(1-\pi'_{il})$, with $\pi'_{ik} \sim \text{Beta}(\eta_1\xi_k, \eta_1(1-\sum_l^k \xi_l))$,

and $\xi_k = \xi'_k \prod_l^{k-1}(1-\xi'_l)$, with $\xi'_k \sim \text{Beta}(1,\eta_2)$, see Sethuraman [29]. The mixture representation highlights the fact that G_i and G_0 share common atoms $\phi_k \sim P_0$ and thus naturally induce clustering of the β_{iv}'s in (2). As a result, the coefficients β_{iv}'s may be effectively shared across active voxels within a subject as well as between subjects. In order to take into account information on the anatomical structure of the brain, in particular the correlation between neighboring voxels, Zhang et al. [38] place a Markov Random Field (MRF) prior on the selection parameter γ_{iv},

$$P(\gamma_{iv}|d,e,\gamma_{ik},k\in N_{iv}) \propto \exp(\gamma_{iv}(d+e\sum_{k\in N_{iv}}\gamma_{ik})), \tag{3}$$

with N_{iv} the set of neighboring voxels of voxel v in subject i. The sparsity parameter $d \in (-\infty,\infty)$ represents the expected prior number of activated voxels. The smoothing parameter $e > 0$ controls the probability of identifying a voxel as active based on the activation of its neighboring voxels. Zhang et al. [38] show that inference via variational Bayes achieves satisfactory results in the selection of activated regions at a much reduced computational costs than using a Markov Chain Monte Carlo algorithm. In an application to case study data, the method successfully detected activations in the occipital areas during presentation of visual stimuli, whereas no activations were detected in the frontal areas. They also showed that a multi-subject modeling strategy leads to a more accurate detection of the activated areas than single-subject models.

3 Capturing the heterogeneity of the distribution of neuronal spikes

Technological advancements in the development of miniaturized fluorescence microscopes—light enough to be worn by a freely behaving rodent—have recently enabled the visualization of the activity of individual neurons recorded over time. In particular, the technique of calcium imaging has been paramount in allowing scientists to visualize the activity of large populations of neurons in awake animals in response to external stimulation [1, 27]. Neurons' firing events (i.e., neuronal activations) are rendered through transient peaks in the intra-cellular calcium levels, and the amplitudes of these peaks can be analyzed to measure the intensity of the response.

The availability of these data, however, has also called the attention to the complexity of neuronal processes while encoding external information, and the need to devise adequate methods for analysis. Even when focusing just on the activity of a single neuron, there are several modeling and computational challenges one has to deal with. The observed calcium trace is only a proxy of the underlying neuronal activity, which has to be extracted through the use of deconvolution techniques [36]: the output of this phase is the so called *spike train*, which is the series of the observed firing events. Then, the series of estimated activation spikes has to be analyzed and

associated with the experimental conditions. Neurons' response to stimulation can indeed be very heterogeneous, and it is of interest to investigate how the frequency and amplitudes of spikes vary over time and across conditions [5].

To this purpose, D'Angelo et al. [15] have recently proposed a Bayesian nested finite mixture model that simultaneously allows deconvolving the signal and analyzing how the extracted activity varies in response to different experimental conditions. First, a biophysical model [35] is used to describe the calcium dynamics at each time $t = 1,\dots,T$ in order to deconvolve the signal. Let y_t, $t = 1,\dots,T$ denote the observed fluorescence trace. Then, the model assumes that the observed fluorescence trace y_t is a noisy realization of the underlying calcium level c_t,

$$y_t = b + c_t + \varepsilon_t, \tag{4}$$

Then, the calcium dynamics are modeled using an autoregressive process with jumps at the neuron's firing events,

$$c_t = \gamma c_{t-1} + A_t + w_t. \tag{5}$$

The parameters A_t are the major focus of the analysis: at each time they describe either the absence of activity ($A_t = 0$) or the spike amplitude when an activation is detected ($A_t > 0$). D'Angelo et al. [15] built a prior distribution for the parameters A_t under the assumption of varying experimental conditions. As motivating application, they considered calcium imaging data from the Allen Brain Observatory [2, 9], a large and public data repository. The experiment investigates how a neuron located in the mouse visual cortex responds to different types of visual stimuli. See Fig. 1 for an illustration of the dataset.

D'Angelo et al. [15] adapt the common atom model of Denti et al. [11] and the generalized mixtures of finite mixtures of Frühwirth-Schnatter et al. [18] to the context of calcium imaging studies. Their modeling framework allows identifying similarities in the distributional patterns of the neuronal responses to different stimuli, and clustering the spikes' intensities within and between experimental conditions. For the parameters A_t it is assumed a spike-and-slab specification, with a Dirac mass at zero modeling the absence of activity, and a Gamma density modeling the positive amplitudes.

Figure 1 shows the fluorescence trace recorded for an illustrative neuron in the Allen Brain Observatory dataset, together with the estimated neuronal activity: it is evident the heterogeneity of the response to the different visual stimuli. The application of the model above led to the estimation of three distributional clusters. More specifically, the neuronal response was similar during two of the four experimental conditions, and different in the others. Finally, the estimated firing rates and spikes' amplitudes were coherent with a known behaviour of the neurons, which exhibit a more intense activity during the more complex stimuli [28].

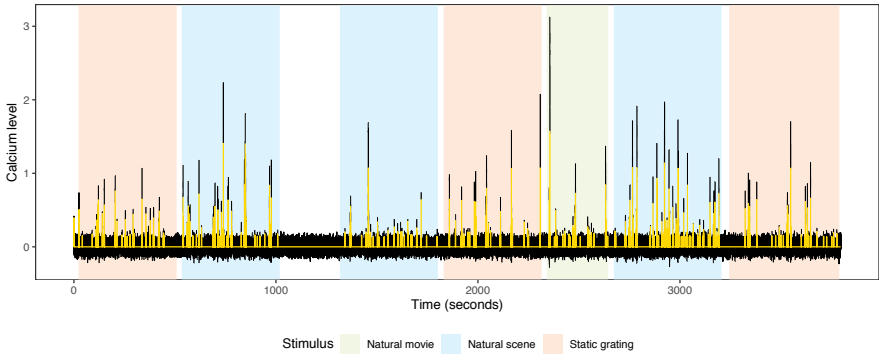


Fig. 1 Observed fluorescence trace y_t from the Allen Brain Observatory data (black line), and visual stimulus (shaded areas). The yellow line represents the estimated spike train in [15].

4 Mixture for improving hypothesis testing: the two-group model and the Horseshoe Mix

The primary objective of many brain-imaging analyses is the identification of brain regions that activate in conjunction with a particular task of interest. Often, the detection problem is tackled from a multiple hypothesis testing perspective, where a null hypothesis (e.g., $H_0^{(i)}$: region i is inactive) is tested for every region of interest (e.g., pixel, voxel, brain subregions, etc.). Over the years, numerous statistical approaches have been developed to account for the multiplicity induced by the large amount of tests, e.g. by adjusting the p-values generated from the testing procedure [3, 30].

The two-group model (2GM) of Efron [16] has received wide attention in the multiple testing literature. Suppose we are evaluating the results of n tests, and consider a vector of z-scores $\boldsymbol{z} = \{z_i\}_{i=1}^n$ to test the i-th null hypothesis $H_0^{(i)}$. The two-group model separates the z-scores using a two-component mixture:

$$f(z) = \pi_0 f_0(z) + (1-\pi_0) f_1(z).$$

Here, $1-\pi_0$ is the (expected) proportion of relevant tests while f_0 and f_1 denote an (empirical) *null* distribution and the *alternative* distribution, respectively. The empirical null distribution should be carefully modeled, as it should reflect the theoretical distribution of the test statistic under the null hypothesis. The alternative distribution f_1 is assumed as longer tailed than f_0. From a mixture model perspective, the model partitions the test statistics into two groups: relevant (when generated by f_1) and irrelevant (when generated from f_0). Many extensions of this modeling framework have appeared in the Bayesian literature [26, 14, 23, 13, 12]. One of the major issues is that the decisions are still dichotomized, whereas a recent push calls for multiple label group testing to ensure better control of false positives [33, 25].

The multi-comparison problem can be also seen from an estimate-regularization

point of view. Recently, Denti et al. [10] have proposed a shrinkage prior obtained as a mixture of Horseshoe distributions [8]. More specifically, their Horseshoe Mix (HSmix) prior assumes a *multi-group model*, where

$$z_i|\beta_i,\sigma^2 \sim \mathcal{N}(\beta_i,\sigma^2), \quad \beta_i|\lambda,\tau,\sigma \sim \sum_{l\geq 1} \pi_l\, \phi(0,\lambda_l^2\,\tau^2\,\sigma^2), \quad \lambda_l \sim \mathscr{C}^+(0,1). \quad (6)$$

Here, each coefficient β_i, $i = 1,\ldots,n$, represents the true underlying signal, whose estimates are shrunk via a continuous scale mixture of Gaussians densities $\phi(\cdot)$. The mixture in (6) can be finite or infinite, according to the specification of the weights $\boldsymbol{\pi}$. A global shrinkage parameter τ (here considered fixed) and a set of mixture-component shrinkage parameters $\boldsymbol{\lambda} = \{\lambda_l\}_{l\geq 1}$ define several shrinkage levels in the mixture. The mixture component characterized by the lowest variance is representative of the null distribution. All the other components are then ranked in increasing order, representing different degrees of statistical relevance. The model allows sharing of information across tests. Most importantly, the induced clustering overcomes the traditional "significant vs. non-significant" paradigm in hypothesis testing, by ranking the regions into tiers of relevance and capturing signals otherwise lost within the canonical binary decision framework.

4.1 Capturing Activations in fMRI studies via the Horseshoe Mix

In this Section, we showcase the use of the HSmix prior in an application to a fMRI dataset. More specifically, we consider the single-subject fMRI data collected during an *attention* experiment by [6] and analyzed more recently in [37]. In this experiment, a subject was asked to closely follow fixed and moving points in the middle of a transparent screen. See Section 3.4 of [37] for more details about the experiment. We focus on a single brain slice containing the primary visual cortex (V1). Thus, we analyze the signal recorded over 2D brain images with a resolution of 53 by 64 pixels replicated over 360 times. To filter out irrelevant pixels (e.g., recorded in the image but outside the patient's head), we first exclude those pixels that present a mean signal over time lower than 300. Then, we mimic the preprocessing steps outlined by [37], using a wavelet-based transformation to filter out the temporal correlation. At the end of the pre-processing, we are left with 2,366 pixels and fMRI data recorded over 320-time points. We fit a multivariate linear model, regressing the transformed fMRI over the convolved stimulus. Differently from [37], in the application of the general linear model of [17] we use a canonical hemodynamic response function to convolve the stimulus pattern over time. We also do not explicitly account for any spatial relationships, which may induce residual correlations across the test statistics.
From the linear model estimation, we compute $n = 2,366$ z-scores, which we further analyze employing a non-parametric version of the HSmix model. Figure 2 displays the magnitude of the test statistics (left panel) and their regularized version

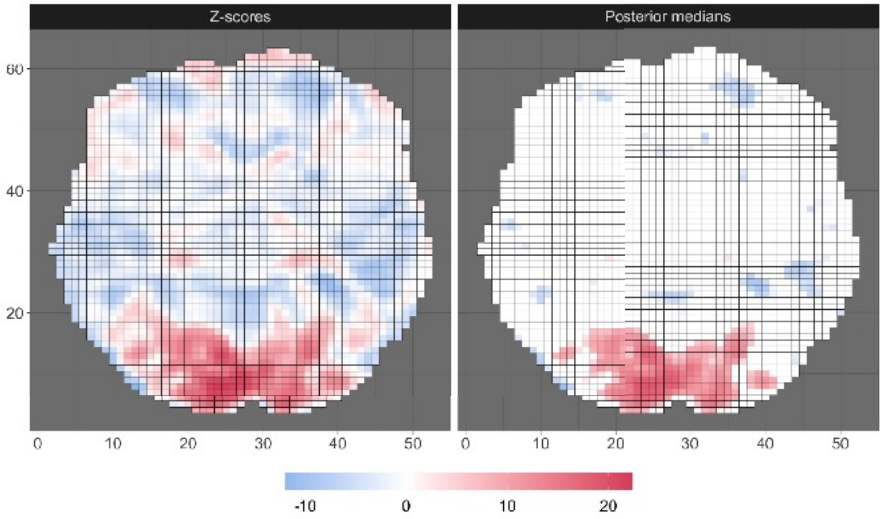


Fig. 2 Heatmaps comparing the pixel-specific z-scores (left panel) versus their regularized estimates obtained with the HSmix model (right panel - posterior medians).

(posterior medians of the coefficients, right panel). Numerous noisy fluctuations are shrunk to zero, and the activated areas appear more evident.

Moreover, from the HSmix model results, we are able to cluster the pixels into three relevance tiers. In Figure 3, we compare the resulting relevance tiers (reported in the bottom-right panel) with other three screening methods: naive thresholding of p-values (level: 0.05, top-left panel), thresholding of Benjamini-Hochberg adjusted p-values (level: 0.05, top-right panel), and thresholding of the local-FDR resulting from the empirical Bayes estimate of the two group model (FDR level: 0.20, bottom-left panel). On the one hand, the local-FDR method tends to be the more conservative, flagging as relevant only the pixels with high-magnitude z-scores. On the other hand, relying on p-values produces numerous discoveries, a result that might be hindered by the unaccounted spatial correlation across the pixels. The HSmix provides a good trade-off, detecting areas of high activation (red) and recovering other areas that can be regarded as only mildly active (e.g., the areas in orange on the left side of the brain).

5 Conclusions

The development of neuroimaging biomarkers for targeted interventions requires to take into account the complexity and heterogeneity of brain functioning. In this paper, we review three distinct mixture-based frameworks for analyzing the variability of brain signals in either humans or animals. Hierarchical Bayesian methods allow

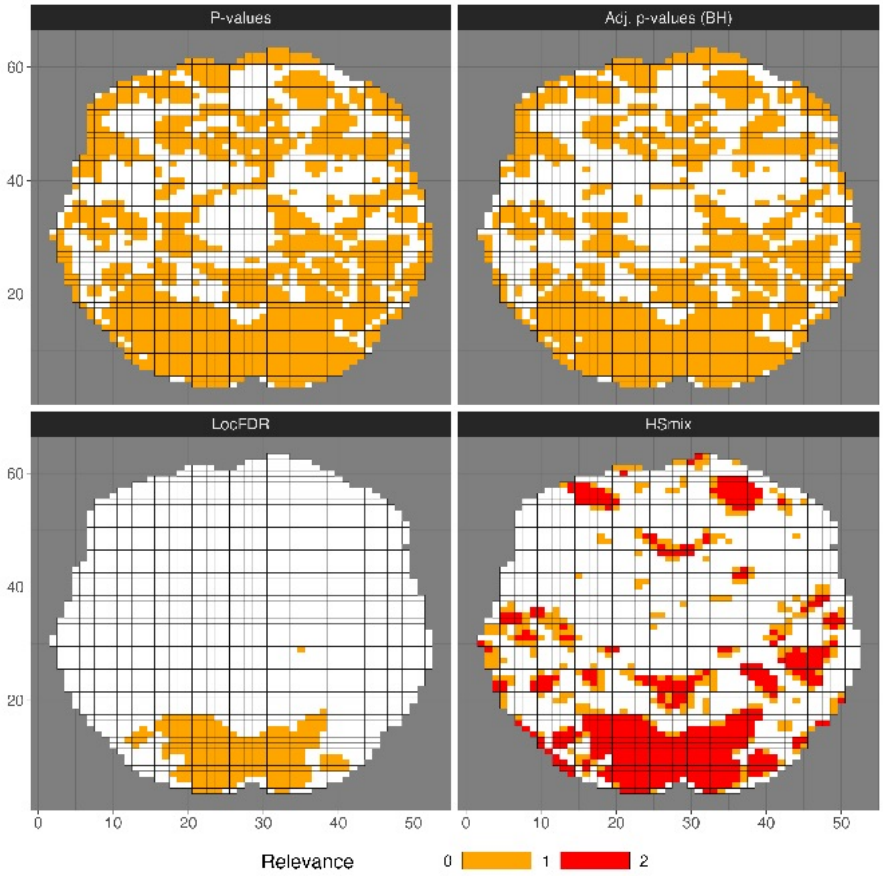


Fig. 3 Comparison of the discoveries obtained with four different screening procedures. The Relevance level 0 indicates no activation.

for an elegant borrowing of information across and within subjects, but they also present challenges. Computational scalability is a major challenge, which may lead to the investigation of dimension reduction techniques and approximate inference solutions. The path ahead is long, complex, and wrought with obstacles. However, the ultimate reward will be highly gratifying: a better understanding of how different humans think and how they respond to external input. Statistical approaches and a close collaboration with neuroscientists are crucial for a successful journey.

References

[1] Daniel Aharoni, Baljit S. Khakh, Alcino J. Silva, and Peyman Golshani. All the light that we can see: a new era in miniaturized microscopy. *Nature Methods*, 16(1):11–13, 2019.

[2] Allen Institute MindScope Program. Allen Brain Observatory – 2-photon visual coding [dataset]. brain-map.org/explore/circuits, 2016.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discover rate – a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 75(1):289–300, 1995.

[4] F. Bowman, B. Caffo, S. Bassett, and C. Kilts. A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, 39(1):146–156, 2008.

[5] Naama Brenner, Oded Agam, William Bialek, and Rob de Ruyter van Steveninck. Statistical properties of spike trains: universal and stimulus-dependent aspects. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 66:031907, Sep 2002.

[6] Christian Büchel and Karl J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7(8):768–778, 1997. ISSN 10473211. doi: 10.1093/cercor/7.8.768.

[7] Vince Calhoun, Tülay Adali, Godfrey Pearlson, and J. T. Group ica of functional mri data: Separability, stationarity, and inference. *Proceedings of the International Conference on ICA and BSS*, 01 2002.

[8] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[9] Saskia de Vries, Jerome Lecoq, Michael Buice, Peter Groblewski, Gabriel Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, Kate Roll, Marina Garrett, Tom Keenan, Chihchau Kuan, Stefan Mihalas, Shawn Olsen, Carol Thompson, Wayne Wakeman, Jack Waters, and Christof Koch. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23 (1):138–151, 2020.

[10] Francesco Denti, Ricardo Azevedo, Chelsie Lo, Damian Wheeler, Sunil P. Gandhi, Michele Guindani, and Babak Shahbaba. A Horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging. *Arxiv Preprint*, 2021.

[11] Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira. A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, 2021.

[12] Francesco Denti, Michele Guindani, Fabrizio Leisen, Antonio Lijoi, William Duncan Wadsworth, and Marina Vannucci. Two-group Poisson-Dirichlet mixtures for multiple testing. *Biometrics*, 77(2):622–633, 2021. ISSN 15410420. doi: 10.1111/biom.13314.

[13] Francesco Denti, Stefano Peluso, Michele Guindani, and Antonietta Mira. Multiple hypothesis screening using mixtures of non-local distributions. *Arxiv Preprint*, 2022.

[14] Kim Anh Do, Peter Müller, and Feng Tang. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(3):627–644, 2005. ISSN 00359254. doi: 10.1111/j.1467-9876.2005.05593.x.

[15] Laura D'Angelo, Antonio Canale, Zhaoxia Yu, and Michele Guindani. Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. *Biometrics*, pages 1–13, 2022. doi: 10.1111/biom.13626.

[16] B. Efron. Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23:1–22, 2008.

[17] K. J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994.

[18] Sylvia Frühwirth-Schnatter, Gertraud Malsiner-Walli, and Bettina Grün. Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279 – 1307, 2021.

[19] S. Kalus, P.G. Sämann, and L. Fahrmeir. Classification of brain activation via spatial Bayesian variable selection in fMRI regression. *Advances in Data Analysis and Classification*, 8:63–83, 2013.

[20] K. Lee, G.L. Jones, B.S. Caffo, and S.S. Bassett. Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis*, 9(3):699–732, 2014.

[21] Xiang Li, Dajiang Zhu, Xi Jiang, Changfeng Jin, Xin Zhang, Lei Guo, Jing Zhang, Xiaoping Hu, Lingjiang Li, and Tianming Liu. Dynamic functional connectomics signatures for characterization and differentiation of PTSD patients. *Human Brain Mapping*, 35(4):1761–1778, 2014.

[22] M.A. Lindquist, J.M. Loh, L.Y. Atlas, and T.D. Wager. Modeling the hemodynamic response function in fMRI: Efficiency, bias, and mis-modeling. *NeuroImage*, 45:187–198, 2009.

[23] Ryan Martin and Surya T. Tokdar. A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–439, 2012. ISSN 14654644. doi: 10.1093/biostatistics/kxr039.

[24] Andrew M. Michael, Mathew Anderson, Robyn L. Miller, Tülay Adalı, and Vince D. Calhoun. Preserving subject variability in group fmri analysis: performance evaluation of gica vs. iva. *Frontiers in Systems Neuroscience*, 8, 2014. ISSN 1662-5137.

[25] Chul Moon and Nicole A. Lazar. Hypothesis testing for shapes using vectorized persistence diagrams, 2020.

[26] Omkar Muralidharan. An empirical Bayes mixture method for effect size and false discovery rate estimation. *Annals of Applied Statistics*, 6(1):422–438, 2012. ISSN 19326157. doi: 10.1214/09-AOAS276.

[27] Miho Nakajima and L. Ian Schmitt. Understanding the circuit basis of cognitive functions using mouse models. *Neuroscience Research*, 152:44 – 58, 2020. ISSN 0168-0102.

[28] Simon Peter Peron and Fabrizio Gabbiani. Role of spike-frequency adaptation in shaping neuronal response to dynamic stimuli. *Biological cybernetics*, 100 (6):505–520, 06 2009.

[29] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[30] J.D. Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (3):347–368, 2007.

[31] S. Su, B. Caffo, E. Garrett-Mayer, and S. Bassett. Modified test statistics by inter-voxel variance shrinkage with an application to fMRI. *Biostatistics*, 10 (2):219–227, 2009.

[32] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[33] Mikael Vejdemo-Johansson and Sayan Mukherjee. Multiple hypothesis testing with persistent homology. In *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*, 2020.

[34] Francesco Versace, David W. Frank, Elise M. Stevens, Menton M. Deweese, Michele Guindani, and Susan M. Schembre. The reality of "food porn": Larger brain responses to food-related cues than to erotic images predict cue-induced eating. *Psychophysiology*, 56(4), 2019.

[35] J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology*, 104(6):3691–3704, 2010.

[36] Joshua T. Vogelstein, Brendon O. Watson, Adam M. Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophysical Journal*, 97(2):636 – 655, 2009.

[37] L. Zhang, M. Guindani, F. Versace, and M. Vannucci. A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, 95:162–175, 2014.

[38] L. Zhang, M. Guindani, F. Versace, J.M. Engelmann, and M. Vannucci. A spatio-temporal nonparametric Bayesian model of multi-subject fMRI data. *Annals of Applied Statistics*, 10(2):638–666, 2016.

2 Specialized Sessions

Advances in Bayesian nonparametric methodology

Repulsive mixture models for high-dimensional data

Modelli mistura con prior di tipo repulsiva per dati altamente dimensionali

Lorenzo Ghilotti, Mario Beraha and Alessandra Guglielmi

Abstract Model-based clustering is customarily achieved in the Bayesian setting through finite or infinite mixture models, assuming that the p-dimensional data are iid generated from homogeneous populations, represented by parametric densities. The poor performance of Bayesian mixtures in the large-p setting is known and they may lead to inconsistent cluster estimates when p increases to infinity. We build on a class of mixtures of latent factor models, similar to the model in [4], mixing over the latent parameters. Our main contribution to the model is the assumption of a repulsive point process as mixing measure. The matrix of factor loadings drives the anisotropic behavior, so that separation is indeed induced between the high-dimensional centers of different clusters. We also propose a MCMC algorithm which extends a conditional algorithm for repulsive mixture models, introduced previously in the literature.

Abstract *In ambito bayesiano, i modelli per il clustering sono solitamente i modelli mistura, con un numero finito od infinito di componenti. In pratica, si assume che i dati p dimensionali siano iid da popolazioni omogenee, rappresentate da densità parametriche. Tuttavia, quando p è grande, tali modelli non sono particolarmente efficaci e quando p tende ad infinito si possono ottenere stime del numero dei cluster che sono inconsistenti. Qui presentiamo una classe di misture di modelli a fattori latenti come in [4], misturando rispetto ai fattori latenti. Il nostro principale contributo modellistico è l'assunzione di un processo di punto repulsivo come misura misturante. La matrice dei fattori latenti guida il comportamento anisotropo che vogliamo includere nel modello, ma viene anche inclusa nell'indurre la separazione tra i cluster. Proponiamo anche un algoritmo MCMC che estende un algoritmo condizionale per le misture repulsive, che è già apparso in letteratura.*

Lorenzo Ghilotti[1], Mario Beraha [2,3]and Alessandra Guglielmi[3]
[1] Department of Economics, Management and Statistics, Università degli Studi di Milano Bicocca, Milano, Italy
[2] Department of Computer Science, Università di Bologna, Bologna, Italy
[3] Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: l.ghilotti@campus.unimib.it, {mario.beraha, alessandra.guglielmi}@polimi.it

Key words: latent factor models, determinantal point processes, model-based clustering.

1 Introduction

In this paper, we consider model-based clustering for high-dimensional data. Let $y_1, \dots, y_n \in \mathbb{R}^p$ represent the data we aim at clustering. In this talk, we focus on the large-p setting, i.e. for instance when p is in the order of hundreds or thousands, and possibly larger than the sample size n. Cluster analysis might be particularly useful for such high-dimensional datasets, as it provides a straightforward procedure to explore the data by exploiting the *latent* structure arising from similar observations. Model-based clustering is customarily achieved in the Bayesian setting through finite or infinite mixture models; see [5] for a recent review on mixtures. Specifically, the conditional distribution of data, given parameters, under the mixture model takes the form

$$y_1, \dots, y_n \mid \boldsymbol{w}, \boldsymbol{\theta} \overset{\text{iid}}{\sim} p(\cdot) = \sum_{h=1}^{m} w_h f_{\theta_h}(\cdot). \tag{1}$$

Under the Bayesian approach, suitable priors are assumed for the weights $\boldsymbol{w} = (w_1, \dots, w_m)$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$, and m itself. Here $f_{\theta_h}(\cdot)$, the h-th *component* of the mixture, denotes a parametric density for some parameter $\theta_h \in \Theta$. Weights $\boldsymbol{w} = (w_1, \dots, w_m)$ ($w_h \geq 0$, $\sum w_h = 1$) specify the relative frequency of each population f_{θ_h}.

The poor performance of Bayesian mixtures when p is large is known. Not only MCMC algorithms for mixture models scale poorly in general, but, as shown in [4], the choice of the mixture kernel f_θ in (1) might affect consistency. Specifically, [4] show that Gaussian mixtures lead to inconsistent cluster estimates when p increases to infinity: if the covariance matrix is cluster-specific, then with probability one each observation will be clustered into a singleton cluster, while if the covariance matrix is shared through all the clusters, only one cluster is detected. To overcome the problem, [4] propose a latent factor model, where clustering is performed at the latent level, specifically on d-dimensional latent parameters, for d much smaller than p.

In general, when a mixture model is not well specified we can identify a trade-off between the accuracy of cluster detection and density estimate: better density estimates necessarily yield poorer cluster estimates and vice versa. As shown in [3], traditional mixture models tend to favor density over cluster estimates. Repulsive mixture models (see [1] and the references therein) are an attempt to reverse the trade-off in favor of better cluster estimates: by encouraging well-separated components, repulsive mixtures usually have a poorer density estimates but do not overestimate the number of clusters.

Here, we build on a class of mixtures of latent factor models, similar to the model in [4], mixing over the latent parameters. Our main contribution to the model is the assumption of a repulsive point process as mixing measure. The matrix of fac-

tor loadings drives the anisotropic behavior, so that separation is indeed induced between the high-dimensional centers of different clusters. We propose a MCMC algorithm which extends the conditional algorithm introduced in [1] for repulsive mixture models. To sample from the full conditional of the factor loadings, we replace the standard Metropolis step by a Metropolis adjusted Langevin algorithm. We test the model and the algorithm on a simulated data example.

In the next section we describe the repulsive mixture model we propose.

2 Anisotropic repulsive point process latent mixture models

Let $y_1,\dots,y_n \in \mathbb{R}^p$, $\Lambda \in \mathbb{R}^{p\times d}$ a factor loadings matrix, $\eta_1,\dots,\eta_n \in \mathbb{R}^d$ a set of latent factors, and $\Sigma = \text{diag}(\sigma_1^2,\dots,\sigma_p^2)$ $(\sigma_j^2 > 0)$ a diagonal covariance matrix. Let $\mathcal{N}_p$ denote the p–dimensional Gaussian distribution. As in [4], we assume the following model

$$\begin{aligned} y_i \mid \eta_i,\Lambda,\Sigma &\overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda\eta_i,\Sigma), && i=1,\dots,n \\ \eta_i \mid \boldsymbol{w},\boldsymbol{\theta} &\overset{\text{iid}}{\sim} p(z)=\sum_{h=1}^{m} w_h f_{\theta_h}(z), && i=1,\dots,n \end{aligned} \tag{2}$$

where the prior for $\boldsymbol{w}=(w_1,\dots,w_m)$, $\boldsymbol{\theta}=(\theta_1,\dots,\theta_m)$, Λ and the σ_j^2's will be specified later. Here f_{θ_h} is the d-dimensional Gaussian density. Note that, following [4], we assume a mixture model of the latent scores η_i's, instead of modeling the data y_i's themselves from the mixture. Introducing a set of latent cluster indicator variables c_i, $i=1,\dots,n$, such that $P(c_i=h \mid \boldsymbol{w})=w_h$, we can equivalently state the prior for the η_i's in (2) as

$$\eta_i \mid c_i=h,\boldsymbol{\theta} \overset{\text{ind}}{\sim} \mathcal{N}_d(\mu_h,\Delta_h), \qquad i=1,\dots,n \tag{3}$$

where $\theta_h=(\mu_h,\Delta_h)$, being μ_h the mean parameter and Δ_h the (positive definite) covariance matrix. Therefore, a cluster model is induced among the y_i's through the latent variables η_i's. In particular, y_i and y_j belong to the same clusters if η_i and η_j do, that is if $c_i=c_j$.

In general, Bayesian mixture models assume that the cluster specific parameters θ_i's are iid from some fixed distribution P_0. Here instead, in order to obtain more accurate estimate of the number of clusters, we assume a prior that encourages clusters to be well separated, with repulsion between the locations of the mixture, thereby obtaining well separated components. Since equations (2)-(3) imply

$$\{y_i : c_i=h\} \mid \boldsymbol{c},\boldsymbol{\mu},\boldsymbol{\Delta},\Lambda \overset{\text{iid}}{\sim} \mathcal{N}_p(\Lambda\mu_h,\Lambda\Delta_h\Lambda^\top), \tag{4}$$

it is clear from (4) that we should encourage a priori the distance between any couple $(\Lambda\mu_h,\Lambda\mu_l)$ to be large in order to get well separated clusters of datapoints. In our working paper [6], we define an anisotropic *determinantal point process* (DPP),

which is able to induce repulsion between the $\Lambda\mu_h$'s. Note that assuming such a prior yields that the number m of components in the mixture is random. For a thorough review of determinantal point processes, see [7]. Summing up, the prior for our likelihood (4) is given by

$$\{\mu_1,\dots,\mu_m\} \mid \Lambda \sim \text{DPP}(\rho,\Lambda,K_0) \tag{5}$$

$$w_1,\dots,w_m \mid m \sim \text{Dirichlet}(\alpha,\dots,\alpha) \tag{6}$$

$$\Delta_1,\dots,\Delta_m \mid m \overset{\text{iid}}{\sim} \text{inv-Wishart}(\nu_0,\Psi_0) \tag{7}$$

$$\sigma_1^2,\dots,\sigma_p^2 \overset{\text{iid}}{\sim} \text{inv-Gamma}(a_\sigma,b_\sigma), \tag{8}$$

with a Dirichlet-Laplace prior for Λ (as in [4]). Note that ρ is a positive parameter expressing the degree of repulsiveness, while K_0 is a covariance kernel, $\alpha,\nu_0,a_\sigma,b_\sigma>0$ and Ψ_0 represent the mean matrix in the inverse-Wishart distribution.

We propose a Gibbs sampler (with Metropolis-Hasting steps when needed) for this model. Some of the full-conditionals are standard in this type of research, but there are challenging steps, as, for instance, the sampling of the cluster-specific parameters μ_h's and Δ_h's and the sampling of the Λ matrix of parameters. Moreover, the DPP density has an infinite series representation that can be truncated as in [2] and [1].

In the next section we show some preliminary output from our algorithm tested on a simulated dataset.

3 A simulated example

We simulate $n=200$ datapoints with $p=50$ from the likelihood (1) where $\eta_i\in\mathbb{R}^d$, $d=2$ are generated as iid from $p(\cdot)=\sum_{h=1}^4 w_h f_{\theta_h}\mathcal{N}_d(\cdot;\mu_h^{true},I_2)$; here we assume

$$\mu_1^{true}=(7.5,7.5)^\top\ \mu_2^{true}=(2.5,2.5)^\top\ \mu_3^{true}=(-2.5,-2.5)^\top\ \mu_4^{true}=(-7.5,-7.5)^\top.$$

We also fix Λ^{true}, the 50×2 matrix of factor loadings, and simulate data in each of the four cluster from a 50-dimensional t-distribution with location parameters $\Lambda^{true}\eta_i^{(h)}$, covariance matrix $1.5I_{50}$ and 3 degree of freedom, for $h=1,2,3,4$ and $i=1,\dots,50$.

Figure 1 shows the posterior distributions of the number of clusters k, for our model, under three different settings of hyperparameters of the DPP that corresponds to the prior mean of m in (5) (or (6) or (7)) equal to 2.55 (S_3), 6.37 (S_4) and 25.46 (S_5). The posterior mode shown in Figure 1 recovers the true value of the number of clusters in each setting. In the talk, we will see a comparison with the model in [4].

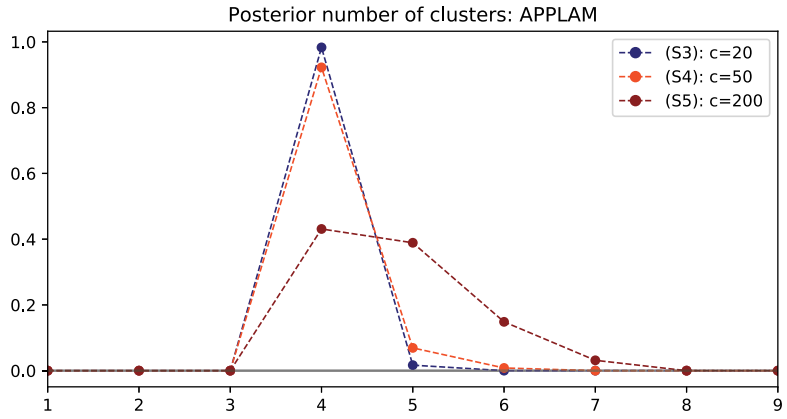


Fig. 1: Posterior distributions of the number of clusters

References

1. M. Beraha, R. Argiento, J. Møller, and A. Guglielmi. Mcmc computations for bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, pages 1–14, 2022.
2. I. Bianchini, A. Guglielmi, and F. A. Quintana. Determinantal point process mixtures via spectral density approach. *Bayesian Analysis*, 15(1):187–214, 2020.
3. D. Cai, T. Campbell, and T. Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169. PMLR, 2021.
4. N. K. Chandra, A. Canale, and D. B. Dunson. Escaping the curse of dimensionality in bayesian model based clustering. *arXiv preprint arXiv:2006.02700*, 2020.
5. S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of mixture analysis*. CRC press, 2019.
6. L. Ghilotti, M. Beraha, and A. Guglielmi. Bayesian clustering of high-dimensional data via latent repulsive mixtures, 2022.
7. F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:853–877, 2015.

Bayesian nonparametric mixtures of directed acyclic graph models

Misture bayesiane non parametriche di modelli grafici orientati

Federico Castelletti and Guido Consonni

Abstract Estimating dependence relations among variables is a pervasive issue in multivariate statistical analysis. In this context, graphical models provide a useful framework, which adopts a synthetic graph-based representation to encode conditional independence statements between variables. However the network of dependencies is typically unknown and interest lies in estimating the graph structure from the data. In addition, data are often collected in heterogeneous settings, each characterized by a specific dependence structure. In this contribution we propose a Bayesian nonparametric methodology for structure learning of directed acyclic graphs which naturally accounts for possibly heterogeneous dependence relations due to latent clusters in the data.

Abstract *La stima di relazioni di dipendenza tra variabli rappresenta un importante ambito d'indagine nell'analisi statistica di problemi multivariati. In questo contesto i modelli grafici forniscono un utile strumento di analisi, che sintetizza tali relazioni di dipendenza tramite una rappresentazione del problema basata su un grafo. Poiché tipicamente la vera struttura di dipendenza è ignota, l'obiettivo principale riguarda la stima della struttura (grafo) attraverso i dati. Questi ultimi sono spesso raccolti in contesti caratterizzati da eterogeneità, la quale implica l'esistenza di gruppi di unità statistiche, ciascuno caratterizzato da specifiche relazioni di dipendenza tra variabili. In questo contributo si propone una metodologia bayesiana non parametrica per l'apprendimento di grafi aciclici orientati che contempla l'esistenza di gruppi latenti, responsabili di strutture di dipendenza eterogenee.*

Key words: Directed acyclic graph, Mixture model, Bayesian model selection

Federico Castelletti
Università Cattolica del Sacro Cuore, federico.castelletti@unicatt.it

Guido Consonni
Università Cattolica del Sacro Cuore, guido.consonni@unicatt.it

1 Introduction

Graphical models based on Directed Acyclic Graphs (DAGs) provide an effective framework for the statistical analysis of complex dependence relations among variables [2], also from a causal perspective [9]. Since the data generating model is typically unknown, the task is to learn it from the data, a problem known as structure learning.

An assumption underlying many structure learning methods is that the available observations are collected from a homogeneous population, implying that all individuals share the same graphical model. When distinct groups are known beforehand, hierarchical methods based on multiple graphs can be implemented; see for instance [8] and [1] for a frequentist and Bayesian approach to multiple directed graphical models respectively. However, in several applied domains and especially genomics, groups are not given in advance, although it is expected that the same phenotype manifests through different *subtypes* each characterized by specific aberrations at gene level. In addition and importantly, discovering disease subtypes is of interest in itself for the development of more precise and targeted therapies.

We tackle this problem through a Bayesian non-parametric mixture of Gaussian DAG models. Our model formulation is based on a Dirichlet Process (DP) prior with base distribution given by a joint prior over the space of DAGs and DAG parameters. Accordingly, the resulting framework allows to identify clusters of units in the data, as well as dependence relations, represented by DAG structures and allied parameters, at subject-specific level.

2 Bayesian model formulation

In this section we summarize our Bayesian model. This is based on a mixture model, where each mixture component corresponds to a Gaussian DAG model. Accordingly, we first introduce Gaussian DAG models in terms of sampling distribution and priors for model parameters (Section 2.1) and then extend our framework to a Dirichlet Process mixture of Gaussian DAGs (Section 2.2).

2.1 Gaussian DAG models

Let $\mathscr{D} = (V, E)$ be a DAG, with set of nodes $V = \{1, \ldots, q\}$ and set of edges $E \subseteq V \times V$. Let also $(X_1, \ldots, X_q)$ be a collection of random variables, each associated with a node in $\mathscr{D}$. We assume that the joint distribution of the q variables is multivariate Normal and *Markov* w.r.t DAG $\mathscr{D}$, meaning that the conditional independencies implied by $\mathscr{D}$ are encoded in the sampling distribution. Specifically,

$$X_1 \ldots, X_q \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega} \sim \mathscr{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}), \quad \boldsymbol{\mu} \in \mathbb{R}^q, \boldsymbol{\Omega} \in \mathscr{P}_{\mathscr{D}}, \tag{1}$$

where $\mathscr{P}_{\mathscr{D}}$ is the set of all symmetric positive definite (s.p.d.) precision matrices Markov w.r.t. $\mathscr{D}$. Notice that the elements of $\boldsymbol{\Omega}$ must satisfy constraints beyond those required for the matrix to be s.p.d. Equation (1) can be equivalently written as a Structural Equation Model (SEM) of the form

$$\boldsymbol{\eta} + \boldsymbol{L}^{\top}(X_1,\ldots,X_q)^{\top} = \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon} \sim \mathscr{N}_q(\mathbf{0},\boldsymbol{D})$. In particular, $\boldsymbol{L}$ is a (q,q) matrix of regression coefficients such that for $u \neq v$ $\boldsymbol{L}_{u,v} \neq 0$ if and only if $(u,v) \in E$, while $\boldsymbol{L}_{u,u} = 1$ for each $u = 1,\ldots,q$. Moreover, $\boldsymbol{D}$ is a (q,q) diagonal matrix whose (j,j)-element $\boldsymbol{D}_{jj}$ corresponds to the conditional variance of X_j, $\mathbb{V}\mathrm{ar}(X_j \,|\, \boldsymbol{x}_{\mathrm{pa}_{\mathscr{D}}(j)})$, where $\mathrm{pa}_{\mathscr{D}}(j)$ are the parents of node j in $\mathscr{D}$. Also, $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_q)^{\top}$ is an intercept term. Given $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, a SEM implies the following re-parameterization

$$\boldsymbol{L}_{\prec j]} = \boldsymbol{\Sigma}^{-1}_{\prec j \succ}\boldsymbol{\Sigma}_{\prec j]}, \quad \boldsymbol{D}_{jj} = \boldsymbol{\Sigma}_{jj\,|\,\mathrm{pa}_{\mathscr{D}}(j)}, \quad \eta_j = \mu_j + \boldsymbol{L}^{\top}_{\prec j]}\boldsymbol{\mu}_{\mathrm{pa}_{\mathscr{D}}(j)}, \tag{3}$$

for $j = 1,\ldots,q$, where $\boldsymbol{\Sigma}_{jj|\mathrm{pa}_{\mathscr{D}}(j)} = \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{[j\succ}\boldsymbol{\Sigma}^{-1}_{\prec j \succ}\boldsymbol{\Sigma}_{\prec j]}$, $\prec j] = \mathrm{pa}_{\mathscr{D}}(j) \times j$, $[j \succ = j \times \mathrm{pa}_{\mathscr{D}}(j)$, $\prec j \succ = \mathrm{pa}_{\mathscr{D}}(j) \times \mathrm{pa}_{\mathscr{D}}(j)$. Accordingly, the joint density of $(X_1,\ldots,X_q)$ can be equivalently written as

$$f(x_1,\ldots,x_q \,|\, \boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L},\mathscr{D}) = \prod_{j=1}^{q} d\mathscr{N}(x_j \,|\, \eta_j - \boldsymbol{L}^{\top}_{\prec j]}\boldsymbol{x}_{\mathrm{pa}_{\mathscr{D}}(j)}, \boldsymbol{D}_{jj}), \tag{4}$$

where $d\mathscr{N}(x|\,\mu,\sigma^2)$ denotes the Normal density of $\mathscr{N}(\mu,\sigma^2)$ and we also emphasize the dependence on DAG $\mathscr{D}$. Expression (4) is an instance of the usual DAG factorization.

We complete our model formulation by assigning a prior to DAG $\mathscr{D}$ and parameters $(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L})$. Since $(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L})$ are specific to each DAG model under consideration, we structure our prior as

$$p(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L},\mathscr{D}) = p(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L}\,|\,\mathscr{D})p(\mathscr{D}). \tag{5}$$

Let $\mathscr{S}_q$ be the space of all DAGs on q nodes. We assign a prior to $\mathscr{D} \in \mathscr{S}_q$ through independent Bernoulli distributions $\mathrm{Ber}(\pi)$ on the collection of $q(q-1)/2$ 0-1 elements, each indicating the absence/presence of a link between two distinct nodes in the DAG. In addition, to account for multiplicity correction [12] we assign hierarchically a Beta prior on π. The resulting prior can be written as

$$\begin{aligned} p(\boldsymbol{S}^{\mathscr{D}} \,|\, \pi) &= \pi^{|\boldsymbol{S}^{\mathscr{D}}|}(1-\pi)^{\frac{q(q-1)}{2}-|\boldsymbol{S}^{\mathscr{D}}|} \\ \pi &\sim \mathrm{Beta}(a,b), \end{aligned} \tag{6}$$

where $\boldsymbol{S}^{\mathscr{D}}$ is the (q,q) adjacency matrix of the skeleton (underlying undirected graph) of DAG $\mathscr{D}$, and $|\boldsymbol{S}^{\mathscr{D}}|$ is the number of edges in $\mathscr{D}$. By integrating w.r.t. π we then obtain

$$p(\boldsymbol{S}^{\mathscr{D}}) = \frac{\Gamma\left(|\boldsymbol{S}^{\mathscr{D}}|+a\right)\Gamma\left(\frac{q(q-1)}{2}-|\boldsymbol{S}^{\mathscr{D}}|+b\right)}{\Gamma\left(\frac{q(q-1)}{2}+a+b\right)}\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}. \tag{7}$$

Finally we take $p(\mathscr{D}) \propto p(\boldsymbol{S}^{\mathscr{D}})$.

Consider now the prior $p(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L}\,|\,\mathscr{D})$, where parameters $(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L})$ correspond to the re-parameterization of $(\boldsymbol{\mu},\boldsymbol{\Omega})$ in (3), and $\boldsymbol{\Omega}$ is Markov w.r.t. DAG $\mathscr{D}$. To assign $p(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L}\,|\,\mathscr{D})$, one can follow the elicitation procedure introduced by [4]. In particular, starting from a Normal-Wishart prior on the parameters of a *complete* DAG model, $\mathscr{N}_q(\boldsymbol{\mu},\boldsymbol{\Omega}^{-1})$, with $\boldsymbol{\Omega} \in \mathscr{P}$ *unconstrained*, the prior under any other (incomplete) DAG can be derived automatically starting from suitable assumptions. Specifically, one can show that $(\boldsymbol{\mu},\boldsymbol{\Omega}) \sim \mathscr{NW}(a_\mu,\boldsymbol{m},a_\Omega,\boldsymbol{U})$, a Normal-Wishart distribution, induces a prior of the form

$$\begin{aligned} p(\boldsymbol{\eta},\boldsymbol{D},\boldsymbol{L}\,|\,\mathscr{D}) &= \prod_{j=1}^{q} p(\eta_j,\boldsymbol{L}_{\prec j]},\boldsymbol{D}_{jj}) \\ &= \prod_{j=1}^{q} p(\eta_j\,|\,\boldsymbol{L}_{\prec j]},\boldsymbol{D}_{jj})p(\boldsymbol{L}_{\prec j]}\,|\,\boldsymbol{D}_{jj})p(\boldsymbol{D}_{jj}) \end{aligned} \tag{8}$$

where in particular

$$\begin{aligned} \boldsymbol{D}_{jj} &\sim \text{I-Ga}\left(\frac{1}{2}a_j^{\mathscr{D}},\frac{1}{2}\boldsymbol{U}_{jj|\text{pa}_{\mathscr{D}}(j)}\right), \\ \boldsymbol{L}_{\prec j]}\,|\,\boldsymbol{D}_{jj} &\sim \mathscr{N}_{|\text{pa}_{\mathscr{D}}(j)|}\left(-\boldsymbol{U}^{-1}_{\prec j\succ}\boldsymbol{U}_{\prec j]},\boldsymbol{D}_{jj}\boldsymbol{U}^{-1}_{\prec j\succ}\right), \\ \eta_j\,|\,\boldsymbol{L}_{\prec j]},\boldsymbol{D}_{jj} &\sim \mathscr{N}\left(m_j+\boldsymbol{L}^{\top}_{\prec j]}\boldsymbol{m}_{\text{pa}_{\mathscr{D}}(j)},\boldsymbol{D}_{jj}/a_\mu\right), \end{aligned} \tag{9}$$

with I-Ga denoting an inverse-gamma distribution and $a_j^{\mathscr{D}} = a_\Omega + |\text{pa}_{\mathscr{D}}(j)| - q + 1$. Most importantly, the resulting prior is conjugate to the sampling density in (4). Hence, for given n i.i.d. observations $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ from the Gaussian DAG model (1), the marginal likelihood $p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\,|\,\mathscr{D})$ is analytically available, a feature which dramatically reduces the computational burden of the sampling scheme summarized in Section 3.

2.2 Mixture of Gaussian DAG models

In this section we introduce our mixture model. This is based on a *Dirichlet Process* (DP) prior [7], which we can write using the following hierarchical structure

$$\begin{aligned} \boldsymbol{x}_i \mid \boldsymbol{\theta}_i &\sim p\big(\boldsymbol{x}_i \mid \boldsymbol{\theta}_i\big), \\ \boldsymbol{\theta}_i \mid M &\sim M, \\ M &\sim DP(M_0, \alpha), \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}_i = (\boldsymbol{\eta}_i, \boldsymbol{D}_i, \boldsymbol{L}_i, \mathscr{D}_i)$ and $DP(M_0, \alpha)$ represents the Dirichlet Process with base distribution M_0 and concentration parameter α. In particular, we take $p(\boldsymbol{\eta}, \boldsymbol{D}, \boldsymbol{L}, \mathscr{D})$ (5) as the base distribution M_0.

Let now $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,q})^\top$, $i = 1, \ldots, n$, be n independent draws from (10). In a DP mixture each sample $\boldsymbol{x}_i$ has potentially a distinct parameter $\boldsymbol{\theta}_i$. If we let $K \leq n$ be the number of *unique* values among $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ and $\xi_1, \ldots, \xi_n$, $\xi_i \in \{1, \ldots, K\}$, a sequence of indicator variables such that $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_{\xi_i}$, we can equivalently write model (10) in terms of the random partition induced by the $\{\xi_i\}$'s,

$$f(\boldsymbol{X} \mid \xi_1, \ldots, \xi_n, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \prod_{k=1}^{K} \left\{ \prod_{i:\xi_i=k} f(\boldsymbol{x}_i \mid \boldsymbol{\eta}^*_k, \boldsymbol{D}^*_k, \boldsymbol{L}^*_k, \mathscr{D}^*_k) \right\}. \tag{11}$$

The previous expression decomposes the likelihood into K homogeneous groups, with observations within each group corresponding to i.i.d. draws from a Gaussian DAG model of the form (4).

3 Posterior inference and illustrations

Posterior inference for model (10) can be performed by resorting to a *slice sampler* [13] which maintains the structure of a blocked Gibbs sampler. This is based on two main steps.

Conditionally on a partition of the n individuals into K clusters, each cluster-specific parameter $\boldsymbol{\theta}^{(k)} = \big(\boldsymbol{\eta}^{(k)}, \boldsymbol{D}^{(k)}, \boldsymbol{L}^{(k)}, \mathscr{D}^{(k)}\big)$, $k = 1, \ldots, K$, is updated through an MCMC scheme based on a Partial Analytic Structure (PAS) algorithm [5]. The latter proceeds by i) updating DAG $\mathscr{D}^{(k)}$ using a Metropolis Hastings (MH) step where a new candidate DAG is drawn from a suitable proposal distribution and accepted with probability given by the MH acceptance ratio; ii) sampling the DAG-dependent parameters $\big(\boldsymbol{\eta}^{(k)}, \boldsymbol{D}^{(k)}, \boldsymbol{L}^{(k)}\big)$ conditionally on the accepted DAG from their full conditional distribution. The latter corresponds to a Normal-DAG-Wishart distribution because of conjugacy of the prior (9) with the Gaussian DAG model (4).

In the second step, cluster indicators $\xi_1, \ldots, \xi_n$ are then updated from their full conditional distribution, augmented by auxiliary variables representing the weights of the mixture model; we refer the reader to [6] for full details.

We illustrate the proposed methodology on a simple simulated dataset with number of clusters $K = 2$ and $q = 10$ variables. For each cluster $k = 1, 2$, we randomly generate a *sparse* DAG structure by fixing a probability of edge inclusion 0.2, and the corresponding parameters as

$$\eta_j^{(k)} \sim \text{Unif}(-\delta,\delta), \quad \boldsymbol{L}_{l,j}^{(k)} \sim \text{Unif}(-1,1), \quad \boldsymbol{D}_{jj}^{(k)} = 1, \tag{12}$$

independently across $j = 1,\ldots,q$, $l \in \text{pa}_{\mathscr{D}}(j)$, $k \in \{1,2\}$ and varying $\delta \in \{1,2,5\}$. The latter choice identifies three distinct scenarios characterized by different levels of cluster separation, due to the expected (increasing) difference in mean across variables and among groups. For each $k = 1,2$, $n^{(k)} = 50$ i.i.d. observations are finally generated following (4). The two DAG structures are represented in Figure 1.

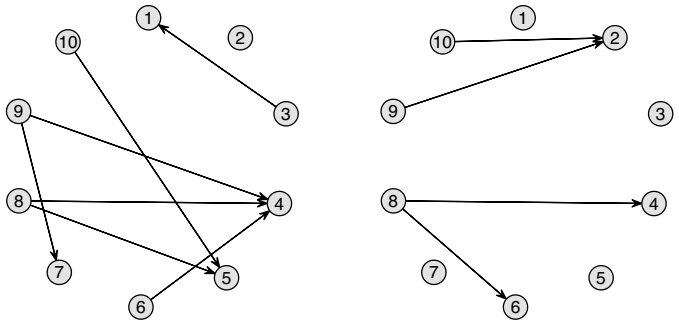


Fig. 1: Simulation study. True randomly generated DAG structures, $\mathscr{D}_1$, $\mathscr{D}_2$.

The output of our MCMC scheme is a collection of S draws approximately sampled from the posterior of $(\boldsymbol{\xi},\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^*$ corresponding to $(\boldsymbol{\eta}^*,\boldsymbol{D}^*,\boldsymbol{L}^*,\mathscr{D}^*)$. More specifically, for each MCMC iteration $t = 1,\ldots,S$, our algorithm returns an n-dimensional vector of cluster indicators $\boldsymbol{\xi}^{(t)} = \left(\xi_1^{(t)},\ldots,\xi_n^{(t)}\right)$, with $\xi_i^{(t)} \in \left\{1,\ldots,K^{(t)}\right\}$ where $K^{(t)}$ is the number of distinct clusters, together with a collection of $K^{(t)}$ distinct cluster-specific parameters $\left\{\boldsymbol{\theta}_1^{(t)},\ldots,\boldsymbol{\theta}_{K^{(t)}}^{(t)}\right\}$.

We first construct an (n,n) similarity matrix $\boldsymbol{S}$, whose (i,i')-element corresponds to the posterior probability that individuals i and i' are assigned to the same cluster. Results, obtained under each scenario defined by $\delta \in \{1,2,5\}$, are reported in Figure 2, where subjects labelled as 1:50 and 51:100 belong to (true) clusters 1 and 2 respectively. By visual inspection, it appears that recovery of the true clustering structure improves as the degree of separation between groups due to $\delta \in \{1,2,5\}$ grows.

We now focus on graph learning. Starting from the MCMC output, we can compute for each subject $i = 1,\ldots,n$ a (q,q) matrix $\hat{\mathbf{P}}_i$ collecting *subject-specific* posterior probabilities of edge inclusion. Specifically, the (u,v)-element of $\hat{\mathbf{P}}_i$ is

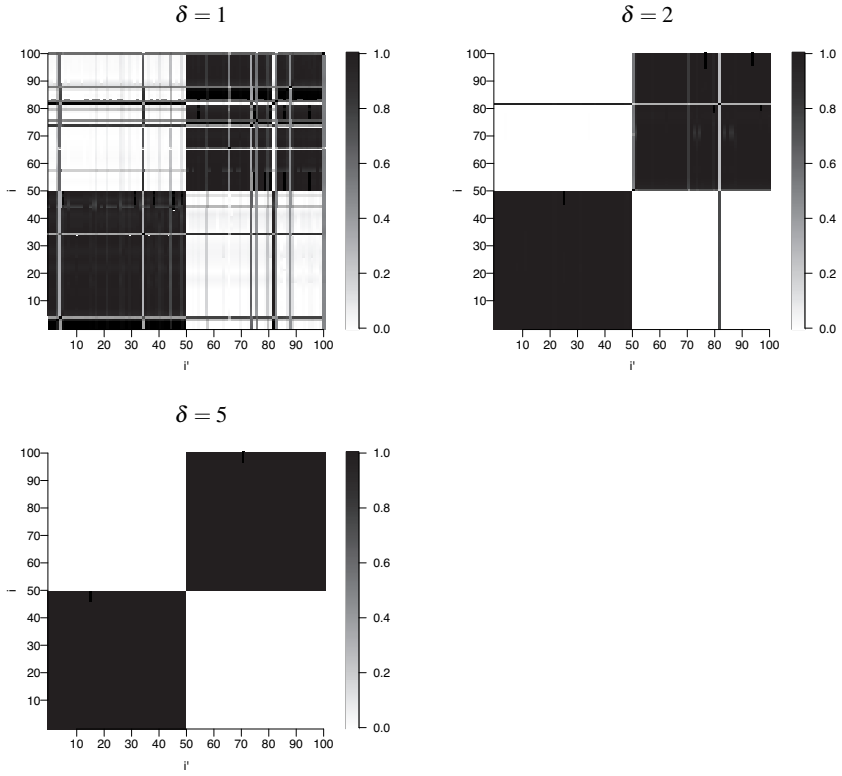


Fig. 2: Simulation study. Posterior similarity matrix with subjects arranged by true cluster membership (1:50 from cluster 1, 51:100 from cluster 2) for values of $\delta \in \{1,2,5\}$.

$$\hat{p}_i(u \to v \,|\, \boldsymbol{X}) = \frac{1}{S}\sum_{t=1}^{S} \mathbf{1}_{u\to v}\left\{\mathscr{D}_{\xi_i}^{(t)}\right\} \tag{13}$$

where $\mathbf{1}_{u\to v}\{\mathscr{D}\} = 1$ if DAG $\mathscr{D}$ contains the edge (u,v), and zero otherwise. Posterior probabilities of edge inclusion for two subjects $i \in \{10,60\}$ belonging to (true) clusters 1 and 2 respectively are reported as heatmaps in Figure 3. The output refers to the scenario defined by $\delta = 2$, where the two individuals were assigned to distinct clusters with probability one; see also Figure 2.

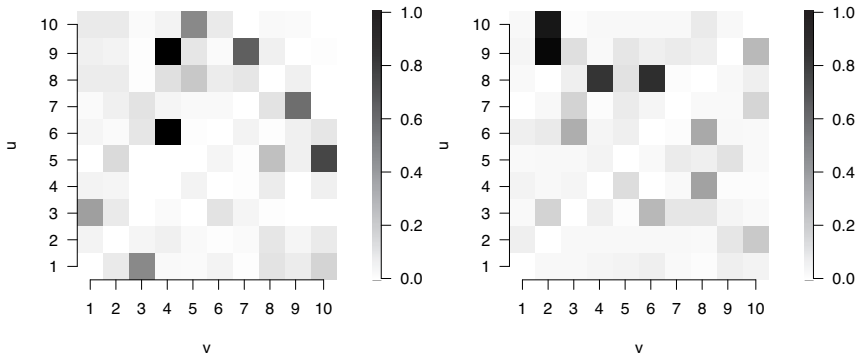


Fig. 3: Simulation study. Posterior probability of edge inclusion for each directed edge (u,v), for subjects $i \in \{10,60\}$ whose true cluster memberships are $\xi_{10} = 1$ and $\xi_{60} = 2$.

References

1. Castelletti, F., La Rocca, L., Peluso, S., Stingo, F. C., Consonni, G.: Bayesian learning of multiple directed networks from observational data. Statistics in Medicine **39**, 4745–4766 (2020)
2. Cowell, R. G., Dawid, P. A., Lauritzen, S. L., Spiegelhalter, D. J.: Probabilistic Networks and Expert Systems. New York: Springer (1999)
3. Escobar, M. D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association **90**, 577–588 (1995)
4. Geiger, D., Heckerman, D.: Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics **30**, 1412–1440 (2002)
5. Godsill, S. J.: On the relationship between Markov chain Monte Carlo methods for model uncertainty. Journal of Computational and Graphical Statistics **10**, 230–248 (2012)
6. Kalli, M., Griffin, J., Walker, S.: Slice sampling mixture models. Statistics and Computing **21**, 93–105 (2011)
7. Müller, P., Mitra, R.: Bayesian nonparametric inference: why and how. Bayesian Analysis **8**, 269–302 (2013)
8. Oates, C. J., Smith, J. Q., Mukherjee, S., Cussens, J.: Exact estimation of multiple directed acyclic graphs. Statistics and Computing **26**, 797–811 (2016)
9. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
10. Peterson, C., Stingo, F. C., Vannucci, M.: Bayesian inference of multiple Gaussian graphical models. Journal of the American Statistical Association **110**, 159–174 (2015)
11. Rodriguez, A., Lenkoski, A., Dobra, A.: Sparse covariance estimation in heterogeneous samples. Electronic Journal of Statistics **5**, 981–1014 (2011)
12. Scott, J. G., Berger, J. O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics, **38**, 2587–2619 (2010)
13. Walker, S. G.: Sampling the Dirichlet mixture model with slices. Communications in Statistics - Simulation and Computation **36**, 45–54 (2007)

Bayesian Clustering of Brain Regions via Extended Stochastic Block Models

Clustering Bayesiano delle Regioni Cerebrali attraverso Modelli a Blocchi Stocastici Estesi

Sirio Legramanti, Tommaso Rigon, and Daniele Durante

Abstract Brain networks typically exhibit clusters of nodes with similar connectivity patterns. Moreover, for each node (brain region), attributes are available in the form of hemisphere and lobe memberships. Clustering brain regions based on their connectivity patterns and their attributes is then of substantial statistical interest when analyzing brain networks. However, the algorithms available for this task lack uncertainty quantification. Even traditional model-based solutions present some critical issues, namely in the specification of the number of clusters and the incorporation of node attributes. Hence, to analyze the considered brain network, we opt for the extended stochastic block model by Legramanti et al. (2022b), which allows to infer the number of clusters and to incorporate node attributes.

Abstract *Le reti cerebrali mostrano tipicamente gruppi di nodi con connettività simili. Inoltre per ciascun nodo (regione cerebrale) sono disponibili come attributi emisfero e lobo di appartenenza. Il clustering delle regioni cerebrali sulla base delle loro connessioni e attributi è quindi di grande interesse nell'analisi statistica delle reti cerebrali. Tuttavia gli algoritmi per questo compito mancano di quantificazione dell'incertezza. Anche le tradizionali soluzioni modellistiche presentano delle criticità, in particolare nello specificare il numero di gruppi e nell'incorporare gli attributi dei nodi. Per analizzare la rete cerebrale considerata, optiamo quindi per il modello a blocchi stocastici esteso di Legramanti et al. (2022b), che permette di inferire il numero di gruppi e di incorporare gli attributi dei nodi.*

Key words: Bayesian nonparametrics, Gibbs-type prior, Network

---

Sirio Legramanti
Università degli Studi di Bergamo, via dei Caniana 2, 24127 Bergamo (Italia), e-mail: sirio.legramanti@unibg.it

Tommaso Rigon
Università degli Studi di Milano-Bicocca, piazza dell'Ateneo Nuovo 1, 20126 Milano (Italia), e-mail: tommaso.rigon@unimib.it

Daniele Durante
Università Bocconi, via Roentgen 1, 20136 Milano (Italia), e-mail: daniele.durante@unibocconi.it

1 Introduction

As is often the case with network data, brain networks typically exhibit clusters of nodes sharing similar connectivity patterns. Moreover, for each node (brain region) attributes are available in the form of hemisphere and lobe memberships. Clustering brain regions based on their connectivity patterns and their attributes is then of major statistical interest when analyzing brain networks. However, the algorithms available for this task (Blondel et al., 2008; Girvan and Newman, 2002; Newman and Girvan, 2004; Newman, 2006) lack uncertainty quantification and focus on partitions characterized by dense within-cluster connectivity and sparser connections between different clusters. Also traditional model-based solutions such as spectral clustering (Von Luxburg, 2007) and stochastic block model (SBM) (Holland et al., 1983; Nowicki and Snijders, 2001) present some critical issues, namely in the specification of the number of clusters and the incorporation of node attributes.

The extended stochastic block model (ESBM) proposed by Legramanti et al. (2022b) addresses these issues through a model-based framework that: (i) quantifies uncertainty in the inferred clustering through a Bayesian approach; (ii) allows the number of clusters to be fixed or random, and asymptotically finite or infinite, depending on the application; (iii) facilitates the incorporation of node attributes, favoring clusters that are homogeneous with respect to such attributes.

A peculiarity of brain networks is that a growth in the number of nodes does not represent the addition of new entities (like, e.g., in social networks), but just a more refined subdivision in brain regions. Hence, we cannot assume the number of clusters to grow indefinitely with the number of nodes. On the contrary, it is more reasonable to assume that the number of clusters remains finite even on a more refined division into regions. Moreover, the number of clusters is typically unknown. These observations rule out both a traditional SBM based on a Dirichlet-multinomial with a fixed pre-specified number of clusters (Holland et al., 1983; Nowicki and Snijders, 2001) and an infinite relational model (Kemp et al., 2006) based on a Chinese-restaurant process that would yield infinitely many clusters for infinitely many nodes.

To analyze the motivating brain network described in Section 2, we then opt for a Gnedin-process specification of the ESBM, thus allowing for a random but finite number of clusters. Such a behavior is shared with the mixture-of-finite-mixtures solution proposed by Geng et al. (2019), which however is based on a different prior for the number of clusters and does not provide a solution for the inclusion of node attributes.

The rest of the paper is organized as follows: in Section 2 we describe the considered brain network data; in Section 3 we recall the ESBM framework; finally, in Section 4 we analyze the considered brain network through ESBM.

2 Data

We consider the data provided by Sulaimany et al. (2017), focusing on the matrix representing the brain network of healthy subjects. In this network, nodes represent the 68 anatomical regions in the Desikan atlas (Desikan et al., 2006), while edges encode the presence of white matter fibers among such regions. The corresponding matrix is then binary and symmetric. Self-loops are not considered, as not informative for our clustering goal. Each node (brain region) comes with its hemisphere and lobe memberships as additional attributes. The two hemispheres (left, right) and six lobes (frontal, insular, limbic, occipital, parietal, temporal) result in 12 hemisphere-lobe combinations that will be employed as node attributes.

Figure 1 provides a graphical representation of the considered brain network, with node positions obtained via force-directed placement (Fruchterman and Reingold, 1991). The fact that nodes in the same hemisphere and/or lobe are placed close to each other suggests that the hemisphere-lobe combination may be informative for node clustering, even if not sufficient to fully characterize the network block structures; in this regard, see also Legramanti et al. (2022a).

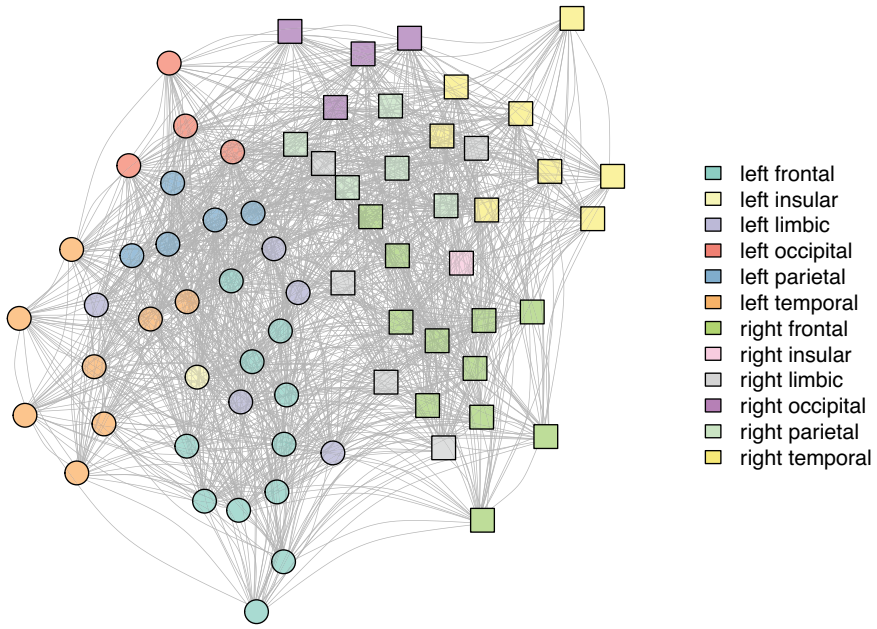


Fig. 1 Graphical representation of the considered network of brain regions. Node positions are obtained via force-directed placement (Fruchterman and Reingold, 1991), node shapes denote hemispheres (round=left, square=right), and colors correspond to the 12 hemisphere-lobe combinations.

3 Extended Stochastic Block Models

In this section, we briefly recall the ESBM framework proposed by Legramanti et al. (2022b) and employed here to analyze the brain network described in Section 2. For further details on ESBMs please refer to Legramanti et al. (2022b).

Let us denote with V the number of nodes in the considered binary undirected network, and let $\mathbf{Y}$ be its $V \times V$ symmetric adjacency matrix, with elements $y_{vu} = y_{uv} = 1$ if nodes v and u are connected, and $y_{vu} = y_{uv} = 0$ otherwise. SBMs (Holland et al., 1983; Nowicki and Snijders, 2001) partition the nodes into H disjoint clusters, with nodes in the same cluster sharing a common connectivity pattern.

In particular, for a binary undirected network without self-loops like our motivating brain network, SBMs assume that the sub-diagonal entries y_{vu} $(v = 2,\ldots,V; u = 1,\ldots,v-1)$ of the symmetric adjacency matrix $\mathbf{Y}$ are conditionally independent Bernoulli random variables with probabilities $\theta_{z_v,z_u} \in (0,1)$ depending only on the cluster memberships z_v and z_u of the two involved nodes.

Let $\mathbf{z} = (z_1,\ldots,z_V)^\intercal \in \{1,\ldots,H\}^V$ be the vector of node memberships associated to the node partition $\{Z_1,\ldots,Z_H\}$, so that $z_v = h$ if and only if $v \in Z_h$. Moreover, denote with Θ the $H \times H$ symmetric matrix whose generic element $\theta_{hk} \in (0,1)$ is the probability of an edge between a node in cluster h and a node in cluster k. Then, the likelihood for $\mathbf{Y}$ given $\mathbf{z}$ and Θ is $p(\mathbf{Y} \mid \mathbf{z}, \Theta) = \prod_{h=1}^{H} \prod_{k=1}^{h} \theta_{hk}^{m_{hk}} (1-\theta_{hk})^{\overline{m}_{hk}}$, where m_{hk} and $\overline{m}_{hk}$ denote the number of edges and non-edges between nodes in clusters h and k, respectively.

However, the block probabilities Θ are not of direct interest in our application. Hence, we follow the common practice of treating Θ as a nuisance parameter. We then assign independent Beta(a,b) priors to the block probabilities θ_{hk}, and marginalize them out in $p(\mathbf{Y} \mid \mathbf{z}, \Theta)$, thus obtaining

$$p(\mathbf{Y} \mid \mathbf{z}) = \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{\mathrm{B}(a+m_{hk}, b+\overline{m}_{hk})}{\mathrm{B}(a,b)}. \tag{1}$$

The likelihood in (1) is common to several SBM formulations, which differ in the choice of a prior for $\mathbf{z}$. Several options have been considered in the context of SBMs as priors for $\mathbf{z}$, including the Dirichlet-multinomial (Nowicki and Snijders, 2001), the Dirichlet process (Kemp et al., 2006), and mixtures of finite Dirichlet mixtures (Geng et al., 2019).

Notably, these are all examples of Gibbs-type priors (e.g., De Blasi et al., 2013). This motivates the unifying ESBM framework in Legramanti et al. (2022b), which assumes a generic Gibbs-type prior for $\mathbf{z}$. Within the Gibbs-type family, besides the options listed above, Legramanti et al. (2022b) explored the use of the Gnedin process for SBMs. The Gnedin process (Gnedin, 2010) depends on a single parameter $\gamma \in (0,1)$, and yields the following urn scheme

$$\mathrm{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} (n_h+1)(V-H+\gamma) & \text{for } h = 1,\ldots,H, \\ H^2 - H\gamma & \text{for } h = H+1, \end{cases} \tag{2}$$

where n_h is the number of nodes in cluster h.

The Gnedin process is particularly suited to our motivating brain network application, since it assumes that the number of clusters is random and finite, even for infinitely many nodes. This makes it preferable to the Dirichlet and Pitman-Yor processes, which instead yield infinitely many clusters for infinitely many nodes, and to the Dirichlet-multinomial, which needs a pre-specified fixed number of clusters. In contrast to the Dirichlet-multinomial, the Gnedin process induces a prior (and allows to infer a posterior) on the number of clusters. This feature is shared with the mixed-of-finite-mixtures approach of Geng et al. (2019). However, the Gnedin process yields the simple urn scheme in (2), which facilitates posterior sampling, and induces a prior on the number of clusters with mode at 1 and heavy tails, thus favoring parsimonious representations while also allowing for richer structures.

When node attributes x_v are available for each node v, like in our brain network application, this information can support inference on block structures, reducing both posterior bias and uncertainty. The ESBM framework allows to leverage node attributes in a principled manner, by assuming a model for the attributes given cluster memberships. In the motivating brain data, each node attribute $x_v \in \{1,\dots,C\}$ is a single categorical variable denoting a hemisphere-lobe combination ($C = 12$). In this setting, following (Müller et al., 2011), Legramanti et al. (2022b) recommend a Dirichlet-multinomial model for the attributes

$$p(\mathbf{X}_h) \propto \frac{1}{\Gamma(n_h+\alpha_0)} \prod_{c=1}^{C} \Gamma(n_{hc}+\alpha_c), \tag{3}$$

where n_{hc} is the number of nodes in cluster h with attribute value c, and $\alpha_0 = \sum_{c=1}^{C} \alpha_c$, with $\alpha_c > 0$ for $c = 1,\dots,C$. Including (3) in (2) yields, in the case of a Gnedin process prior for $\mathbf{z}$, to the following supervised urn scheme

$$\text{pr}(z_{V+1}{=}h|\mathbf{X}, x_{V+1}, \mathbf{z}) \propto \begin{cases} \frac{n_{hx_{V+1}}+\alpha_{x_{V+1}}}{n_h+\alpha_0}(n_h+1)(V-H+\gamma) & \text{for } h{=}1,\dots,H, \\ \frac{\alpha_{x_{V+1}}}{\alpha_0}(H^2-H\gamma) & \text{for } h{=}H+1, \end{cases} \tag{4}$$

where $n_{hx_{V+1}}$ is the number of nodes in cluster h with the same covariate value $c = x_{V+1}$ as node $V+1$, whereas $\alpha_{x_{V+1}}$ is the parameter associated with the category $c = x_{V+1}$ of node $V+1$. This favors the attribution of each node to the cluster(s) containing a higher fraction of nodes with its same attribute value.

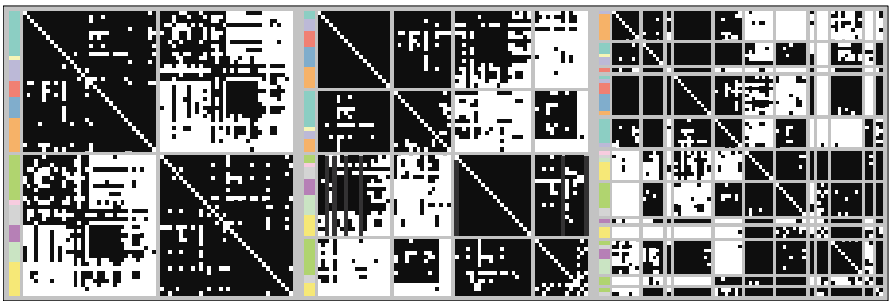
The availability of the urn schemes (2) and (4) allows to derive a collapsed Gibbs sampler. This actually holds for the whole class of Gibbs-type priors, and hence for any ESBM. See Legramanti et al. (2022b) for details on the Gibbs sampler, and https://github.com/danieledurante/ESBM for an R implementation.

4 Brain Network Analysis

We analyze the data described in Section 2 through the ESBM framework proposed in Legramanti et al. (2022b) and summarized in Section 3. Motivated by the discussion in the previous sections, we opt for a Gnedin process specification of the ESBM, supervised with the hemisphere-lobe membership of each brain region. We set the Gnedin process parameter to $\gamma = 0.3$, which corresponds to a prior expectation of approximately 17 clusters. The analyses are performed via a slight modification of the R code publicly available at https://github.com/danieledurante/ESBM.

In Figure 2, the posterior estimate obtained under this ESBM specification is visually compared to the output of the Louvain algorithm (Blondel et al., 2008), and to the estimate provided by spectral clustering (Von Luxburg, 2007). The number of clusters for the latter is obtained via a combination of the model selection procedures in the R package `randnet`.

The Louvain algorithm identifies only two clusters which correspond to hemispheres, thus providing a coarsened representation that fails to capture connectivity patterns among the two hemispheres. This tendency toward coarsening is partially replicated by spectral clustering, which identifies 4 clusters. However, Figure 2(b) clearly shows some residual structure, e.g. in the block corresponding to edges among the second and fourth clusters. In contrast, the supervised Gnedin-based ESBM exhibits an improved ability in learning the block structures within the considered brain network. It recovers 12 clusters, whose composition is partially coherent with hemisphere-lobe memberships, but also departs from them when this is required in order to capture connectivity patterns.



(a) (b) (c)

Fig. 2 Adjacency matrix of the considered brain network with rows/columns (representing nodes) ordered and partitioned according to the clustering estimated under three different methods: (a) Louvain algorithm; (b) spectral clustering; (c) ESBM with Gnedin process prior supervised with hemisphere-lobe memberships. Black and white cells represent edges and non-edges, respectively, while side colors correspond to hemisphere-lobe combinations (see the legend of Figure 1).

References

V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 10:P10008, 2008.

P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero. Are Gibbs–type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:212–229, 2013.

R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21:1129–1164, 1991.

J. Geng, A. Bhattacharya, and D. Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114:893–905, 2019.

M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.

A. Gnedin. Species sampling model with finitely many types. *Electronic Communications in Probability*, 15:79–88, 2010.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, pages 381–388, 2006.

S. Legramanti, T. Rigon, and D. Durante. Bayesian testing for exogenous partition structures in stochastic block models. *Sankhya A*, 84:108–126, 2022a.

S. Legramanti, T. Rigon, D. Durante, and D. B. Dunson. Extended stochastic block models with application to criminal networks. *Annals of Applied Statistics, in press*, 2022b.

P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103:8577–8582, 2006.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

S. Sulaimany, M. Khansari, P. Zarrineh, M. Daianu, N. Jahanshad, P. M. Thompson, and A. Masoudi-Nejad. Predicting brain network changes in Alzheimer's disease with link prediction algorithms. *Molecular BioSystems*, 13(4):725–735, 2017.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Data Science skills for next generation statisticians

Cluster based oversampling for imbalanced learning

Sovracampionamento basato sui cluster per l'apprendimento con dati sbilanciati

Gioia Di Credico and Nicola Torelli

Abstract Oversampling is a widespread remedy used when there is data imbalance in classification problems. Some oversampling techniques amount to generating new cases in the minority class which are similar to the observed ones. ROSE (Random OverSampling Examples) is an algorithm for generating new data, both in minority and majority classes, by using ideas from kernel density estimation and bootstrap resampling. In this paper, we show that a new strategy which couples density-based clustering methods with ROSE can improve the performance of supervised classification methods with data imbalance. Evidence from some simulation experiments shows that the new procedure is promising and solves some issues related to the use of ROSE.

Abstract *Il sovracampionamento è una soluzione spesso utilizzata in presenza di sbilanciamento dei dati nei problemi di classificazione. Alcune tecniche di sovracampionamento consentono di generare nuovi casi nella classe minoritaria simili a quelli osservati. ROSE (Random OverSampling Examples) è un algoritmo per generare nuovi dati, sia nella classe di minoranza che in quella di maggioranza, basato sull'idea di stima della densità col metodo del nucleo e del ricampionamento bootstrap. In questo lavoro mostriamo che una nuova strategia che abbina i metodi di clustering basati sulla densità con ROSE può migliorare le prestazioni dei metodi di classificazione supervisionati in caso di dati sbilanciati. Le prime evidenze sull'uso della procedura proposta, basate su uno studio di simulazione, sono promettenti.*

Key words: Density-based clustering, tuning parameters, resampling, ROSE, SMOTE

Gioia Di Credico
Department of Economics, Business, Mathematics and Statistics, Università di Trieste, Piazzale Europa 1, 34127 Trieste e-mail: gioia.dicredico@deams.units.it

Nicola Torelli
Department of Economics, Business, Mathematics and Statistics, Università di Trieste, Piazzale Europa 1, 34127 Trieste e-mail: nicola.torelli@deams.units.it

1 Introduction

When dealing with imbalanced classification problems undersampling and oversampling are the solutions proposed more often and successfully applied proved successful in building a supervised learning algorithm. ROSE (Random OverSampling Examples) [9] is a procedure for generating synthetic samples by using ideas from kernel density estimation and bootstrap resampling. It has proved to be a sound alternative to other methods for generating synthetic data such as those based on SMOTE (Synthetic minority oversampling techniques) [3] and its many variants. The popularity of these methods is due to the simplicity of their application since they are implemented as a step of the data preparation phase in popular software for managing the statistical and machine learning pipeline (see for instance the `caret` package in `R`, or SciKitLearn in Python). For this reason, it is easy to neglect that generation of new data depends crucially on some choices, the most relevant is probably related to the choice of tuning (hyper)parameters. How far are new synthetic data from the observed ones depends on tuning parameters which in turn depend on the local density of the data. If data points (especially those in the minority class) are clustered, the default choice of the tuning parameters, which is the most common solution adopted, could lead to synthetic data which are too distant from observed data then leading to poor performance of classifiers when applied to augmented data. A possible solution is to use unsupervised classification before oversampling (for both the two classes) and to generate a synthetic sample conditional on the group the units belong to. This will make the standard choice of the tuning parameters less critical. In this paper, we will consider the application of some clustering techniques before using ROSE. We will consider density-based clustering techniques and more specifically (a) a modal clustering procedure based on kernel density estimation [1, 2] or (b) DBScan which is probably the most popular density-based clustering technique within the machine learning community. Both these methods can work with a default specification of extra hyper-parameters. The paper is organized as follows. In Sect. 2 we introduce the class imbalance problem, the ROSE algorithm, and some other alternative algorithms for oversampling the rare class. Sect. 3 illustrates the procedure called ROSEclust along with some density clustering procedures. Sect. 4 describes and reports some results from two simulation experiments. Sect. 5 contains some concluding remarks.

2 The class imbalance problem and oversampling techniques

Class imbalance is a largely recognized issue in supervised classification problems. Some events of potential major interest for the analysis are very rare and, limiting ourselves to the simplest case of a dichotomous response variable, for one class (the minority class) we observe only a limited, often very limited, number of data points, while for the majority class the number of data points observed could be, and often actually is, very large. Observing events only in a small number of cases makes

the classification problem troublesome and it is well recognized that standard classification models such as logistic regression, classification tree and Support Vector Machine (SVM) often provide unsatisfactory results.

In the past two decades, many approaches were developed to deal with supervised classification when observation in one of the classes is extremely rare. They can be broadly classified into 3 categories: (i) cost-sensitive learning, (ii) data preprocessing and (iii) algorithm modifications which can include also ensemble methods. For a comprehensive review of the imbalanced classification problem and of the various approaches to deal with, see [7].

Methods referring to the data pre-processing approach are likely the most used in real-life analyses. This is probably due to the fact that their application is generally straightforward and can be often carried out automatically by simply using some algorithmic recipes available within statistical and machine learning software.

The data pre-processing methods are aimed at modifying the proportions of the classes to obtain a balanced data set. This is done by using sampling or resampling techniques such as undersampling and oversampling. Before using data for training a classification algorithm, undersampling refers to the idea of taking a random sample from the majority class, while keeping all (or most of) the data in the minority class. The new sample is such that its size matches approximately the size of the minority class. In sampling the majority class some auxiliary information can be used to better represent the characteristics of the data. Undersampling is very successful especially with very large data sets. In this situation, the actual number of minority cases representing only a small proportion of the available data, is large enough to obtain a balanced data set that is suitable for successive analyses.

Oversampling aims at augmenting the minority class examples. Using classical bootstrap is also a viable option, but it might lead to unsatisfactory results and overfitting. Most of the methods proposed are then designed to generate new data points of the minority class which are similar to the rare observed cases.

ROSE is one of the methods which have been proposed to generate new synthetic data and it proved to be successful in many contexts. In the next subsection, the main characteristic of ROSE will be described.

2.1 ROSE and other oversampling methods

ROSE is the acronym of Random OverSampling Examples, as it builds on the generation of new artificial examples from (both) the classes, according to a smoothed bootstrap approach [5].

Consider a training set $\mathbf{T}_n$, of size n, whose generic row is the pair $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. The class labels y_i belong to the set $\{\mathscr{Y}_0, \mathscr{Y}_1\}$, and the $\mathbf{x}_i$ are some related attributes supposed to be realizations of a random vector $\mathbf{x}$ defined on R^d, with unknown probability density function $f(\mathbf{x})$. Denote with $n_j < n$ the number of units in class $\mathscr{Y}_j, j = 0, 1$. The ROSE procedure for generating one new artificial example consists of the following steps:

1. select $y^* = \mathscr{Y}_j$ with probability π_j
2. select $(\mathbf{x}_i, y_i) \in \mathbf{T}_n$, such that $y_i = y^*$, with probability $\frac{1}{n_j}$;
3. sample $\mathbf{x}^*$ from $K_{\mathbf{H}_j}(\cdot, \mathbf{x}_i)$, with $K_{\mathbf{H}_j}$ a probability distribution centered at $\mathbf{x}_i$ and having covariance matrix $\mathbf{H}_j$.

Essentially, we draw from the training set an observation belonging to one of the two classes, and generate a new example $(\mathbf{x}^*, y^*)$ in its neighbourhood, where the shape of the neighbourhood is determined by the shape of the contour sets of K and its width is governed by $\mathbf{H}_j$.

It can be easily shown that once a label class has been selected, the generation of new examples from class $\mathscr{Y}_j$, according to ROSE, corresponds to the generation of data from the kernel density estimate of $f(\mathbf{x}|\mathscr{Y}_j)$, with kernel K and smoothing matrix $\mathbf{H}_j$ (see [9]).

The package ROSE considers Gaussian kernels with diagonal smoothing matrices $\mathbf{H}_j = diag(h_j^{(1)}, \ldots, h_j^{(d)})$ where the vector of h_js' is selected as optimal under the assumption that the true conditional densities underlying the data follow a Normal distribution. This leads to

$$h_j^{(q)} = (4/((d+2)n_j))^{1/(d+4)} \hat{\sigma}_j^{(q)}, \quad j = 0,1, \quad q = 1, \ldots, d \tag{1}$$

where $\hat{\sigma}_j^{(q)}$ is a sample estimate of the standard deviation of the q-th dimension of the observations belonging to the class $\mathscr{Y}_j$. It is worthwhile to note that, for $\mathbf{H}_j \to 0$, ROSE produces a standard bootstrap resampling for the minority class. The package `ROSE` in `R` allows the user to set different values for h_j by multiplying it. So that one can force the generation of the new data very close to observed data points or more dispersed around the observed data. In practical applications, it has been noted that could be more efficient to use values for h_j much smaller (or much larger) than the default value. This largely depends on the structure of the data and actually, these tuning parameters should depend on the local density of the data.

Note that ROSE has been designed for using the same generation procedure also for the prevalent class.

The point we address with the method presented in the next sections is related to obtaining a more robust procedure in case of data which are far from being consistent with the assumption above.

SMOTE is the most popular alternative for generating synthetic cases, it is very simple but has no sound theoretical justification. In SMOTE a new point is generated in the minority class (only) by choosing randomly among available data points and then considering the K nearest neighbours according to an appropriate distance. The lines connecting the data point to its neighbours are then considered and new points are selected picking a point on one of these lines. The exact position on the lines is determined by randomly choosing a value between 0 and 1. The value K is a tuning parameter and for smaller K the generated values will be very close to observed points while for large K some synthetic points could be very far from the actual data. Being based on nearest neighbours, the method takes into account the local density of the data. There have been suggested a large number of variants of

SMOTE to tackle possible problems. For a more detailed description of SMOTE and some of its variants refer again to the book of [7].

3 The ROSEclust strategy

To improve the estimation performance of the models in the case of highly unbalanced data and with the presence of subgroups, we propose to combine clustering and balancing techniques. The idea is to verify if instances of the minority class belong to subgroups and then apply the balancing method to them. It should be added that similar ideas have already been put forward for improving on SMOTE algorithm (see, for instance, [4]). We propose a solution which is aimed instead at making the ROSE procedure more flexible to improve classification performance when data exhibits a more complex structure (especially in the minority class).

Under very general conditions, ROSE has proved to offer a reasonable solution, often overperforming other methods. A difficulty might emerge when data in the minority class are such that the value of the smoothing parameters set by default is not appropriate. In the case of clustering of the data, it could lead to the generation of synthetic data which are overly dispersed. A viable solution, but possibly computational demanding, then is a fine-tuning of the smoothing parameter aimed at obtaining better classification results. This problem could also emerge when data in the input space have a structure that requires the variable smoothing parameters to adapt to the local density. To this end, implementation of the ROSE algorithm in `R` [8] offers the possibility of setting the smoothing parameter to a multiple of the default value (`h.mult.mino`=1).

An alternative solution is proposed and explored in the sequel. The idea is to first detect possible clusters in the data and then use the default smoothing parameter separately for each cluster. When choosing an appropriate clustering algorithm it seemed consistent to select those which detect clusters as regions of high density. The natural notion of a cluster consisting of points gathered together in regions of high probability is very intuitive and forms the basis of density-based clustering. These methods gained large popularity in the last three decades with the consequent development of a very large number of algorithms.

Among the density-based clustering techniques, we have chosen DBScan, possibly the most popular density-based clustering algorithm within the machine learning community [6] and pdfCluster [1, 2]. This last method is based, like ROSE, on a kernel density estimate. For both of these methods, parameter tuning is not required, although some of them can be changed to deal with specific problems. Thus, the entire strategy works without choosing any parameters and could hopefully give good results by using standard default choices set into the software.

Once subgroups into the minority class, and possibly into the majority class at least for ROSE, have been identified the oversampling algorithms can be directly applied to each subgroup.

We expect that the performance of the classification algorithms tested on balanced data accounting for the presence of subgroups could be even better than those obtained by using ROSE (or SMOTE) after a grid search for an "optimal" setting of the tuning parameters.

4 Evidence from two simulation studies

We present a preliminary study of our approach on two simulated data sets. The main difference between the two data sets concerns the simulation of the explanatory variables for the instances of the minority class. In the first case, these were simulated by a mixture of bivariate normal distributions, so the presence of subgroups in the minority class is evident. In the second case, the definition of the minority class data does not directly include subgroups. Indeed, the explanatory variables of the minority class derive from a half-circle depleted filled with the prevalent class. These have been introduced with the acronym hacide (half-circle depleted) and have already been used to study the application of ROSE [9]. Thus they represent a benchmark for the ROSEclust methodology to study its effect in the absence of subgroups in the minority class. However, the rare class appears elongated and using the default values for the smoothing parameters might lead to overly dispersed synthetic data.

In the mixture data example, the features related to the majority class were simulated from a bivariate normal distribution defined as

$$\binom{x_1}{x_2}|y=0 \sim \mathscr{N}\left(\binom{0}{0}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right)$$

While those of the minority class follow a mixture of normal bivariate with three components

$$\begin{aligned}\binom{x_1}{x_2}|y=1 \sim\ & 0.3\mathscr{N}\left(\binom{-1}{-2}, \begin{pmatrix} 0.2 & -0.12 \\ -0.12 & 0.2 \end{pmatrix}\right) + 0.5\mathscr{N}\left(\binom{-0.5}{1}, \begin{pmatrix} 0.2 & -0.12 \\ -0.12 & 0.2 \end{pmatrix}\right) \\ & + 0.2\mathscr{N}\left(\binom{1.5}{-1}, \begin{pmatrix} 0.1 & -0.06 \\ -0.06 & 0.1 \end{pmatrix}\right)\end{aligned}$$

In this framework, we simulated three data sets with 5000 data points, each with the minority class proportion p equal to 0.01, 0.03, and 0.05.

The second study follows the definition of filled semi–hypersphere data in the bivariate case presented in [9]. For this second example, we simulated 1250 data points with an imbalance of the minority class equal to $p = 0.02$.

The data sets were randomly divided into a training set (70% of the mixture data and 80% of the hacide data) to fit the models and a test set (30% and 20%, respectively) to evaluate the performances of the algorithms. The clustering procedures pdfCluster and DBScan were applied only to the minority class of the training sets.

Next, we obtained the balanced training sets applying the ROSE methodology with different smoothing parameters on the minority class, and through ROSE on the subgroups identified by the clustering procedure using the default setting for

the smoothing parameters. With the same scheme, we replicated the balancing step using SMOTE with different values on the K parameter and SMOTE on the two classifications with default K value. Some examples of the mixture and hacide data obtained under different settings of the balancing methodologies are shown in Fig 1.

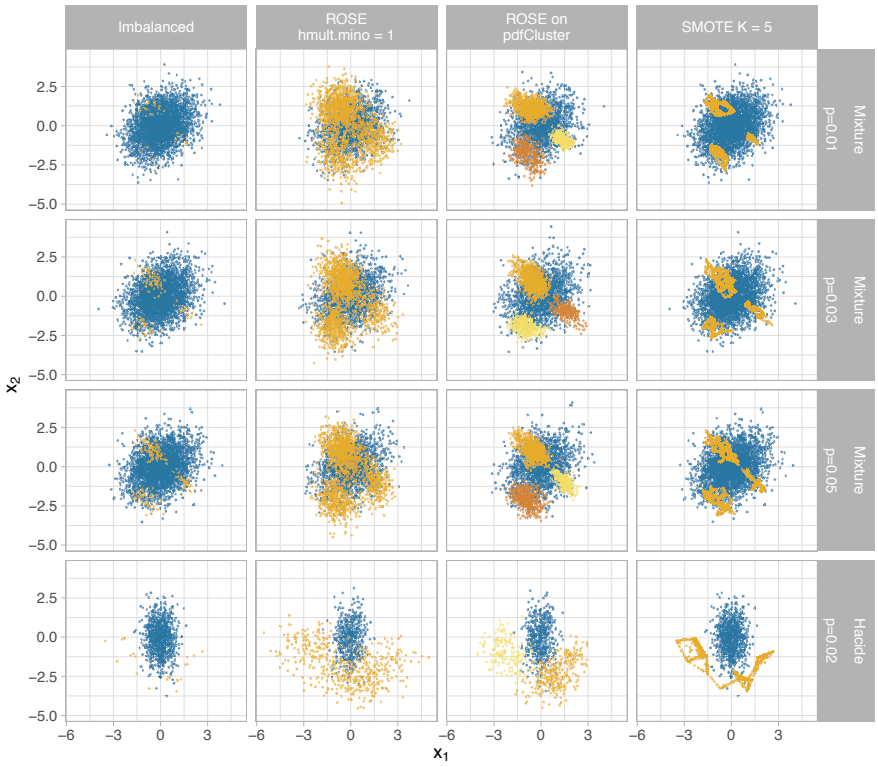


Fig. 1 Training sets with different proportion of the minority class (rows: $p = (0.01, 0.03, 0.05)$ for mixture data and $p = 0.02$ for hacide data). Each training set contains 3500 data points for the three mixture data examples and 1000 data points for the hacide data. The first column shows the imbalanced data set, while the remaining three columns depict data sets balanced with different methodologies: ROSE with smoothing parameter for the minority class equal to the default value (`h.mult.mino` = 1), ROSE applied to clusters identified by pdfCluster, and SMOTE with a number of nearest neighbours used during sampling process equal to the default value ($K = 5$). Blue points represent the majority class, orange (yellow, and red) the minority class (and clusters).

The image unveils that when ROSE accounts for the clusters, the simulated data of the minority class are markedly less variable if compared with ROSE. The difference between the two approaches, with and without clusters, increases as the imbalance of the data grows. It is also worth noting the different oversampling ap-

proaches of the ROSE and SMOTE algorithms in their default settings. The data generated for the minority class by SMOTE are more concentrated on the observed values even without considering clusters.

4.1 Results

About the clustering algorithms, both algorithms identified the three groups in the minority class of mixture data for $p = 0.05$ and $p = 0.03$. DBScan classified some instances as noise (4 for $p = 0.05$ and 5 for $p = 0.03$) that were excluded from the subsequent steps involving the DBScan clusters. On mixture data with $p = 0.01$, pdfClusters identified 3 clusters, while DBScan found 4 clusters and no noise. The two classification techniques returned the same result on the hacide data identifying 2 clusters in the minority class. Running a small simulation study on 10000 hacide data sets, we found that in about 18% of cases, the two clustering methodologies did not detect subgroups on the training sets. These cases are not of particular interest for the present study, as the results between the original and the proposed techniques do not change. To better highlight the effect of the methodology when clusters are identified even if they are not defined, we have selected an example in which pdfCluster and DBScan found two subgroups of the minority class. For the sake of completeness, we report that, in our simulations study, pdfCluster identified two or more groups in 53% of cases, while DBscan in 66%.

As for the classification methods adopted we considered: logistic regression, classification tree, random forest, and the boosting, specifically AdaBoost.M1. Predictive performances of the classification algorithms were evaluated on the test set, containing 30% points from the imbalanced data set, through the Area under the ROC Curve (AUC) and other metrics computed from the confusion matrix. For each model, the threshold value, namely the cutoff point for classifying an unseen case to be in the minority class, maximizes the Youden's J statistic, that is, maximizes both sensitivity and specificity, as defined later.

The most popular validation metric derived from the confusion matrix is the accuracy that evaluates the proportion of correctly predicted values, not differentiating between positives of the majority and minority class but it is well known that this measure is the least suitable in case of imbalanced data. Performance metrics beyond accuracy allow us to highlight interesting aspects of the quality of the prediction performance in dealing with highly imbalanced data where the focus is on the minority class. Sensitivity (or recall) measures the proportion of the correctly classified instances in the minority class over the observed instances in the minority class. Ideally, in the imbalance classification context, a high sensitivity should be combined with high precision, that is the proportion of correctly classified in the minority class over the instances predicted in the minority class. These two aspects are summarized in their harmonic mean called F1 measure.

Results shown in Table 1 refer to the prediction obtained for the test sets of the four models (logistic, classification tree, random forest, and boosting) considering

three different proportions of imbalance in the mixture data. Models were fitted on the imbalanced and balanced training sets using ROSE and SMOTE in their default settings and in evaluating the clusters identified in the minority class.

Table 1 AUC and F1 measure for the models fitted on imbalanced data varying the proportion of minority class ($p = 0.01, 0.03,$ and 0.05), on balanced data using ROSE and SMOTE with their default setting (`h.mult.mino`=1 and $K = 5$, respectively), and on balanced data using ROSE and SMOTE on the classification results of the pdfCluster and DBScan algorithms.

Model	Data set	Clustering	p=0.01		p=0.03		p=0.05	
			AUC	F1	AUC	F1	AUC	F1
Logistic								
	Imbalanced		0.627	0.033	0.575	0.079	0.568	0.072
	ROSE		0.697	0.039	0.601	0.08	0.579	0.177
	ROSE	pdfCluster	0.697	0.039	0.605	0.078	0.579	0.177
	ROSE	DBScan	0.698	0.039	0.598	0.08	0.581	0.121
	SMOTE		0.696	0.039	0.596	0.082	0.579	0.177
	SMOTE	pdfCluster	0.689	0.043	0.592	0.082	0.58	0.12
	SMOTE	DBScan	0.696	0.039	0.587	0.082	0.585	0.122
Tree								
	Imbalanced		0.5		0.5		0.5	
	ROSE		0.73	0.097	0.751	0.213	0.847	0.311
	ROSE	pdfCluster	0.789	0.071	0.863	0.175	0.859	0.293
	ROSE	DBScan	0.767	0.06	0.836	0.178	0.867	0.295
	SMOTE		0.76	0.09	0.809	0.228	0.855	0.333
	SMOTE	pdfCluster	0.769	0.076	0.839	0.206	0.866	0.339
	SMOTE	DBScan	0.769	0.078	0.782	0.2	0.871	0.38
Rdm Forest								
	Imbalanced		0.588	0.04	0.858	0.184	0.894	0.293
	ROSE		0.722	0.037	0.865	0.176	0.856	0.234
	ROSE	pdfCluster	0.83	0.084	0.87	0.149	0.903	0.33
	ROSE	DBScan	0.814	0.051	0.881	0.167	0.9	0.3
	SMOTE		0.821	0.094	0.884	0.206	0.889	0.293
	SMOTE	pdfCluster	0.866	0.082	0.879	0.206	0.882	0.322
	SMOTE	DBScan	0.795	0.054	0.864	0.196	0.887	0.328
Boosting								
	Imbalanced		0.627	0.03	0.9	0.187	0.917	0.376
	ROSE		0.821	0.052	0.886	0.191	0.901	0.341
	ROSE	pdfCluster	0.868	0.084	0.909	0.17	0.921	0.327
	ROSE	DBScan	0.854	0.053	0.893	0.199	0.913	0.395
	SMOTE		0.797	0.078	0.891	0.196	0.92	0.377
	SMOTE	pdfCluster	0.813	0.08	0.888	0.193	0.92	0.387
	SMOTE	DBScan	0.802	0.085	0.889	0.214	0.909	0.335

Regardless of the degree of data imbalance and the balancing technique, the logistic model appears to be the least suitable of the four to estimate the data under consideration. Even if the AUC seems higher for the mixture data with $p = 0.01$ compared to the other two scenarios, the F1 measures reveal a low capacity of the models in predicting the minority class accurately. The other three methods

show increasing predictive performances in both the AUC and the F1 measure as the proportion of the minority class increases. The remarkable difference between the measures obtained from the models estimated on imbalanced and balanced data sets emerges in almost all cases. Furthermore, the differences between the measures obtained using the balanced data sets and the cluster-based balanced ones are consistently greater for ROSE than for SMOTE. We conjecture that this could be due to the different oversampling procedures used by the two methods. SMOTE, even without considering clusters, generates synthetic data closer to the observed one. Therefore, the impact of the oversampling step conditioned on the clustering partition is more relevant for ROSE than for SMOTE. Except for the case of $p = 0.01$, both random forest and boosting appear to work well with appropriate threshold selection. Overall, the best results for the mixture data are recorded by applying one of the two oversampling techniques to the clusters.

Table 2 shows more detail on the predictive performance of the four models as the ROSE and SMOTE parameters vary for an imbalance proportion of $p = 0.01$ in the mixture data. In particular, for ROSE 11 values of the smoothing parameter `h.mult.mino` were considered (from 0, equivalent to bootstrap, to 2.5); while for SMOTE an incremental number of nearest neighbours K from 1 to 11 were selected. Except for the logistics model, the results reveal a decreasing trend in the predictive performances of the models estimated on balanced data for parameter values equal (for ROSE) or greater than the default values. Indeed, the best predictive results for the mixture data are obtained by reducing the parameter default values, e.g. in the random forest model, or applying the balancing techniques to the identified clusters. Following considerations about Fig.1, the effect is more evident for ROSE than for SMOTE.

Table 3 reports the AUC of the four models fitted on the hacide training set. It is worth noting that also in this case the predictive capabilities of models on balanced data are always higher than that of models estimated on unbalanced data. Furthermore, although the groups were not present by definition, the predictive performance of the models fitted on balanced data conditioning on the stated clusters does not appear to be compromised. Also, the ROSE and SMOTE results in all their variations are comparable, except for the classification tree models.

Finally, some general considerations on the applied methods that make them more or less suitable for different contexts. Of the two clustering procedures tested, DBScan has the highest number of parameters to set and can identify points as noise. In some situation s, it can prove to be an advantage, but when dealing with highly imbalanced data it can lead to an additional reduction of the minority class instances or a required fine-tuning of the algorithm parameters. Furthermore, the undeniable impact of the balancing techniques on predictive performances is confirmed. Also, is clear how the effect of oversampling techniques conditional to clusters is weaker on SMOTE than on ROSE. On the other hand, ROSE on clusters results to be a tuning-free method that can also handle categorical data.

Table 2 AUC results of the models fitted on the imbalanced data with minority class proportion p=0.01, on balanced data using ROSE and SMOTE varying the algorithm parameters (h.mino.mult from 0=bootstrap to 2.5; *K* from 1 to 11), and on balanced data using ROSE and SMOTE on the minority class clusters identified by pdfCluster and DBScan methods.

Data set		Clustering	AUC			
			Logistic	Tree	Rdm Forest	Boosting
Imbalanced			0.627	0.5	0.588	0.627
ROSE	h					
	0		0.703	0.767	0.717	0.701
	0.25		0.702	0.811	0.810	0.846
	0.5		0.702	0.750	0.840	0.855
	0.75		0.700	0.773	0.814	0.853
	1		0.697	0.730	0.722	0.821
	1.25		0.697	0.651	0.685	0.784
	1.5		0.697	0.652	0.655	0.771
	1.75		0.694	0.645	0.746	0.783
	2		0.696	0.602	0.611	0.757
	2.25		0.691	0.586	0.534	0.679
	2.5		0.691	0.597	0.593	0.706
	1	pdfCluster	0.697	0.789	0.830	0.868
	1	DBScan	0.698	0.767	0.814	0.854
SMOTE	*K*					
	1		0.695	0.762	0.742	0.760
	2		0.694	0.715	0.799	0.776
	3		0.697	0.759	0.832	0.810
	4		0.697	0.776	0.847	0.849
	5		0.696	0.760	0.821	0.797
	6		0.701	0.770	0.846	0.812
	7		0.691	0.727	0.793	0.797
	8		0.681	0.716	0.748	0.782
	9		0.677	0.715	0.776	0.756
	10		0.676	0.721	0.735	0.784
	11		0.657	0.695	0.763	0.784
	5	pdfCluster	0.689	0.769	0.866	0.813
	5	DBScan	0.696	0.769	0.795	0.802

Table 3 AUC results of the models fitted on the imbalanced hacide data with minority class proportion p=0.02, on balanced data using ROSE and SMOTE with default setting, and on balanced data using ROSE and SMOTE on the minority class clusters identified by pdfCluster and DBScan methods.

Data set	Clustering	AUC			
		Logistic	Tree	Rdm Forest	Boosting
Imbalanced		0.896	0.832	0.98	0.989
ROSE		0.904	0.935	0.987	0.993
ROSE	pdfCluster	0.904	0.984	0.994	0.991
ROSE	DBScan	0.903	0.957	0.989	0.995
SMOTE		0.904	0.819	0.989	0.986
SMOTE	pdfCluster	0.904	0.819	0.982	0.976
SMOTE	DBScan	0.904	0.985	0.995	0.991

5 Concluding remarks

Introducing variants of the base oversampling technique could be appropriate in many cases (and indeed many variants have been proposed for SMOTE), but it is worth looking for a more general technique which can work in a large majority of the applications. We aim to make the data preparation step, including oversampling via the generation of new synthetic data, as simple as possible, leaving more room for the modelling step. ROSEclust is a generalization of the ROSE procedure to deal with a possible more complex structure of the data. Preliminary results show us that it can be also used as a technique to select the hyper-parameters according to the local density of the data. Admittedly the presented results are largely preliminary but they encourage us to confirm the promising results by enlarging simulation studies and, more importantly, future work will focus on applying the ROSEclust procedure to various other real-world problems.

References

1. Azzalini, A., Torelli, N.: Clustering via nonparametric density estimation. Stat. Comput. **17(1)**, 71–80 (2007)
2. Azzalini, A., Menardi, G.: Clustering via Nonparametric Density Estimation: The R Package pdfCluster. J. Stat. Softw. **57(11)**, 1–26 (2014)
3. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-Sampling Technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
4. Douzas, G., Bacao, F., Last F.: Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE Inf. Sci. **465**, 1–20 (2018)
5. Efron, B., Tibshirani, R.: An introduction to the bootstrap. Chapman and Hall, New York (1993)
6. Ester, M., Kriegel, H. P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd **96(34)**, 226–231 (1996, August)
7. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer International Publishing, (2018)
8. Lunardon, N., Menardi, G., Torelli, N.: ROSE: a Package for Binary Imbalanced Learning. The R Journal, **6(1)**, 79–89 (2014)
9. Menardi G., Torelli, N.: Training and Assessing Classification Rules with Data Imbalance. Data Min. Knowl Disc. **28(1)**, 92–122 (2014)

Estimating the effect of remote teaching for university students through generalised linear mixed models

Stima dell'effetto della didattica a distanza per gli studenti universitari mediante modelli lineari misti generalizzati

Silvia Bacci and Bruno Bertaccini and Simone Del Sarto and Leonardo Grilli and Carla Rampichini

Abstract The present paper aims at analysing the effects of remote teaching on university students' careers, as a consequence of university closures due to the COVID-19 pandemic. For this purpose, we use administrative data of the University of Florence on students' careers and compare their performance in terms of the probability of passing specific exams. In particular, using a random intercept logit model, we compare the group of students enrolled in the academic year 2018/2019 – who received classic face-to-face teaching – with the group of students enrolled in the subsequent academic year, who experimented remote teaching during the second semester. Results obtained on different degree programs show that the effect of remote teaching at the course level is markedly heterogeneous, with different sign and magnitude.

Abstract *Il presente lavoro si propone di analizzare gli effetti della didattica a distanza sulla carriera degli studenti universitari, a seguito della chiusura delle università dovute alla pandemia di COVID-19. A tal fine, utilizziamo i dati amministrativi dell'Università di Firenze sulla carriera degli studenti e confrontiamo la loro*

Silvia Bacci
Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
e-mail: silvia.bacci@unifi.it

Bruno Bertaccini
Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
e-mail: bruno.bertaccini@unifi.it

Simone Del Sarto
Department of Political Science - University of Perugia
e-mail: simone.delsarto@unipg.it

Leonardo Grilli
Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
e-mail: leonardo.grilli@unifi.it

Carla Rampichini
Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
e-mail: carla.rampichini@unifi.it

produttività in termini di probabilità di superare specifici esami. In particolare, utilizzando un modello logit a intercetta casuale, confrontiamo il gruppo di studenti iscritti all'a.a. 2018/2019 – che hanno frequentato la classica didattica frontale – con il gruppo di studenti iscritti all'a.a. successivo, che hanno sperimentato la didattica a distanza nel secondo semestre. I risultati ottenuti sui diversi corsi di studio mostrano che l'effetto della didattica a distanza a livello di insegnamento è marcatamente eterogeneo, con segno e grandezza differenti.

Key words: COVID-19, distance learning, logit model, random effects, university exams

1 Introduction

The COVID-19 pandemic has had a resonant impact in all aspects of social life. The educational field has been affected by the pandemic as well, due to schools and universities closure. Focusing on the academic world, universities are intensively dealing with the pandemic, whose effects consisted in an impact on enrolments, demands for large tuition cuts, redesign of courses and learning approaches by teachers. On the other side, students had to face with an uncertain environment due to financial and health shocks, as well as to the introduction of remote teaching activities, which contributed to jeopardise their university performance, educational plans, labour market participation, and in general expectations about future [2].

Several studies attempt to evaluate the impact of COVID-19 on higher education students' experiences, in particular on their academic performance [4, 6, 7, 5]. Although all these studies highlight their overall difficulties, they detect opposing results in terms of effects on students' performance.

However, existing works on the effects of remote teaching on the university students' performance are mainly based on specifically designed surveys. In this paper, we contribute to the research on the present topic by proposing a modelling approach that allows us to exploit use administrative data on students' career.

The approach is applied to data on passed exams about students' belonging to different degree programs at the University of Florence. In particular, we aim at comparing two cohorts of students: *i*. those enrolled in academic year 2018/2019, who did not experience remote teaching at all, and *ii*. those enrolled in the subsequent academic year 2019/2020, who experimented remote teaching only as regards the courses held in the second semester.

The comparison is performed separately for each degree program, by analysing students' performance in terms of probability of passing the exams of the courses envisaged by the degree program. Given that the same student has to take several exams (hierarchical data structure, with exams nested within students), a random intercept logit model is estimated and the remote teaching effect for each course is detected in terms of difference in the probability of passing the exams between the two cohorts.

The paper is organised as follows. Section 2 is dedicated to a description of the data and the statistical model used for the analyses, whose results are presented in Section 3. Finally, Section 4 draws some concluding remarks.

2 Data and proposed model

Data from the administrative archive of the University of Florence on students' careers are considered, including information on passed exams together with some students' background details (e.g., gender, high school type and grade). Data about students of the two cohorts (2018/2019 and 2019/2020) are extracted from the archive as regards the the following five bachelor degree programs: *i*. Chemistry; *ii*. Industrial design; *iii*. Law; *iv*. Mechanical engineering; *v*. Psychology. Descriptive statistics on these data are reported in Table 1.

Table 1 Share of students (%) that passed the exams related the courses envisaged in each degree program (second semester of the first year): comparison of observed raw outcomes between cohorts (sizes of cohorts in parenthesis).

			Cohort		
Degree program	Credits	Course	2018	2019	Diff.
Chemistry	6	CHEM1	11.2	5.4	−5.8
($n_{2018} = 98$, $n_{2019} = 112$)	12	CHEM2	42.9	29.5	−13.4
	6	CHEM3	9.2	14.3	5.1
	6	CHEM4	27.6	23.2	−4.4
Industrial design	6	DES1	80.5	77.2	−3.3
($n_{2018} = 149$, $n_{2019} = 167$)	6	DES2	61.1	60.5	−0.6
	12	DES3	63.8	56.9	−6.9
Law	12	LAW1	57.3	53.0	−4.3
($n_{2018} = 347$, $n_{2019} = 421$)	9	LAW2	36.9	41.3	4.4
	9	LAW3	58.8	55.1	−3.7
Mechanical engineering	9	ENG1	47.1	57.2	10.1
($n_{2018} = 325$, $n_{2019} = 318$)	6	ENG2	14.8	22.0	7.2
	12	ENG3	44.0	23.6	−20.4
	12	ENG4	21.2	20.1	−1.1
Psychology	9	PSY1	77.5	72.1	−5.4
($n_{2018} = 427$, $n_{2019} = 426$)	9	PSY2	75.4	76.1	0.7
	6	PSY3	69.6	62.7	−6.9
	6	PSY4	66.0	71.4	5.4

Given the hierarchical structure of our data (exams to take nested within students), we consider a mixed model formulation. In particular, the response variable is the exam outcome (passed/not passed) of each student as regards the courses envisaged in the second semester of the corresponding degree program study plan

(first year compulsory courses). Moreover, in order to control for differences among students of the two cohorts, we include a control variable in the model, namely the student's performance at the first semester.

Given degree program p with N_p enrolled students and M_p courses envisaged at the second semester of the first year, our dichotomous response variable, denoted by Y_{ij}, is equal to 1 if student i passes exam j, and 0 otherwise, with $i = 1,\ldots,N_p$ and $j = 1,\ldots,M_p$. In order to consider the correlation between exams of the same student, a generalised linear mixed model is employed for modelling the probability of passing exam j by student i, $P(Y_{ij} = 1)$:

$$\text{logit}(P(Y_{ij} = 1|\boldsymbol{x}_i, D_i)) = \gamma_j + \delta_j D_i + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \tag{1}$$

where D_i is a dummy variable for the cohort (reference level is cohort 2018/2019) and $\boldsymbol{x}_i$ is the vector of student covariates. Specifically, two covariates are included in this model as regards the student's performance at the first semester, namely the proportion of gained credits and a dummy variable for students getting zero credits during the first semester.

The model at issue has an exam-specific intercept γ_j, while parameter δ_j represents the effect of remote teaching on exam j, as it is the variation in the model intercept between the two cohorts of students, that is, the difference on the logit of passing exam j between the two cohorts. Finally, random intercepts u_i are independent normally distributed, with zero mean and constant variance σ_u^2.

3 Results

Model (1) is fitted separately for students belonging to each degree program by means of the R package `lme4`[3]. However, given that the parameter of greatest interest δ_j – which represents the effect of the remote teaching on exam j – is on the logit scale, we compute the average marginal effect (AME)[1], that is, the average discrete difference in the probability of passing the exam between 2018/2019 and 2019/2020.

AMEs are reported in Table 2, together with the corresponding 95% confidence intervals. For example, looking at the first course of the Psychology degree program (PSY1), a negative and significant effect of remote teaching is detected. In fact, the related AME is -0.082, hence, when comparing cohort 2019/2020 with respect to cohort 2018/2019, the probability of passing this exam decreases, on average, by 8.2% (first semester performance being equal).

As can be noticed, both positive and negative effects can be highlighted in each degree program, although they are significant in few cases. In fact, in Chemistry, Industrial Design and Law, no significant effect (at 5%) emerges from the analysis, whereas the most pronounced effects are detected as regards students of Mechanical engineering and Psychology. Students of the former program have both positive (courses ENG1 and ENG2) and negative (ENG3) significant performance: in partic-

Table 2 Average Marginal Effects (AME) and 95% confidence interval (95% CI) by degree program from model (1)

Course	AME	95% CI
	Chemistry	
CHEM1	$-0.066^{\cdot}$	$(-0.135, 0.002)$
CHEM2	-0.034	$(-0.089, 0.022)$
CHEM3	$0.079^{\cdot}$	$(-0.003, 0.161)$
CHEM4	0.017	$(-0.048, 0.081)$
	Industrial design	
DES1	0.064	$(-0.026, 0.155)$
DES2	$0.066^{\cdot}$	$(-0.009, 0.141)$
DES3	0.013	$(-0.068, 0.094)$
	Law	
LAW1	$-0.053^{\cdot}$	$(-0.108, 0.002)$
LAW2	0.036	$(-0.018, 0.090)$
LAW3	-0.046	$(-0.102, 0.009)$
	Mechanical engineering	
ENG1	0.096^{**}	(0.040, 0.152)
ENG2	0.108^{**}	(0.043, 0.174)
ENG3	-0.167^{***}	$(-0.225, -0.110)$
ENG4	0.010	$(-0.054, 0.073)$
	Psychology	
PSY1	-0.082^{**}	$(-0.133, -0.030)$
PSY2	-0.017	$(-0.065, 0.031)$
PSY3	-0.086^{***}	$(-0.134, -0.038)$
PSY4	0.025	$(-0.019, 0.068)$

Significance levels: *** = 0.001; ** = 0.01; * = 0.05; $^{\cdot}$ = 0.10

ular, for this latter course, the greatest (in absolute value) effect of remote teaching is outlined, equal to -0.167. Finally, two significant and negative effects of the same magnitude (around -0.08) are detected for Psychology.

4 Conclusions

Due to COVID-19 pandemic, universities have had to employ emergency strategies to carry out teaching activities, such as switching from face-to-face to remote teaching. However, its implementation within the same university can be very heterogeneous, as teachers could customise their online courses, despite the availability of general guidelines at the university level. Consequently, remote teaching effect on students' performance can be pretty multifaceted.

Relying on the generalised linear mixed modelling framework, we studied the effect of remote teaching at the single course level, by exploiting the administrative archive on students' careers of the University of Florence. Specifically, we compared the performance (in terms of probability of passing an exam) of students belonging to different bachelor's degree programs and from two separate cohorts, of which only one experienced remote teaching.

As expected, our analysis underlined negative and positive remote teaching effects among different degree programs and, also, within the same degree program, even if the detected effects were not significant in most cases.

The present work presents some drawbacks. Firstly,the outcome is based on exam results, but we are aware that passing the exam can only be considered a proxy of learning achievement. Secondly, the analysis does not allow us to separate remote teaching effect from the impact of new exam rules, as the data do not include details on the examination modalities.

References

1. Agresti, A., Tarantola, C.: Simple ways to interpret effects in modeling ordinal categorical data. Stat. Neerl., **72**(3), 210–223 (2018)
2. Aucejo, E. M., French, J., Ugalde Araya, M. P., Zafar, B.: The impact of COVID-19 on student experiences and expectations: evidence from a survey. J. Public Econ. **191**, 104271 (2020)
3. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed effects models using lme4. J. Stat. Softw., **67**(1), 1–48 (2015)
4. Gonzalez, T., Rubia, M. A. de la, Hincz, K. P., Comas-Lopez, M., Subirats, L., Fort, S., Sacha, G. M.: Influence of COVID-19 confinement on students' performance in higher education. PLoS One, **15**(10), e0239490 (2020)
5. Iglesias-Pradas, S., Hernández-García, Á., Chaparro-Peláez, J., Prieto J.L.: Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: a case study. Comput. Hum. Behav., **119**, 106713 (2021)
6. Mahdy, M. A. A.: The impact of COVID-19 pandemic on the academic performance of veterinary medical students. Front. Vet. Sci, **7**, 594261 (2020)
7. Realyvasquez-Vargas, A., Aracely Maldonado-Macias, A., Cecilia Arredondo-Soto, K., Baez-Lopez, Y., Carrillo-Gutierrez, T., Hernandez-Escobedo, G.: The impact of environmental factors on academic performance of university students taking online classes during the COVID-19 pandemic in Mexico. Sustainability, **12**(21), 9194 (2020)

Perceived stress across EU countries: does working from home impact?
La percezione dello stress lavorativo in Europa: gli effetti del telelavoro

Abstract The study concerns the relationship between self-assessed occupational stress and workers' characteristics, stemming from the Sixth European Working Conditions Survey. Specific tasks, which are generally performed more often by women, such as caregiving and house working activities, are also considered, as well as home-based teleworking condition. The analysis, carried out by means of a heteroskedastic Ordered Probit model, provides results that are partially expected, such as the effects of gender and age on the response patterns, and the influence of the presence of children in the household. Besides, some unexpected findings are presented, as the statistical non-significance of specific family care commitments. Future research will be directed at investigating the role of gender and the distinction between employees and the self-employed respondents.

Abstract *Lo studio analizza la relazione tra la percezione dello stress da lavoro e le principali caratteristiche dei lavoratori, utilizzando i dati della sesta edizione dell'Indagine Europea sulle Condizioni di Lavoro. Il focus è rivolto al genere, alle attività di cura familiare e al lavoro domestico, più spesso svolti dalle lavoratrici, anche considerando l'effetto del telelavoro. Taluni dei risultati emersi dall'applicazione di un modello Ordered Probit eteroschedastico possono considerarsi parzialmente attesi, come l'effetto del genere e dell'età dei rispondenti; altre evidenze sono invece meno attese, come la non rilevanza statistica di alcuni oneri di cura familiare. Futuri approfondimenti della ricerca saranno diretti ad approfondire il ruolo svolto dal genere, anche distinguendo tra lavoratori dipendenti e autonomi.*

Key words: Occupational stress, Ordered Probit; EWCS, Teleworking

Introduction

Aim of the paper is investigating work-related stress across European countries, explicitly considering specific duties and tasks which are performed more often by women, such as caregiving and house working activities, together with the circumstance to telework from home (Del Boca et al., 2020).

In addition to the complex privacy issues and the recently acknowledged right to disconnect, the enhanced flexibility and autonomy implied by home-based teleworking frequently come with greater work intensity and longer working hours (European Parliament, 2021). Furthermore, the associated detrimental effects on workers' work-life balance are more often registered in case of women with caring

responsibilities and especially of working mothers (Chung and Van der Horst, 2018; Pascucci et al., 2021). Therefore, in order to provide insights in a perspective of integrated European policies towards a healthier, happier and more sustainable quality of life, our research question is to comprehend whether undertaking those extra duties could exert an effect on occupational stress, as suggested by the literature (among many others: Repetti et al., 1989; Eurofound and ILO, 2017; Messenger, 2019).

The present study is conducted employing data from the Sixth European Working Condition Survey (EWCS), carried out in 2015, which is the most recent representative information source at EU level on working conditions so far. Of course, related findings are to be interpreted in light of a pre-Covid-19 scenario.

The paper is organized as follows: the next Section presents the employed data, the main descriptive statistics, and the selected model; the results of the modelling implementation and related discussion are in Section 3; finally, Section 4 presents brief concluding remarks.

Data and methods

Data from the EWCS are used focusing on EU-28 countries and, more specifically, considering responses to question Q61M: “You experience stress in your Work?”, as measured over a 5-point Likert scale.

The EWCS provides comprehensive evidence on a wide range of topics related to workers and workplaces, including exposure to physical and psychosocial risks, work organization, balance between private and professional life, as well as perceived health and well-being. Most recent EWCS data refer to 2015 determining that the available information may offer a picture of “pre-COVID 19 Europe at work” when home-based teleworking was still a quite marginal feature of labour market and telework arrangements were made mostly on a voluntary basis. Additionally, in our modelling implementation other specific work-life balance features connected with home-based work are considered.

A preliminary screening for missing values of the selected explanatory variables lowers the original sample size to 22,864 respondents (52.72% are women). When distinguishing 5 age classes for respondents, in accordance with the sampling design, 14.37% are aged under 30 years; 24.72% are between 30 and 40 years old; 28.43% are between 40 and 50 years old; 25.82% are in between 50 and 60 years old, while those over 60 are 6.65%.

In such target sample, 3.47% of the workforce declares to “daily” work from home, and a similar proportion (4.40%) states to work from home “several times a week”, with no remarkable difference by gender. Considering the latter respondents as home-based workers, about 33% of them declare to experience “always” or “most of the time” occupational stress. This sub-set response pattern seems to be quite different as compared to that of the interviewees who do not work from home on a regular basis. In fact, descriptive results for the whole sample indicates that about 27% of workers claim to experience occupational stress constantly, since the distribution of the answers is: “Always” (10.47%); “Most of the time” (16.79%); “Sometimes”

(40.14%); "Rarely" (20.15%); "Never" (12.43%). Furthermore, women report a higher level of perceived stress, although these proportions are not so largely different across gender.

To interpret responses with respect to specific determinants and workers' characteristics, different modelling approaches can be implemented. Aiming to analyse ordinal data, Agresti (2010), Tutz (2012), and Piccolo and Simone (2019) may be considered as the main references. Due to the nature of the available data, the most straightforward choice is a heteroskedastic Ordered Probit Model (Agresti 2010) in order to detect the effects of subjective, environmental, and economic variables on reported occupational stress.

Drivers of interest, selected from the available information set according to the established literature on the topic (among others: Eurofound and ILO, 2017; Messenger, 2019), encompass three main areas: basic socio-demographics; workers' family management issues; job-related features. The considered socio-demographic variables are: gender, age classes, education level (using 2 dummies: one for tertiary education and one for high school degree), and number of family components. Some variables related to respondents' family management are: making-ends-meet, house working, caregiving, leisure activities (measured on a 5 point Likert scale ranging from 1=*never* to 5=*always*). Work-related features are expressed by dummies referred to permanent contract, full-time job, private sector and home-based work.

Additionally, we consider several self-registered assessments measured again on a 5-point Likert scales (from 1=*never* to 5=*always*) for work-life balance, fitting of working time, autonomy of decision at work, and a continuous variable for the amount of working hours per week. Finally, a dummy is used to discriminate for geographical aspects, considering countries of Northern Europe=1.

Results and discussion

The estimated model (Table 1) with self-assessed occupational stress being the response variable shows that gender, age class, and number of components of the household turn out to be statistically significant. Respondents' level of education does impact significantly when considering university degree. With respect to work-related characteristics, significant effects are those related to the number of working hours, permanent job contract and working in the private sector. Likewise, having some influence on decision at work and regularly working from home seem to exert some impact on stress perception, while having a full-time job is not statistically significant. The fitting of working hours with family or social commitments outside work, as well as making-ends-meet also exert some effects. Enjoying leisure activities turns out to be slightly significant. Somewhat unexpected results come from caregiving, childcare and house working activities, which do not seem to influence respondents' experience of work-related stress. Finally, living in a Northern European country, where welfare settings and job regulation systems are supposed to be more homogeneous, is also a significant covariate.

Since in the Ordered Probit model neither the sign nor the magnitude of the coefficients provides any information about the partial effects of a given explanatory variable, it could be useful to consider marginal effects of selected variables on the dependent one.

Table 1: Heteroskedastic Ordered Probit estimated coefficients

Stress	Coef.	Std. Err.	z	P>z	
Gender	0.118	0.015	8.07	0.000	***
Age class	-0.031	0.006	-5.43	0.000	***
Highschool education	0.014	0.015	0.97	0.330	
Tertiary education	-0.031	0.015	-2.03	0.042	**
Household components	-0.020	0.006	-3.37	0.001	***
Children	0.014	0.019	0.74	0.457	
Permanent job	0.125	0.018	6.95	0.000	***
Private sector	-0.049	0.013	-3.84	0.000	***
Full time job	-0.022	0.019	-1.19	0.233	
Working hours	0.009	0.001	10.53	0.000	***
Make-ends-meet	-0.016	0.005	-2.92	0.003	***
Childcare	0.003	0.005	0.53	0.597	
House working	0.002	0.006	0.36	0.717	
Caregiving	0.006	0.006	1.14	0.253	
Working hours fit	-0.282	0.013	-21.49	0.000	***
Home-based Telework	0.193	0.023	8.31	0.000	***
Influence on decisions	0.039	0.005	7.82	0.000	***
Leisure	0.009	0.005	1.78	0.076	*
D_North	0.113	0.014	7.99	0.000	***
lnsigma					
Make-ends-meet	-0.033	0.005	-6.94	0.000	***
Household components	0.011	0.005	2.37	0.018	**
Permanent job	-0.098	0.016	-6.02	0.000	***
Working hours	0.000	0.001	-0.74	0.461	
Age class	0.008	0.005	1.57	0.116	
/cut1	-1.428	0.076	-18.78	0.000	***
/cut2	-0.804	0.063	-12.78	0.000	***
/cut3	0.128	0.055	2.33	0.020	**
/cut4	0.700	0.060	11.72	0.000	***

***: significant at 1%; **: significant at 5%; *: significant at 10%

Specifically, we investigate the predicted probabilities of "stress"=never (Table 2) and "stress"=always (Table 3), for a worker profile holding high-school diploma, a full-time permanent job in the private sector, married with children and declaring to be in a household made of 3 components, working 40 hours per week, with the ordinal considered variables being fixed at their modal values. The profile is also distinguished by gender, home-based/non-home-based working condition and living in a Northern/Southern country.

Among respondents who state to not regularly work from home (non-home-based), men present a higher estimated probability to perceive lower occupational stress as compared to women, with similar results for both Northern and Southern European Union countries.

Women are more likely to report they perceive "always" stress at work, and this is more evident for those who usually work from home, and especially in a Northern country. In addition to well-known work/family reconciliation issues, among the reasons explaining why female respondents report higher occupational stress, it could be said that women may deal with workplace sexism more often than men, and they must demonstrate that they are as capable as men to perform their jobs. Furthermore, women often obtain lower wages. Thus, results on perceived occupational stress for women working from home are consistent with current literature (see, among others: European Parliament, 2021).

Table 2: Estimated probability to be *never stressed at work*

	South		North	
	Non-home-based	Home-based	Non-home based	Home-based
Men	0.121 (0.005)	0.080 (0.005)	0.096 (0.005)	0.062 (0.004)
Women	0.095 (0.004)	0.061 (0.004)	0.074 (0.004)	0.046 (0.004)

Table 3: Estimated probability to be *always stressed at work*

	South		North	
	Non-home-based	Home-based	Non-home based	Home-based
Men	0.080 (0.004)	0.120 (0.007)	0.102 (0.005)	0.150 (0.008)
Women	0.103 (0.004)	0.151 (0.008)	0.130 (0.006)	0.186 (0.095)

Concluding remarks

Although our study focuses on a pre-pandemic era, it could provide useful insights on gender-based differences in the perception of occupational stress. The issue of working from home has become of most concerns since it has been the "new normal" during the last two years. However, for most employees working remotely has not been a choice, but a necessity imposed by the pandemic. By contrast, the present analysis focuses on a period when working from home was quite unusual, and most likely an option. This could help to explain our results on caregiving, caring for children and house working activities that do not seem to influence respondents' experience of work-related stress.

Interestingly, given the characteristics of the welfare settings in the Northern Europe, our results show that women living in a Northern country report they perceive

work-related stress “always” more often than those living in the South. Such evidence may be interpreted considering the greater awareness of psycho-social risk factors at work in Northern countries, where the prospects of social protection are likely to be stronger. As a matter of fact, formal care services for elderly and kids are more efficient in the North as compared to the South of Europe, whereas in the latter countries’ norms concerning intergenerational responsibilities are still stronger and this could make women participation to the labour market more demanding and stressful. Therefore, as it is also the case for life satisfaction and other related topics, it can be assumed that the expectations of the respondents play an important role (Russell et al., 2018; Nappo, 2020).

In the light of changing working conditions, further research is necessary to examine the effect of female domestic (unpaid) work with respect to their main paid work, also considering the effects of geographical and welfare differences. In addition, more investigation is needed with respect to the distinction between employees and self-employed respondents.

References

1. Agresti, A.: Analysis of ordinal categorical data. Wiley, Hoboken, N.J. (2010)
2. Chung, H., Van der Horst, M.: Women’s employment patterns after childbirth and the perceived access to and use of flexitime and teleworking. Human Relations, Studies towards the integration of the social sciences, **71**(1), 47-72 (2018)
3. Del Boca, D., Oggero, N., Profeta, P., Rossi, M.: Women’s and men’s work, housework and childcare, before and during COVID-19. Review of Economics of the Household, **18**, 1001-1017 (2020)
4. Eurofound and the ILO (International Labour Office): Working anytime, anywhere: the effects on the world of work. Publications Office of the European Union, Luxembourg; the International Labour Office (ILO), Geneva, Switzerland (2017)
5. European Parliament: The impact of teleworking and digital work on workers and society. Samek Lodovici, M. et al. (eds.) Publication for the Committee on Employment and Social Affairs, Policy Department for Economic, Scientific and Quality of Life Policies (IPOL). Luxembourg (2021)
6. Messenger, J.: Conclusions and recommendations for policy and practice. In Telework in the 21st century, Messenger, J. (ed), International Labour Organisation (ILO), Geneva (2019)
7. Nappo, N.: Job stress and interpersonal relationships, cross country evidence from the EU15: a correlation analysis. BMC Public Health, **20,** 1143 (2020)
8. Pascucci, T., Hernàndez Sanchèz, B., Sanchéz Garcia, J.C.: Being stressed in the family or married with work? A literature review and clustering of work-family conflict. European Journal of Management and Business Economics, doi: 10.1108/EJMBE-06-2021-0191 (2021)
9. Piccolo, D., Simone, R.: The class of CUB models: statistical foundations, inferential issues and empirical evidence. Statistical Methods and Application, **28**, 389-435 (2019)
10. Repetti, R. L., Matthews, K. A., Waldron, I.: Employment and Women’s Health: Effects of Paid Employment on Women’s Mental and Physical Health. American Psychologist, **44**, 1394-1401. (1989)
11. Russell, H, Bertrand, M, Watson, D., Éamonn, F.: Job stress and working conditions: Ireland in comparative perspective-an analysis of the European working conditions survey. Research series, economic and social research institute (ESRI), RS84, Dublin (2018)
12. Tutz, G.: Regression for categorical data, Cambridge University Press, Cambridge (2012)

Investigating effects of air pollution on health: a challenge for statisticians

Investigating effect of air pollution on health via Spatial-Resolution Varying Coefficient Models

Garritt L. Page and Massimo Ventrucci

Abstract We focus on observational studies of environmental epidemiology where the goal is to estimate the effect of an exposure variable, such as air pollution, on a health outcome using spatial area-level data. We describe a novel framework introduced in [1] seeking estimation of the exposure effect at different spatial resolution and show how these methods behave under different types of conditionally autoregressive spatial models defined by the user. We illustrate the methods in a study on the association between COVID-19 mortality and air pollution.

Abstract *Uno degli obiettivi negli studi osservazionali di epidemiologia ambientale quello di stimare l'effetto dell'esposizione ad inquinanti sulla salute a partire da dati aggregati a livello areale. Descriveremo una nuova classe di modelli introdotti in [1], che permettono di stimare l'effetto dell'esposizione a diverse risoluzioni spaziali, e studieremo la loro performance per diverse specificazioni di modelli condizionali autoregressivi spaziali. Illustreremo il metodo su un caso studio riguardante l'associazione fra inquinamento atmosferico e mortalit dovuta a COVID-19.*

Key words: Conditionally autoregressive prior, Spatial confounding, Spatial causal inference

1 Background

A fundamental task in environmental epidemiology is to estimate the effect of a treatment variable (or exposure variable) on a health-related response variable. It is

Garritt L. Page
Department of Statistics, Brigham Young University, USA e-mail: page@stat.byu.edu

Massimo Ventrucci
Department of Statistical Sciences, University of Bologna, Italy e-mail: massimo.ventrucci@unibo.it

now common for environmental and epidemiological studies to be spatially varying in addition to being observational. On the one hand, these studies are conveniently carried out based on routinely collected data at the area-level, e.g. counts of cases/deaths and average air pollution levels in administrative areas. On the other hand, the observational nature of the study makes it challenging to take into account relevant (unobserved) confounding variables, hence putting into question the reliability of the obtained effect estimates.

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{X} = (X_1, ..., X_n)^T$ and $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ be, respectively, the count of cases/deaths, the exposure, and the unobserved confounder at spatial units $1, \ldots, n$. Spatial generalized linear models introduce a link function g such that $g\{\mathrm{E}(Y_i \mid X_i, Z_i)\} = \theta_i = \beta_0 + \beta_x X_i + \beta_z Z_i$. Letting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ we have that

$$\boldsymbol{\theta} = \beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \beta_z \boldsymbol{Z}. \tag{1}$$

If $\boldsymbol{Z}$ is observed and the model correct, then it is straight forward to estimate the causal effect β_x. In many real case studies $\boldsymbol{Z}$ is unobserved. A common recipe in these types of studies is to add a random effect to the spatial generalized linear model just described, where it is hoped that the spatial random effect will account for unmeasured confounders and perform the desired adjustment. Recent works have shown that adding random effects can instead lead to biased estimates due to so-called spatial confounding [2, 3].

1.1 Spectral adjustment for spatial confounding

We are interested in cases where the $\boldsymbol{Z}$ acts as a spatial confounder, in the sense that it is a spatially varying factor that influences both the treatment and response. We follow Guan et al. [1] who propose modeling the exposure and unmeasured confounder in the spectral domain, which permits deriving the coherence function and determine the assumptions necessary to establish a causal interpretation of exposure. Under this spectral framework, removing spatial confounding bias is possible provided the *unconfoundedness at high-frequencies* assumption, which states that exposure $\boldsymbol{X}$ and confounder $\boldsymbol{Z}$ are independent at the small spatial-resolution scale, holds. Along these lines, Guan et al. [1] introduce a spatial-resolution varying coefficient framework for area-level data. They focus on modelling jointly the spectral projections of $\boldsymbol{X}$ and $\boldsymbol{Z}$ using the conditionally autoregressive (CAR) spatial model proposed in Leroux et al. [4], that we will denote hereafter as *Leroux*. For instance, under Leroux

$$\boldsymbol{Z} \sim \text{Normal}\left(\mathbf{0}, \sigma_z^2 \boldsymbol{\Gamma}\left[(1-\lambda_z)\boldsymbol{I}_n + \lambda_z \boldsymbol{W}\right]^{-1} \boldsymbol{\Gamma}^T\right) \tag{2}$$

where σ_z^2 is the variance, λ_z is the spatial smoothing parameter and $\boldsymbol{\Gamma W \Gamma}^T = \boldsymbol{R}$ is the spectral decomposition of the structure matrix $\boldsymbol{R}$, specifying adjacency relationships between regions. Matrix $\boldsymbol{\Gamma}$ contains eigenvectors and $\boldsymbol{W}$ is a diagonal matrix

with eigenvalues $\omega_1, \ldots, \omega_n$. Then $\boldsymbol{X}^* = \boldsymbol{\Gamma}'\boldsymbol{X}$ and $\boldsymbol{Z}^* = \boldsymbol{\Gamma}'\boldsymbol{Z}$ project $\boldsymbol{X}$ and $\boldsymbol{Z}$ into the spectral domain. Guan et al. [1] then proceed to model X_i^* and Z_i^* jointly using a Gaussian distribution with covariance matrix informed by the Leroux model. Then marginalizing over $\boldsymbol{Z}^*$ and projecting back into the spatial domain produces

$$\boldsymbol{\theta} \mid \boldsymbol{X} \sim \text{Normal}\left(\beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^T\boldsymbol{X}, \sigma_z^2 \boldsymbol{\Gamma}\left[(1-\lambda_z)\boldsymbol{I}_n + \lambda_z \boldsymbol{W}\right]^{-1}\boldsymbol{\Gamma}^T\right). \quad (3)$$

The term $\boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^T\boldsymbol{X}$ in (3) adjusts for missing spatial confounders with

$$\boldsymbol{A} = \text{diag}\left(\alpha(\omega_1), \ldots, \alpha(\omega_n)\right),$$

where $\alpha(\omega_k), k = 1, \ldots, n$ are the *adjustment factors*. The terms $\alpha(\omega_k)$'s are functions of the Leroux model parameters (i.e. $\lambda_z, \lambda_x, \sigma_z, \sigma_x$) and the eigenvalues $\omega_k, k = 1, \ldots, n$; see [1] sec 4 for details. Given that $\boldsymbol{A}$ depends on the eigenvalues, the exposure effect $\beta(\omega) = (\beta_x + \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^T)\boldsymbol{X}$ can be interpreted as the effect of exposure $\boldsymbol{X}$ as a function of spatial resolution ω.

1.2 Spatial-resolution varying coefficient models

Spatial-resolution varying coefficient models can be defined by modelling the adjusting factor from (3) as $\boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^T = \sum_{l=1}^{L} \boldsymbol{Z}_l b_l$, where $\boldsymbol{Z}_l = \boldsymbol{\Gamma}\boldsymbol{B}_l\boldsymbol{\Gamma}^T\boldsymbol{X}$, $\boldsymbol{B}_l$ is a diagonal matrix with spline basis functions, $\{B_l(\omega_1), ..., B_l(\omega_n)\}$ and $\boldsymbol{b} = (b_1, \ldots, b_L)$ the associated spline coefficients. We fit this model by assuming a random walk prior on the spline coefficients, $\boldsymbol{b} \sim RW(\tau)$, with a penalized complexity prior on τ following [5]. We obtain the curve $\hat{\beta}(\omega_k), k = 1, \ldots, n$, representing the estimated effect of exposure for varying spatial-resolutions $\omega_1, \ldots, \omega_k$. Under the assumption of unconfoundedness at small spatial-resolutions (i.e. large eigenvalues), we can take $\hat{\beta}(\omega_n)$ as the adjusted estimate unaffected by spatial confounding.

Model (3) can be extended to areal data models other than Leroux, including the simple intrinsic CAR [6] hereafter denotes as *ICAR*, and the Besag York and Mollié model [7] with the intuitive parametrization proposed by Dean et al. (2001) [8, 9], hereafter denoted as *Dean*. We therefore extend the spectral framework by [1] to these other CAR specifications, to understand whether the ability to reduce spatial confounding may be driven by the type of CAR model adopted by the user.

2 Application

Wu et al. [10] noticed that many co-morbidities associated with COVID-19 had connections to being exposed to higher concentrations of ambient fine particular matter ($PM_{2.5}$). Due to this, they conducted a study to determine if an increase in $PM_{2.5}$ resulted in a higher COVID-19 mortality rate. They found that an increase of

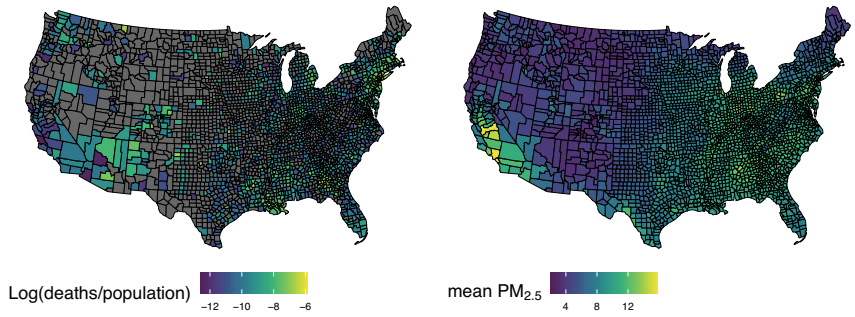


Fig. 1 $PM_{2.5}$ exposure and COVID-19 mortality by US county: (left) the log COVID-19 mortality rate (i.e., log(deaths/population)) through May 12, 2020 (counties with no deaths are shaded gray) and (right) Average $PM_{2.5}$ ($\mu g/m^3$) over 2000-2016.

$1\ \mu g/m^3$ in ambient fine particulate matter ($PM_{2.5}$) is associated with a 15% increase in the COVID-19 mortality rate. The response variable is the cumulative COVID-19 mortality counts through May 12, 2020 for US counties. County-level exposure to $PM_{2.5}$ was calculated by averaging results from an established exposure prediction model for years 2000-2016. This resulted in mortality counts and $PM_{2.5}$ measures for $n = 3109$ counties, see Fig. 1). The long-term average $PM_{2.5}$ is the highest in the Eastern US and California, while the mortality response is the highest in the New York, Los Angeles and Seattle areas. The average $PM_{2.5}$ rate is a smoother spatial process than mortality, likely because the $PM_{2.5}$ exposure estimates are generated from predictive models. In addition to $PM_{2.5}$ exposure, 20 potential confounding variables (e.g., the percentage of the population at least 65 years old) are included in our modeling (see [10] for the complete set of potential confounding covariates).

2.1 Modelling and results

Let's denote Y_i as the number of deaths attributed to COVID-19, E_i as the population, X_i as the average $PM_{2.5}$ and $\boldsymbol{C}_i$ as the vector of 20 known confounding variables, for county i. Similar to [10], we fit a Negative-Binomial regression model $Y_i \mid X_i, Z_i, \boldsymbol{C}_i \overset{indep}{\sim} \text{NegBin}\{r_i, p_i\}$, where r_i and p_i are, respectively, the size parameter and the probability of success in each trial. Under this model the mean is $\text{E}(Y_i \mid X_i, Z_i, \boldsymbol{C}_i) = \lambda_i = r_i(1-p_i)/p_i$. We parameterize the model in terms of λ_i and the over-dispersion parameter r_i. The mean is linked to the linear predictor as $\log(\lambda_i) = \log(E_i) + \theta_i$ where $\theta_i = \beta_0 + \beta_x X_i + Z_i + \boldsymbol{C}_i'\boldsymbol{\beta}_c$ and the offset term E_i is the county population and $\boldsymbol{\beta}_c$ is a vector of regression coefficients associated with the

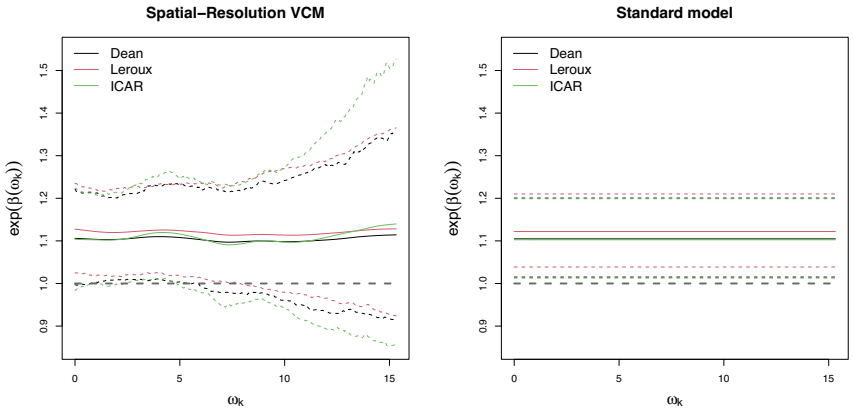


Fig. 2 The left plot contains estimated coefficient values as a function of spatial resolution based on the spatial-resolution varying coefficient model for the Dean, Leroux, and ICAR areal data models. The right plot is the standard spatial generalized linear mixed model approach.

confounding variables. We follow the spectral approach and model $(\boldsymbol{X}^*, \boldsymbol{Z}^*)$ jointly by using different CAR assumptions on $\boldsymbol{X}$ and $\boldsymbol{Z}$, namely Leroux, Dean and ICAR. All models were fitted using R-INLA [11].

Fig. 2 reports the posterior mean and 95% credible bands of $\exp(\beta_x)$, i.e. the mortality rate ratio associated to an increase of 1 $\mu g/m^3$ of $PM_{2.5}$, both for the Spatial-resolution VCM and the standard spatial generalized linear mixed model approach. We can see a general agreement between the Leroux and Dean model, while the credible bands for the ICAR are quite a bit larger. The point estimates of β_x from all the models substantially agree with the standard analysis (right plot in Fig. 2). However, different from the standard analysis, the estimated effect at small spatial-resolution (i.e. large eigenvalues) show large uncertainty and overlaps with no effect (grey dashed line).

3 Discussion

It appears in the case study considered by Wu et al. [10] that the choice of CAR specification employed impacts the confounding adjustment based on the spectral methods developed in Guan et al. [1]. The reasonableness of assumptions under each of the CAR models is the topic of future research. Additionally, it seems plausible that the priors assumed on the parameters controlling spatial residual variability (i.e. σ_z, λ_z) may have an impact as well. Thus, including good prior information for these parameters is important. Since λ_z in the Dean model has an intuitive interpretation as the proportion of spatially structured variance over the the total residual variance,

it seems that eliciting prior information from experts under this model may be more straightforward.

References

1. Guan, Y., Page, G.L., Reich, B.J., Ventrucci, M., Yang, S.: A spectral adjustment for spatial confounding. arXiv preprint arXiv:2012.11767 (2020)
2. Hodges, J.S., Reich, B.J.: Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. The American Statistician **64**, 325-334 (2010)
3. Page, G.L., Liu, Y., He, Z., Sun, D.: Estimation and prediction in the presence of spatial confounding for spatial linear models. Scandinavian Journal of Statistics, **44**, 780-797 (2017)
4. Leroux B.G., Lei X., Breslow N.: Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Statistical Models in Epidemiology, the Environment, and Clinical Trials. (2000)
5. Franco-Villoria, M., Ventrucci, M., Rue, H.: A unified view on Bayesian varying coefficient models. Electronic Journal of Statistics, **13**, 53345359 (2019)
6. Besag, J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of the Royal Statistical Society Series B, **36**, 192225 (1974)
7. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics **43**, 120 (1991)
8. Dean, C., Ugarte, M., Militino, A.: Detecting interaction between random region and fixed age effects in disease mapping. Biometrics **57**, 197202 (2001)
9. Riebler, A., Srbye, S., Simpson, D., Rue, H.: An intuitive Bayesian spatial model for disease mapping that accounts for scaling Statistical Methods in Medical Research, **25** (2016)
10. Wu, X., Nethery, R., Sabath, M., Baun, D., Dominici, F.: Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. Science Advances, **6**, (2020)
11. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B, **71**, 319-392 (2009)

A statistical framework for evaluating the health effects of PM sources

Un approccio statistico per valutare gli effetti sulla salute delle fonti di particolato

Monica Pirani, Georges Bucyibaruta, Gary Fuller, David Green, Anja Tremper, Christina Mitsakou, Marta Blangiardo

Abstract The association between exposure to particulate matter (PM) and health outcomes is well established. Nevertheless, PM is a mixture of different sources, which might have a different toxicity. Identifying ambient PM sources is key for developing strategies to reduce PM through targeted actions. Current methods to identify sources of particulate pollution typically require *a priori* specification of the number of sources and do not include information on covariates in the source allocations. In this study, we propose a Bayesian nonparametric approach to overcome these limitations. We specify a probabilistic source apportionment model, and we will use a regression model to investigate the link between the source contributions and hospital admissions for respiratory diseases in London between 2012 and 2016.

Abstract *L'associazione tra esposizione a particolato (PM) e salute è ben stabilita. Tuttavia, il PM è una mistura di diverse fonti, che potrebbero avere una diversa tossicità. Identificare le fonti di PM è fondamentale per lo sviluppo di strategie per ridurre il PM attraverso azioni mirate. I metodi attuali per l'identificazione di fonti di inquinamento da particolato in genere richiedono una specificazione* a priori *del numero di fonti e non includono informazioni sulle covariate nell'allocazione delle fonti. In questo studio, proponiamo un approccio Bayesiano non parametrico per superare queste limitazioni. Specifichiamo un modello probabilistico di ripartizione*

Monica Pirani, Department of Epidemiology and Biostatistics, Imperial College London (UK); email: monica.pirani@imperial.ac.uk
Georges Bucyibaruta, Department of Epidemiology and Biostatistics, Imperial College London (UK); email: g.bucyibaruta@imperial.ac.uk
Gary Fuller, Environmental Research Group, Imperial College London (UK); email: g.fuller@imperial.ac.uk
David Green, Environmental Research Group, Imperial College London (UK); email: d.green@imperial.ac.uk
Anja Tremper, Environmental Research Group, Imperial College London (UK); email: anja.tremper@imperial.ac.uk
Christina Mitsakou, UK Health Security Agency; email: christina.mitsakou@phe.gov.uk
Marta Blangiardo, Department of Epidemiology and Biostatistics, Imperial College London (UK); email: m.blangiardo@imperial.ac.uk

delle fonti e utilizzeremo un modello di regressione per studiare il legame tra le diverse fonti di particolato e i ricoveri ospedalieri per malattie respiratorie a Londra tra il 2012 e il 2016.

Key words: Bayesian inference, Dependent Dirichlet process, Source apportionment

1 Introduction

The detrimental effects of exposure to ambient particulate matter (PM) on health outcomes are well established. Nevertheless, as the composition of PM is complex, recent studies have questioned and tried to figure out whether mixture of contaminants from different sources can have different harmful effects (e.g. Cassee et al. 2013; Hackstad et al. 2013; Pirani et al. 2015; Samoli et al. 2016); this makes the identification of pollutant sources a key aspect, in order to implement effective policies to improve air quality and population health.

Compositional data with information on the different chemical components within PM concentration can be expensive and not available at every monitoring site. As an alternative, measurements of particle number concentration (PNC) and the related particle number size distribution (PNSD) have received much attention and have been recently considered as a way to investigate PM sources (Hopke, 2022). Typically, this involves considering sizes spanning across both the ultrafine ($\leq$ 100nm diameter) and fine (100 - 2500nm diameter) particle ranges.

Working with PNSD means dealing with a large number of correlated variables; where, the range of sizes is split into a large number of bins and the number of particles in each bin is calculated. These bins are typically correlated, and the main statistical challenge consists in reducing the high dimensional and correlated data into a smaller number of sources. This analysis is known as *source apportionment* (SA; Krall and Chang, 2019). Traditionally, methods for the SA problem have been dominated by two approaches (Viana et al. 2008): source-oriented deterministic models and receptor models, and we will focus on the latter as our proposed method fits in that framework. Commonly, receptor models for SA decompose ambient concentrations of pollutants into components based on how they co-vary, then associating the components with different source labels. Within this framework, positive matrix factorization (PMF; Paatero and Tapper, 1994) is widely used for pollution SA modelling. PMF requires *a priori* specification of the number of factors to be output by the model. However, there are no objective criteria to select this number. Additionally, there is not a principled way of accounting for the uncertainty in the source allocation, and the method requires that the dataset is complete (typically missing data are removed or imputed). Finally, similar to most dimension reduction techniques, PMF relies on the assumption that the source contributions are independent over time. However, this may not be appropriate and temporal dependence can

exist. This dependence could be (partially or completely) explained by covariates, particularly related to meteorology (e.g. Pineda Rojas et al. 2020).

To overcome these limitations, we propose a Bayesian nonparametric modelling framework, which allows to account for temporal dependencies and concomitant processes (e.g. meteorology) in the identification of the sources. Then, we evaluate the impact of the sources on the health of vulnerable population groups.

The Bayesian approach is naturally placed for mixture models, catering for temporal dependencies, able to deal with sparsely sampled data and to model multiple uncertainties. The inference is performed through Markov Chain Monte Carlo (MCMC) methods in the `R` software using `Nimble` probabilistic programming language. To fill the methodological gaps in source characterisation and health effect evaluation, the work pursues the following main steps:
1. Develop mixture models in a Bayesian nonparametric framework which has received a lot of attention in the machine learning community for unsupervised tasks (Murphy, 2012). In particular, we model source contribution using a Dirichlet process (DP; Ferguson 1973) as a prior for source profiles, which allows us to estimate the number of components that contribute to particle concentration rather than fixing this number beforehand. To better characterise these, we also include meteorological covariates via a flexible Gaussian kernel.
2. The apportioned sources are then linked to health outcomes in vulnerable population through a regression model allowing a comparative assessment of the extent to which variations in the apportionment contributed to variability in the source-specific health outcome, which will also simultaneously estimate the effect of gaseous pollutants and of other time-varying confounding factors.

2 Methods

We specify the receptor model for particle size concentration distribution in a probabilistic perspective, and we make it data driven by the use of a dependent Dirichlet processes (Quintana et al. 2022). In particular, our model formulation is based on the approach proposed by Baerenbold et al. 2022. For the p^{th} bin ($p = 1, \ldots, P$) and t^{th} time point ($t = 1, \ldots, T$), we model the concentration $y_{p,t}$ as follows:

$$\log(y_{p,t}) \sim \mathcal{N}\left(\log(\mu_{p,t}), \sigma_p\right)$$
$$\mu_{p,t} = \sum_k \lambda_{p,k} f_{k,t}$$

where σ_p represents the size-specific measurement error, $\lambda_{p,k}$ is the source profile that provides the proportion of particles from source k in size bin p, and $f_{k,t}$ is the source contribution. We specify $f_{k,t}$ as $f_{k,t} = s_{k,t} c_t$, where $s_{k,t}$ is the proportion of the total particle concentration at time t contributed by source k, and c_t is the total particle concentration at time t. We model the parameter c_t as being normally distributed on the log-scale with a common mean μ_c and standard deviation σ_c).

Then, for $s_{k,t}$ we assume a kernel stick-breaking prior (Dunson and Park, 2007). Generically, letting F denote a random probability measure, the prior is formulated as:

$$F_t = \sum_{k=1}^{\infty} s_{k,t}\delta_{\theta_k}$$

where δ_{θ_k} is the Dirac measure (point mass) at θ_k and $\theta_k \overset{iid}{\sim} G_0$, where G_0 is the base measure (i.e. the expected value of the process). The vector $\mathbf{s}$ are the mixture weights, representing the probability of an observation coming from source $1, \ldots, k, \ldots, K \longrightarrow \infty$. In the standard DP mixture model, a stick-breaking prior (Sethuraman, 1994) is usually specified for the mixing weights, which intuitively consists of breaking pieces off from a stick of unit length (to represent the probability scale), where the breakpoints, say $\mathbf{V}$, are randomly sampled from the Beta distribution. In order to acknowledge the intrinsic order of the data in time, we allow dependence among nearby measurements by specifying that the probabilities weights vary temporally, obtaining a flexible time-dependent partitioning. In order to do so, external variables (such as wind speed and direction in our application) are included to model covariate dependent weights. This is done through a kernel stick-breaking process, where kernel functions are introduced to allow $\mathbf{V}$ to change with covariates, inducing a smoothing effect. To ensure flexibility while retaining computational tractability, we will use a truncated DP (Ishwaran and Zarepour, 2000) and assume a maximum number K; this will provide a good approximation of a DP, but avoiding a large number of unnecessary cluster parameters.

3 Data description

We apply the model to apportion particle number size distribution measured in London (UK) and evaluate the short-term effects of the apportioned sources on the health of population groups in a time-series framework. In building the probabilistic model for SA, we use measurements of particle sizes and wind speed/wind direction obtained from an urban background monitoring site located in North Kensington (central London). We consider hourly data covering the period 2012-2016, with the particle size distribution measured in different size bins ranging from 16.55 to 604.3nm. Time-series concentrations of other air pollutants are also measured such as total oxides of nitrogen (NO_X), nitrogen dioxide (NO_2), total PM, base fraction of PM (PMFB), and volatile fraction of PM (PMFR), carbon (black carbon measured in infrared transmission (CBLK), black carbon measured in ultra-violet transmission (CBUV), and carbon from wood burning (CWOD)). The health data are available from the Hospital Episode Statistics registry within the Small Area Health Statistics Unit (SAHSU) at Imperial College London.

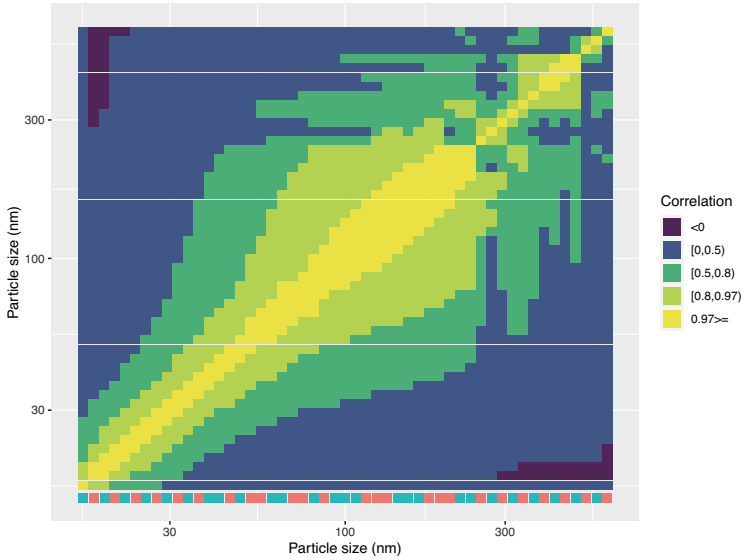


Fig. 1 Correlation matrix between data in original size bins. Resulting aggregated bins are indicated at the bottom by colored segments (preliminary results).

4 Data pre-processing and preliminary results

Here we present some preliminary results for a reduced number of time-points, selecting the summer months. We reduced the number of size bins for computational feasibility. Consecutive bins were aggregated as follows. For the p^{th} bin we computed the correlation between it and q consecutive bins, $q = 1, 2, \ldots$, when this correlation was below a certain threshold, τ, we summed the concentrations and considered this as a new bin where the value represents the total particle concentration across both size bins. Fig. 1 shows the correlation matrix for the different size bins. We show the resulting binned segmentation (colour coded) at the bottom. Setting a threshold of $\tau = 0.97$ leads to 26 distinct size bins. Fig. 2 shows the profiles of five preliminary sources that are estimated by the model to contribute to the total concentrations. The profiles cover the entire range of sizes and all the five sources are clearly distinct. Corresponding wind kernels are shown in Fig. 3 and suggest that the wind plays a role in the source attribution. Note that due to finite approximation of the dependent Dirichlet process, the model estimates only wind kernels for the first $K - 1$ sources.

The extension of the model to the entire dataset (2012-2016), the source labelling, and the link with respiratory health outcomes adopting a two-stages approach are currently on-going.

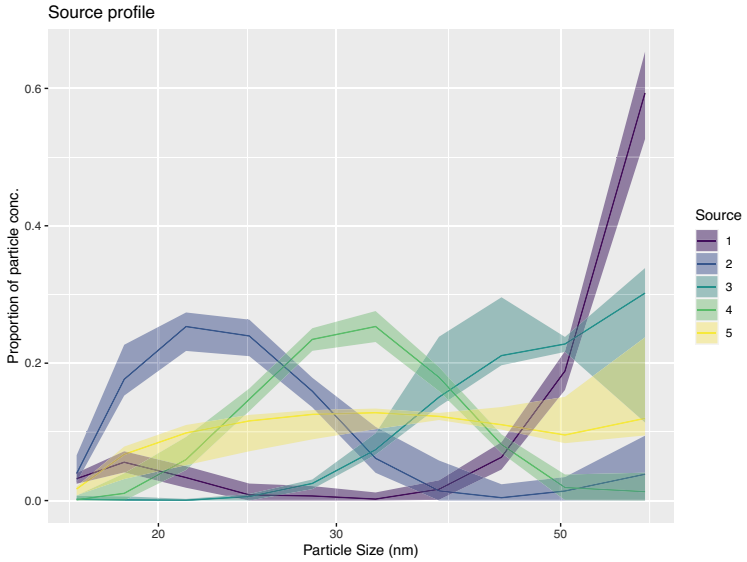


Fig. 2 Particle size distribution for the five sources identified by the model for summer period (preliminary results).

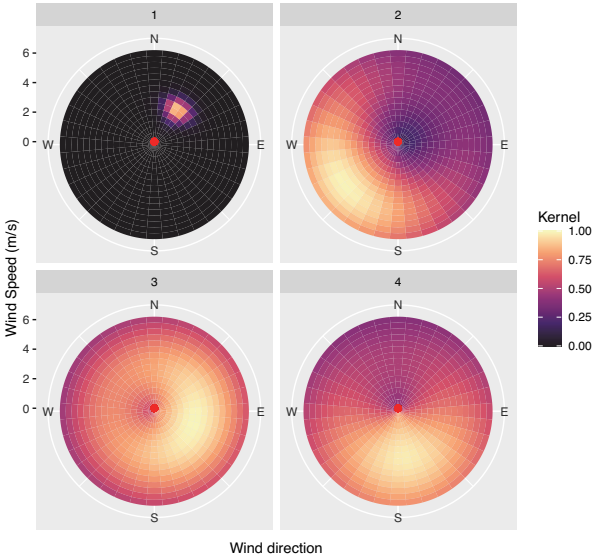


Fig. 3 Wind kernels for $K - 1$ sources (preliminary results).

References

1. Baerenbold, O., Meis M., Martínez-Hernández, I., Euán, C., Burr, W.S., Temper, A., Fuller, G., Pirani, M., Blangiardo, M.: A dependent Bayesian Dirichlet Process model for source

apportionment of particle number size distribution. Submitted to Environmentrics (2022)
2. Cassee, F.R., Héroux, M.E., Gerlofs-Nijland, M.E. and Kelly, F.J.: Particulate matter beyond mass: recent health evidence on the role of fractions, chemical constituents and sources of emission. Inhal Toxicol. **25**, 802-812 (2013)
3. Dunson, D.B., Park, J.: Kernel stick-breaking processes. Biometrika. **95**, 307-323 (2007)
4. Ferguson, T.: A Bayesian analysis of some non-parametric problems. Ann. Stat. **1**, 209-230 (1973)
5. Hackstad, A.J., Peng, R.D.: A Bayesian multivariate receptor model for estimating source contributions to particulate matter pollution using national databases. Environmetrics. **25**, 513–527 (2014)
6. Hopke, P.K., Feng, Y. and Dai, Q.: Source apportionment of particle number concentrations: A global review. Sci. Total Environ. **819**, 153104, (2022)
7. Ishwaran, H., Zarepour, M.: Markov Chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. Biometrika. **87**, 371-390 (2000)
8. Krall, J., Chang, H.: Statistical methods for source apportionment. In: Gelfand, A.E., Fuentes, M., Hoeting, J.A., Lyttleton Smith, R. (eds.) Handbook of Environmental and Ecological Statistics, pp. 523–546. Chapman and Hall/CRC (2019)
9. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press (2012)
10. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. **5**, 111–126 (1994)
11. Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R.W., Fuller, G.W.: Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. Environ. Int. **79**, 56-64 (2015)
12. Quintana, F.A., Müller, P., Jara, A., MacEachern, S.N.: The dependent Dirichlet process and related models. Stat. Sci. **37**, 24-41 (2022)
13. Rojas, A.L.P., Borge, R., Mazzeo, N.A., Saurral, R.I., Matarazzo, B.N., Cordero, J.M., Kropff, E.: High PM_{10} concentrations in the city of Buenos Aires and their relationship with meteorological conditions. Atmos. Environ. **241**, 117773 (2020)
14. Samoli, E., Atkinson, R.W., Analitis, A., Fuller, G.W., Beddows, D., Green, D.C., Mudway, I.S., Harrison, R.M., Anderson, H.R., Kelly, F.J.: Differential health effects of short-term exposure to source-specific particles in London, UK. Environ. Int. **97**, 246-253 (2016)
15. Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)
16. Viana, M., Kuhlbusch, T.A., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A.S., Hueglin, C.: Source apportionment of particulate matter in Europe: A review of methods and results. J Aerosol Sci. **39**, 827–849 (2008)

Adjusting for Unmeasured Spatial Confounding Through Shrinkage Methods

Aggiustamento per fattori confondenti spaziali non misurati tramite metodi di shrinkage

Valentini Pasquale, Schmidt Alexandra M., Zaccardi Carlo and Ippoliti Luigi

Abstract This paper aims to discuss the problem of (unmeasured) spatial confounding, which arises when possible confounders result unmeasured and not included in the model. To adjust for confounding, we propose a semi-parametric regression model based on principal splines under the Bayesian paradigm. We assume spike and slab priors on a subset of regression coefficients in order to achieve dimensionality reduction and reduce the confounding bias.

Abstract *Questo articolo intende trattare il problema della presenza di fattori confondenti spaziali (non misurati), che si manifesta quando le informazioni riguardanti talune variabili confondenti non sono disponibili. Al fine di aggiustare per i fattori confondenti, proponiamo un modello di regressione semiparametrico in ambito bayesiano che utilizza le "principal splines". Assumiamo le prior "spike and slab" su un sottogruppo di coefficienti di regressione affinché si ottenga la riduzione della dimensionalità e del "confounding bias".*

Key words: Bayesian, spatial, confounding, shrinkage, spike and slab, principal splines

---

Pasquale Valentini
University G. d'Annunzio, Chieti-Pescara, Department of Economics, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: pvalent@unich.it

Alexandra M. Schmidt
McGill University, Department of Epidemiology, Biostatistics and Occupational Health, 2001 McGill College Avenue, Montreal, QC, Canada, e-mail: alexandra.schmidt@mcgill.ca

Carlo Zaccardi
University G. d'Annunzio, Chieti-Pescara, Department of Economics, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: carlo.zaccardi@unich.it

Luigi Ippoliti
University G. d'Annunzio, Chieti-Pescara, Department of Economics, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: luigi.ippoliti@unich.it

1 Introduction

In environmental epidemiology, the effect of an exposure on a health outcome is of main interest. Researchers usually apply the tools provided by spatial statistics because the variables have a spatial structure and the neighborhood scheme among the sites should be accounted for. As an introductory example, consider the association between air pollution concentration and mortality counts, when both vary spatially. To estimate the effect size, the outcome is generally regressed on the exposure and a set of other variables that are correlated with both exposure and health outcome, such as temperature, humidity or socioeconomic status [12, 18]. These are known as spatial confounders if they also vary in space. Ideally, any confounder must be included in the regression model, but generally some of them can be unmeasured (i.e. no data are available), so residuals are no longer orthogonal, leading to biased estimators. This problem is known as *spatial confounding*[1] in the literature (see [12, 16] for example) and was firstly recognized by [3].

A rigorous discussion on the mathematical derivation of the bias induced by the absence of information about a confounder was provided by [16]. For point-referenced data, the author treated the exposure and the unmeasured confounder as Gaussian processes and showed that the confounding bias is only reduced when the unconfounded component of the exposure varies at a spatial scale smaller than that of the confounded component. A similar setup was followed by [15, 17]. The former discussed spatial confounding in the case of multilevel data with replications within each location and showed that, even if spatial correlation is absent, the problem still remains. The latter discovered through simulations that an increase of the correlation between exposure and unmeasured confounder can conceivably lead to an improvement of the prediction performance of the outcome model.

The simplest approach for dealing with spatial confounding is to include a spatial random effect (SRE) into the regression analysis, which is generally modeled as a conditional autoregressive (CAR) process in the case of lattice data, or as a continuous Gaussian process in the case of point-referenced data [1, 14, 21]. In the context of areal data, the difficulties of using an SRE for bias reduction were considered by [20] under different scenarios. As a result, the restricted spatial regression (RSR) model was proposed by [9] as a means to reduce the impact of the SRE on the exposure's effect size: this approach restricts the SRE to the orthogonal complement of the other regressors ("fixed effects"). RSR was extended to geostatistical data by [8], who suggested a posterior distribution for the exposure coefficient that do not let the respective credible interval to shrink substantially.

[1] More precisely, one should talk about *unmeasured spatial confounding*, since the term *spatial confounding* indicates the presence of confounders that vary in space. The problem does not exist as far as information about any spatial confounder is available.

RSR has been used by many to recover the unexplained spatial structure [8, 9, 10, 15, 19]. However, others have found this approach too restrictive: as mentioned by [16], the orthogonality assumption is rather strong as fixed and random effects are not orthogonal under generalized least squares (GLS) estimation. Furthermore, [11] demonstrated that RSR provides poorer inference performance when compared to the non-spatial model. Therefore, [24] proposed an approach for areal data based on structural equation modeling techniques, in order to estimate simultaneously an exposure and an outcome model. This translates in a removal of spatial information from both exposure and outcome variables, and in a subsequent regression between the residuals. On the other hand, in the geostatistical case, [4, 12, 16] considered the need of introducing splines to alleviate the spatial confounding bias.

2 The Model

Consider a spatial process $\{Y(\mathbf{s}_i) : \mathbf{s}_i \in \mathscr{S}\}$, where $\mathbf{s}_i$ is a spatial index variable within a spatial domain $\mathscr{S} \subseteq \mathbb{R}^2$. Assuming the presence of unmeasured confounders, the general regression model usually introduces an additional spatial error process, $g(\mathbf{s}_i)$, such that

$$Y(\mathbf{s}_i) = \beta_0 + \beta_x X(\mathbf{s}_i) + g(\mathbf{s}_i) + \varepsilon_y(\mathbf{s}_i), \qquad \varepsilon_y(\mathbf{s}_i) \overset{iid}{\sim} N(0, \sigma_y^2), \tag{1}$$

where $X(\mathbf{s}_i)$ denotes the exposure with unknown effect β_x.

2.1 An Explanatory Example

In order to better understand the sources of spatial confounding bias, we shall consider an example. An easy way to simulate point-referenced data is to assume that the n-dimensional vectors $\mathbf{X}$ and $\mathbf{g}$ follow a linear model of coregionalization (LMC) [23]. Consider the model from (1). Let $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \sigma_x^2 \mathbf{R}_{\phi_x})$ and $\mathbf{g} \sim N(\boldsymbol{\mu}_g, \sigma_g^2 \mathbf{R}_{\phi_g})$, where $\mathbf{R}_{\phi_x}$ and $\mathbf{R}_{\phi_g}$ are defined by parametric correlation function $\rho(|\mathbf{s} - \mathbf{s}'|; \phi)$. Thus, $\mathbf{X}$ and $\mathbf{g}$ are jointly normal with the following:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{g} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_g \end{pmatrix}, \begin{pmatrix} \sigma_x^2 \mathbf{R}_{\phi_x} & \delta \sigma_x \sigma_g \mathbf{R}_{\phi_x}^{1/2} \mathbf{R}_{\phi_g}^{1/2'} \\ \delta \sigma_x \sigma_g \mathbf{R}_{\phi_g}^{1/2} \mathbf{R}_{\phi_x}^{1/2'} & \sigma_g^2 \mathbf{R}_{\phi_g} \end{pmatrix} \right], \tag{2}$$

where $\delta \in (-1, 1)$ is the correlation between $\mathbf{X}$ and $\mathbf{g}$.

Figures 1–4 refer to simulations of $\mathbf{X}$ and $\mathbf{g}$ on a unit-square grid, using the exponential correlation function for $\mathbf{R}_{\phi_x}$ and $\mathbf{R}_{\phi_g}$ with ranges ϕ_x and ϕ_g, respectively.

A first source of bias is the correlation parameter. Figure 1 is obtained by setting $\delta = 0.7$ and $\phi_x = \phi_g = 0.2$, whereas Figure 2 depicts a scenario identical to the first one, but with $\mathbf{X}$ being independent of $\mathbf{g}$ (i.e., $\delta = 0$). By definition, this means that there is no spatial confounding effect in the latter case, thus one would expect the presence of spatial confounding bias only when δ is not null [16].

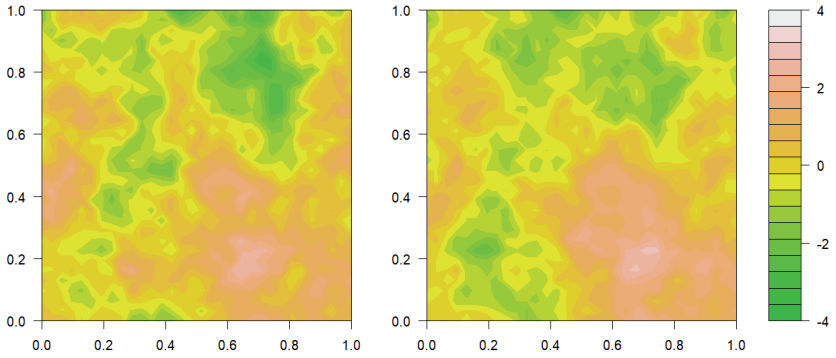


Fig. 1 Exposure (left) and spatial error process $g(\mathbf{s}_i)$ (right) are highly correlated ($\delta = 0.7$). Variables are scaled to have zero mean and unit variance.

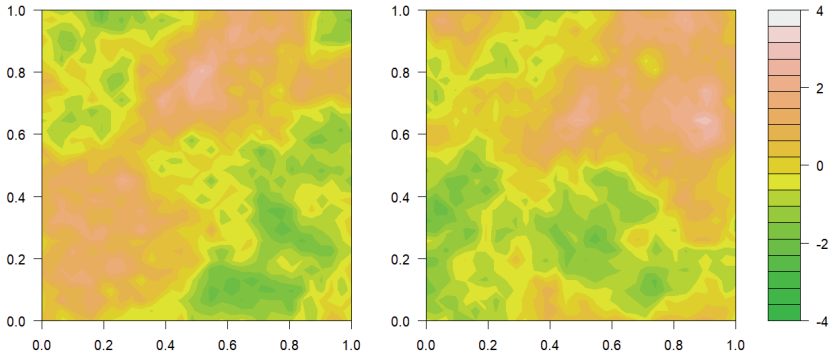


Fig. 2 Exposure (left) and spatial error process $g(\mathbf{s}_i)$ (right) are uncorrelated ($\delta = 0$). Variables are scaled to have zero mean and unit variance.

A second origin of bias is the relationship between spatial ranges. Figure 3 shows two processes such that ϕ_x is much greater than ϕ_g: the spatial error process $g(\mathbf{s}_i)$ and $\varepsilon_y(\mathbf{s}_i)$ are almost indistinguishable. In contrast, ϕ_x is smaller than ϕ_g in Figure 4. The expectation is that the introduction of an SRE component in the regression model would reduce the confounding bias only in scenarios similar to the one presented in Figure 4, because the exposure alone would only be able to explain a small part of the total variability in the outcome [16, 17].

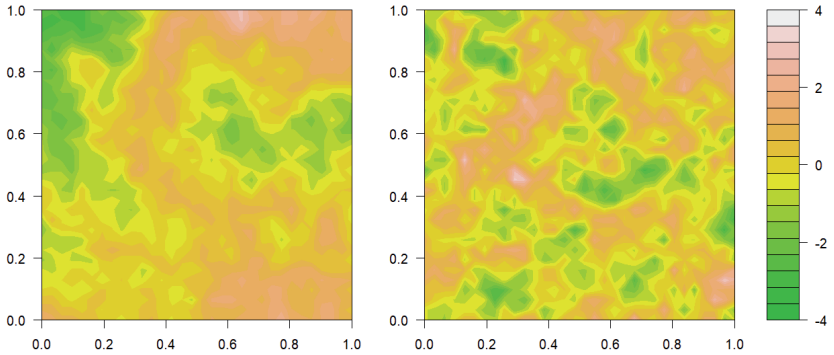


Fig. 3 The exposure (left) is generated using a large spatial range ($\phi_x = 0.5$), while the spatial error process $g(\mathbf{s}_i)$ (right) using a smaller one ($\phi_g = 0.05$). Variables are scaled to have zero mean and unit variance.

3 The Proposed Approach

In contrast to the SRE approach, where $g(\mathbf{s}_i)$ is assumed to follow either Gaussian or CAR process, to model the spatially varying terms in equation (1) we propose a spline representation [2, 5, 6, 22]. Under the Bayesian perspective, our approach is similar in spirit to that discussed by [12]. However, differently from [12], we assume that $g(\mathbf{s}_i)$ can be written as a finite expansion of *principal splines* [2, 5, 6, 22] as follows:

$$g(\mathbf{s}_i) = \mathbf{e}_i'\mathbf{B}\boldsymbol{\xi}\,, \quad i = 1,\ldots,n\,, \tag{3}$$

where $\mathbf{e}_i$ is the unit vector with 1 as the ith element, $\mathbf{B}$ is a matrix collecting a set of basis functions extracted from thin-plate splines, and $\boldsymbol{\xi}$ is a corresponding vector of expansion coefficients. The basis functions are defined as *principal* as they can be ordered in terms of their degrees of smoothness with higher-order functions cor-

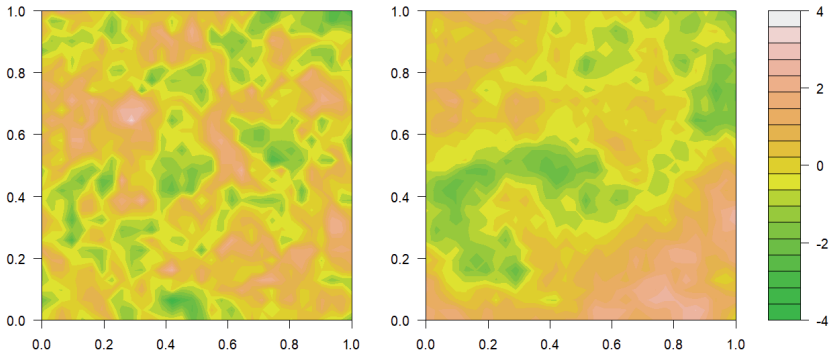


Fig. 4 The exposure (left) is generated using a small spatial range ($\phi_x = 0.05$), while the spatial error process $g(\mathbf{s}_i)$ (right) using a larger one ($\phi_g = 0.5$). Variables are scaled to have zero mean and unit variance.

responding to larger-scale features and lower-order ones corresponding to smaller-scale details, leading to a parsimonious representation of a (nonstationary) spatial covariance function with the number of basis functions representing different spatial variability and resolution. An advantage of our approach is that the proposed class of basis functions avoids the difficult knot allocation or scale selection problems commonly encountered in a spline framework.

To select the number of basis functions to be included in $\mathbf{B}$, different methods are possible. In [12], the number of bases is selected thanks to an information criterion evaluated on an outcome model without exposure. Here, we propose to impose spike and slab priors [7] on the basis coefficients $\boldsymbol{\xi}$ such that all possible models are embodied within a hierarchical formulation and basis selection is carried out model-wise.

A deeper discussion of each scenario presented in Figures 1–4, as well as the conditions under which our model is able to accommodate the statistical issues associated with spatial confounding, will be discussed in an extended version of this paper. In particular, the performance of the proposed model will be tested through both an extensive simulation study and applications to real data.

References

1. Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data (2nd ed.). Chapman and Hall/CRC, New York (2014)
2. Bookstein, F.L.: Morphometric Tools for Landmark Data — Geometry and Biology. Cambridge University Press, Cambridge (1992)
3. Clayton, D.G., Bernardinelli, L., Montomoli, C.: Spatial correlation in ecological analysis. Int. J. of Epidemiology. **22(6)**, 1193–1202 (1993)
4. Dupont, E., Wood, S.N., Augustin, N.: Spatial+: a novel approach to spatial confounding. Biometrics. Accepted/In press (2021)
5. Fontanella, L., Ippoliti, L., Valentini, P.: A Functional Spatio-Temporal Model for Geometric Shape Analysis. In: Torelli N., Pesarin F., Bar-Hen A. (eds.) Advances in Theoretical and Applied Statistics. Studies in Theoretical and Applied Statistics. Springer, Berlin, Heidelberg. (2013)
6. Fontanella, L., Ippoliti, L., Valentini, P.: Predictive functional ANOVA models for longitudinal analysis of mandibular shape changes. Biom. J. **61(4)**, 918–933 (2019) doi: 10.1002/bimj.201800228
7. George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Statistica sinica. 339–373 (1997)
8. Hanks, E.M., Schliep, E.M., Hooten, M.B., Hoeting, J.A.: Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. Environmetrics. **26(4)**, 243–254 (2015)
9. Hodges, J.S., Reich, B.J.: Adding spatially-correlated errors can mess up the fixed effect you love. The Am. Statistician. **64(4)**, 325–334 (2010)
10. Hughes, J., Haran, M.: Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. J. of the Royal Statist. Society: Series B. **75(1)**, 139–159 (2013)
11. Khan, K., Calder, C.A.: Restricted spatial regression methods: Implications for inference. J. Am. Statist. Assoc., 1–13 (2020)
12. Keller, J.P., Szpiro, A.A.: Selecting a scale for spatial confounding adjustment. J. of the Royal Statist. Society: Series A. **183(3)**, 1121–1143 (2020)
13. Mardia, K.V., Goodall, C., Redfern, E.J., Alonso, F.J.: The kriged Kalman filter. Test. **7(2)**, 217–282 (1998)
14. Marques, I., Kneib, T., Klein, N.: A multivariate Gaussian random field prior against spatial confounding. arXiv preprint (2021) arXiv:2106.03737
15. Nobre, W.S., Schmidt, A.M., Pereira, J.B.: On the effects of spatial confounding in hierarchical models. Int. Statist. Rev. **89(2)**, 302–322 (2021)
16. Paciorek, C.J.: The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Statist. Sci. **25(1)**, 107–125. (2010)
17. Page, G.L., Liu, Y., He, Z., Sun, D.: Estimation and prediction in the presence of spatial confounding for spatial linear models. Scandinavian J. of Statistics. **44(3)**, 780–797 (2017)
18. Peng, R.D., Dominici, F.: Statistical Methods for Environmental Epidemiology with R—A Case Study in Air Pollution and Health. Springer-Verlag, New York (2008)
19. Prates, M.O., Assunção, R.M., Rodrigues, E.C.: Alleviating spatial confounding for areal data problems by displacing the geographical centroids. Bayesian Anal. **14(2)**, 623–647 (2019)
20. Reich, B.J., Hodges, J.S., Zadnik, V.: Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics. **62(4)**, 1197–1206 (2006)
21. Reich, B.J., Yang, S., Guan, Y., Giffin, A.B., Miller, M.J., Rappold, A.: A review of spatial causal inference methods for environmental and epidemiological applications. Int. Statist. Rev. **89(3)**, 605–634 (2021)
22. Sahu, S.K., Mardia, K.V.: A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. J. of the Royal Statist. Society: Series C. **54(1)**, 223–244 (2005)
23. Schmidt, A.M., Gelfand, A.E.: A Bayesian coregionalization approach for multivariate pollutant data. J. Geophys. Res. **108(D24)** (2003) doi: 10.1029/2002JD002905
24. Thaden, H., Kneib, T.: Structural equation models for dealing with spatial confounding. The Am. Statistician. **72(3)**, 239–252 (2018)

Explainable Artificial Intelligence methods

Multidimensional Time Series Analysis via Bayesian Matrix Auto Regression

Analisi di Serie Temporali Multidimensionali via Autoregressione Matriciale Bayesiana

Alessandro Celani and Paolo Pagnottoni

Abstract It is often the case that time series observations are generated in matrix form in a wide variety of domains. Standard vector time series analysis may conceal interdependency structures of time series observations which original matrix-valued data may embed. We propose a Bayesian matrix autoregressive model in a bilinear form which presents several advances: i) it leads to a substantial dimensionality reduction and enhanced interpretability; ii) it provides an estimation procedure for covariate and lag structure; iii) it allows the introduction of Bayesian estimators. We propose maximum likelihood and Bayesian estimation of the model, and study their properties through real examples.

Abstract *Spesso le osservazioni legate a serie temporali sono generate in forma matriciale in un ampio spettro di domini. L'analisi vettoriale di serie temporali può celare strutture di interdipendenza delle osservazioni che i dati originali a valori matriciali possono invece includere. Noi proponiamo un modello matriciale autoregressivo Bayesiano in forma bilineare che presenta alcuni avanzamenti: i) porta ad una sostanziale riduzione di dimensionalità e aumentata interpretabilità; ii) prevede una procedura di stima per le covariate e la struttura autoregressiva; iii) permette l'introduzione di stimatori Bayesiani. Proponiamo stime del modello a massima verosimiglianza e Bayesiana, e studiamo le loro proprietà attraverso esempi reali.*

Key words: Autoregressive Models; Bayesian estimation; Bilinear Autoregression; Matrix-valued time series; Multivariate time series

---

Alessandro Celani
Università Politecnica delle Marche, Piazzale Raffaele Martelli 8, 60121, Ancona (AN), e-mail: a.celani@pm.univpm.it

Paolo Pagnottoni
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: paolo.pagnottoni@unipv.it

1 Introduction

Over recent times, there has been an emerging interest in modeling high dimensional time series, and several approaches have been developed for this purpose, such as: a) modelling with regularization (Rothman *et al.*, 2010; Kock & Callot, 2015; Schnücker, 2019); b) statistical and factor models (Bai & Ng, 2002; Forni *et al.*, 2005) among others; c) Bayesian methods (Park & Casella, 2008; Bańbura *et al.*, 2010; Brown & Griffin, 2010; Gefang, 2014; Korobilis, 2021). The common denominator of most of the extant modelling paradigms is to reduce the model parametrization starting from vector-valued time series data.

However, when considering panel data, it seems natural to conceive both the variable and country dimensions as potentially interconnected. In other words, variables related to the same country are allegedly strongly interrelated, as well as there might be dependence between the same time series realizations observed across different countries. Modelling such dependencies can therefore become even more informative when countries and variables are strongly interconnected, a prominent feature of economic and financial time series. The same concept applies to tensor-valued data, such as time-varying multilayer networks, where the relationship between layers might exhibit some dependence structure worthy to be modeled. Despite that, probabilistic properties, estimation procedures and theoretical properties of time series models generated by multidimensional data generating processes (DGPs) are still relatively open questions in the literature.

A recent strand of research has therefore started investigating probabilistic and theoretical properties, along with estimation procedures of matrix-valued time series models (Chen *et al.*, 2021; Wang *et al.*, 2019; Billio *et al.*, 2021, 2022). In particular, Chen *et al.* (2021) propose a first order matrix autoregression (MAR), which exploits the bi-dimensional structure to achieve dimensionality reduction and interpretability. Despite the merit of defining estimation procedures and probabilistic properties of the model, their estimation procedure is limited to the case of a simple MAR(1).

Against this background, we propose a novel matrix autoregressive model which presents three main originalities. Firstly, we design the model so that it allows to explicitely take into account for potential vector-valued covariates of interest. This is of utmost importance in many fields, particularly in macroeconomic analysis, where global exogenous vector-valued covariates might affect the two dimensions of the dependent variables.

Secondly, we extend the MAR(1) model from Chen *et al.* (2021) by providing a suitable estimation procedure for matrix autoregression with lag structure. We generalize the estimation procedure by deriving compact forms for the two dimensions of the model, which can be used to simultaneously estimate the right and left parameter matrices of interest independently from the lag order P.

Thirdly, we propose a fully Bayesian MAR model based on its formulation into right and left compact forms. While Hoff (2015) introduces Bayesian estimation of multidimensional regressions, they exclusively deal with the semi-conjugate prior framework. Differently, our proposal is equipped with Independent-Normal prior

formulation, which relaxes the hypothesis of dependence between conditional mean and variance parameters within each mode.

2 Model formulation and estimation

In this section we propose the generalization of the MAR(1) of Chen *et al.* (2021) to: a) a higher-order lag structure; b) the case of a DGP involving other observable variables which are determined outside the system, i.e. exogenous covariates.

Consider a PVAR including Q lags as well as the contemporaneous effect of K exogenous variables, assumed to be common across countries, i.e. a Panel VARX (PVARX):

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + ... + \Phi_P \mathbf{y}_{t-P} + \Psi_0 \mathbf{x}_t + ... + \Psi_Q \mathbf{x}_{t-Q} + \varepsilon_t, \tag{1}$$

where Ψ_q, $q = 0, ..., Q$ are $GN \times K$ coefficient matrices relating the endogenous variables to the external regressors.

As with autoregressive coefficients, we want to reduce the space of parameters and disentangle potential country and variable effects related to the covariates. This can be done by finding two lower dimensional objects, such that the impact of each $x_{k,t-q} \in \mathbf{x}_{t-q}$ is controlled by two matrices, embedding the variable and covariate effects respectively:

$$\underset{GN\times 1}{\Psi_{k,q}} \approx \underset{N\times 1}{\mathbf{D}_{k,q}} \otimes \underset{G\times 1}{\mathbf{C}_{k,q}}, \tag{2}$$

where $\Psi_{k,q}$ is the k-th column of Ψ_q and $\mathbf{C}_q = [\mathbf{C}_{1,q}, ..., \mathbf{C}_{K,q}] \in \mathbb{R}^{G\times K}$ and $\mathbf{D}_q = [\mathbf{D}_{1,q}, ..., \mathbf{D}_{K,q}] \in \mathbb{R}^{N\times K}$ represent the new left and right coefficient matrices related to the exogenous variables.

Such a structure, coherent with the bilinear nature of the MAR, would be well suited for matrix valued covariates. However, the kronecker product of the two new matrices $\mathbf{D}_q \otimes \mathbf{C}_q$ results in a $NG \times K^2$ dimensional object. Nevertheless, the impact of the covariates is only given by a subset of columns of this product, i.e. the one given by $\mathbf{D}_{i,q} \otimes \mathbf{C}_{j,q}$, where $i = j$, for $i, j = 1, ..., K$. This problem can be easily overcome by reshaping the covariates, turning them into matrices. Let $\mathbf{X}_{t-q} = \text{diag}(\mathbf{x}_{t-q}) \in \mathbb{R}^{K\times K}$ be the matricized version of $\mathbf{x}_{t-q}$, then it follows:

$$\Psi_q \mathbf{x}_{t-q} \approx \text{vec}(\mathbf{C}_q \mathbf{X}_{t-q} \mathbf{D}_q^{'}) = (\mathbf{D}_q \otimes \mathbf{C}_q)\tilde{\mathbf{x}}_{t-q}, \tag{3}$$

where $\tilde{\mathbf{x}}_{t-q} = \text{vec}(\mathbf{X}_{t-q})$. In this way, we are able to find a reasonable approximation of the covariate effects, where the dimensions of the objects are coherent with the bilinear form of the model. The proposed MARX(P,Q) can therefore be expressed in mathematical fashion as:

$$\mathbf{Y}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{Y}_{t-p} \mathbf{B}_p^{'} + \mathbf{C}_0 \mathbf{X}_t \mathbf{D}_0^{'} + \sum_{q=1}^{Q} \mathbf{C}_q \mathbf{X}_{t-q} \mathbf{D}_q^{'} + \mathbf{E}_t, \tag{4}$$

which reads, in vectorized form, as:

(5)

$$\mathbf{y}_t = \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\mathbf{y}_{t-p} + (\mathbf{D}_0 \otimes \mathbf{C}_0)\tilde{\mathbf{x}}_t + \sum_{q=1}^{Q} (\mathbf{D}_q \otimes \mathbf{C}_q)\tilde{\mathbf{x}}_{t-q} + \mathbf{e}_t. \tag{6}$$

Once the model is expressed as in equation (6), the probabilistic properties of the MARX(P,Q) are analogous to those of a VARX(P,Q) - see Lütkepohl (2005).

2.1 Iterative ML estimation

Assuming that time series $\mathbf{Y}_1, ..., \mathbf{Y}_T$ of the $\mathbf{Y}$ variables and $\mathbf{X}_1 = \text{diag}(\mathbf{x}_1), ..., \mathbf{X}_T = \text{diag}(\mathbf{x}_T)$ of the $\mathbf{x}$ variables are available, that is, we have a sample of size T both for each of the G indicators for N countries and for the K exogenous regressors. Since the model is generalized for different lags, for the purposes of estimation it is necessary to rewrite it in compact form, analogously to what is done for the VAR. However, the process is complicated by the fact that the MARX has a set of parameters that premultiply the lagged regressors (the row-wise ones) and another one that postmultiply them (the column-wise ones). As a result, this model admits two compact forms: one as a function of the former set of parameters, where the latter are considered as given, and viceversa. Without loss of generality, in what follows we suppose that $P = \max(P, Q)$. Define:

$$\mathscr{Y}_1 = \underbrace{[\mathbf{Y}_{P+1}\Sigma_2, ..., \mathbf{Y}_T\Sigma_2]}_{G \times \mathscr{J}_1}, \qquad \mathscr{Y}_2 = \underbrace{[\mathbf{Y}'_{P+1}\Sigma_1, ..., \mathbf{Y}'_T\Sigma_1]}_{N \times \mathscr{J}_2},$$

$$\mathscr{E}_1 = \underbrace{[\mathbf{E}_{P+1}\Sigma_2, ..., \mathbf{E}_T\Sigma_2]}_{G \times \mathscr{J}_1}, \qquad \mathscr{E}_2 = \underbrace{[\mathbf{E}'_{P+1}\Sigma_1, ..., \mathbf{E}'_T\Sigma_1]}_{N \times \mathscr{J}_2},$$

$$\mathscr{X}_{1,t} = \begin{bmatrix} \mathbf{Y}_{t-1}\mathbf{B}'_1 \\ \vdots \\ \mathbf{Y}_{t-P}\mathbf{B}'_P \\ \mathbf{X}_t\mathbf{D}'_0 \\ \vdots \\ \mathbf{X}_{t-Q}\mathbf{D}'_Q \end{bmatrix}, \qquad \mathscr{X}_{2,t} = \begin{bmatrix} \mathbf{Y}'_{t-1}\mathbf{A}'_1 \\ \vdots \\ \mathbf{Y}'_{t-P}\mathbf{A}'_P \\ \mathbf{X}'_t\mathbf{C}'_0 \\ \vdots \\ \mathbf{X}'_{t-Q}\mathbf{C}'_Q \end{bmatrix},$$

$$\mathscr{X}_1 = \underbrace{[\mathscr{X}_{1,P+1}\Sigma_2, ..., \mathscr{X}_{1,T}\Sigma_2]}_{\mathscr{K}_1 \times \mathscr{J}_1}, \qquad \mathscr{X}_2 = \underbrace{[\mathscr{X}_{2,P+1}\Sigma_1, ..., \mathscr{X}_{2,T}\Sigma_1]}_{\mathscr{K}_2 \times \mathscr{J}_2},$$

$$\mathscr{B}_1 = \underbrace{[\mathbf{A}_1, ..., \mathbf{A}_P, \mathbf{C}_0, \mathbf{C}_1, ..., \mathbf{C}_Q]}_{G \times \mathscr{K}_1}, \qquad \mathscr{B}_2 = \underbrace{[\mathbf{B}_1, ..., \mathbf{B}_P, \mathbf{D}_0, \mathbf{D}_1, ..., \mathbf{D}_Q]}_{N \times \mathscr{K}_2},$$

where $\mathscr{J}_1 = N(T-P)$, $\mathscr{J}_2 = G(T-P)$, $\mathscr{K}_1 = GP + K(Q+1)$ and $\mathscr{K}_2 = NP + K(Q+1)$. Using this notation, for $t = P+1, ..., T$ the MARX(P,Q) can be compactly rewritten, for $i = 1, 2$, as

$$\mathscr{Y}_i = \mathscr{B}_i \mathscr{X}_i + \mathscr{E}_i, \tag{7}$$

$$\mathscr{E}_i \sim \mathscr{MN}(0, \Sigma_i, \mathbf{I}_{\mathscr{J}_i}), \tag{8}$$

whose log-likelihood is given by

$$\log \mathscr{L}(\theta_1, \theta_2) = -\frac{\mathscr{J}_1 \mathscr{J}_2}{2}\pi - \frac{\mathscr{J}_2}{2}\log|\Sigma_i| - \frac{1}{2}\mathrm{tr}\left[(\mathscr{Y}_i - \mathscr{B}_i\mathscr{X}_i)'\Sigma_i^{-1}(\mathscr{Y}_i - \mathscr{B}_i\mathscr{X}_i)\right], \tag{9}$$

where $\theta_i = \{\mathscr{B}_i, \Sigma_i\}$, $|\cdot|$ denotes the matrix determinant and $\mathrm{tr}(\cdot)$ is the trace operator. Given that the parameters of one compact form are nested into the other and viceversa, the ML estimator cannot be found simultaneously for all the parameters of interest. Nevertheless, given θ_2, the problem of finding the optimal $\mathscr{B}_1$ and Σ_2 is strictly convex and viceversa. In fact, it holds that:

$$\theta_i(\theta_{-i}) = \underset{\mathscr{B}_i, \Sigma_i}{\mathrm{argmin}} - \mathscr{J}_i \log|\Sigma_i| - \mathrm{tr}\left[(\mathscr{Y}_i - \mathscr{B}_i\mathscr{X}_i)'\Sigma_i(\mathscr{Y}_i - \mathscr{B}_i\mathscr{X}_i)\right]. \tag{10}$$

The problem postulated as it is suggests, for each i, the use of a two-stage algorithm where at each step s, the new values of $\theta_i^{[s]}$ are generated by $\hat{\mathscr{B}}_i^{[s]}(\hat{\theta}_{-i}^{[s_i]}, \Sigma_i^{[s-1]})$, $\hat{\Sigma}_i^{[s]}(\hat{\theta}_{-i}^{[s_i]}, \hat{\mathscr{B}}_i^{[s]})$ where $s_i = s-1$ if $i = 1$ and $s_i = s$ if $i = 2$. The optimality conditions are the same as a multivariate regression model, that is:

$$\hat{\mathscr{B}}_i = (\mathscr{Y}_i\mathscr{X}_i')(\mathscr{X}_i\mathscr{X}_i')^{-1}, \tag{11}$$

$$\hat{\Sigma}_i = \mathscr{J}_i^{-1}(\mathscr{Y}_i - \hat{\mathscr{B}}_i\mathscr{X}_i)(\mathscr{Y}_i - \hat{\mathscr{B}}_i\mathscr{X}_i)'. \tag{12}$$

Nevertheless, there is an identification issue regarding the row-wise and the column-wise coefficient matrices: to illustrate, consider that if $\hat{\mathscr{B}}_1$ and $\hat{\mathscr{B}}_2$ are solution of this problem, so are $\alpha_1\hat{\mathscr{B}}_1$ and $\alpha_2\hat{\mathscr{B}}_2$, with $\alpha_1\alpha_2 = 1$. In fact, they yield the same kronecker product $\mathscr{B}_2 \otimes \mathscr{B}_1 = \alpha_2\mathscr{B}_2 \otimes \alpha_1\mathscr{B}_1$, which is always identified. In order to ensure the stability of the iterative process, we select those two constants aiming at keeping the magnitude of the matrices of comparable magnitudes; thereby we renormalize $\mathscr{B}_1$ and $\mathscr{B}_2$ by choosing $\alpha_1 = ||\mathscr{B}_2||_F / ||\mathscr{B}_1||_F$ and $\alpha_2 = 1/\alpha_1$, where $||\cdot||_F$ stands for the Frobenius norm. The same applies for Σ_1 and Σ_2.

Given an a priori estimate of the coefficient matrices related to the corresponding PVARX, i.e. $\{\hat{\Phi}_1, ..., \hat{\Phi}_P, \hat{\Psi}_0, ..., \hat{\Psi}_Q\}$, a reasonable set of starting values for the iterative algorithm is that solving the Nearest Kronecker Product (NKP) problem (Van Loan & Pitsianis, 1993; Loan, 2000). As example, for $\mathbf{A}_1^{[0]} \in \mathscr{B}_1^{[0]}$ and $\mathbf{B}_1^{[0]} \in \mathscr{B}_2^{[0]}$ the problem is:

$$\begin{aligned}\{\hat{\mathbf{A}}_1^{[0]}, \hat{\mathbf{B}}_1^{[0]}\} &= \underset{\mathbf{A}_1, \mathbf{B}_1}{\operatorname{argmin}} \quad \left\| \hat{\Phi}_1 - \mathbf{B}_1 \otimes \mathbf{A}_1 \right\|^2 \\ &= \underset{\mathbf{A}_1, \mathbf{B}_1}{\operatorname{argmin}} \quad \left\| \mathscr{G}(\hat{\Phi}_1) - \operatorname{vec}(\mathbf{A}_1)\operatorname{vec}(\mathbf{B}_1)' \right\|^2,\end{aligned} \tag{13}$$

where $\mathscr{G}(\cdot)$ is a function that permutes the entries of its arguments, meaning that $\mathscr{G}(\hat{\Phi}_1)$ is a rearranged version of $\hat{\Phi}_1$ and $\mathscr{G}(B \otimes A) = \operatorname{vec}(A)\operatorname{vec}(B)'$.

As far as the coefficient matrices related to the covariates are concerned, this preliminary estimation shall be done separately for each k-th column of $\hat{\Psi}_q$, so as to obtain the k-th column of $\hat{\mathbf{C}}_q^{[0]}$ and $\hat{\mathbf{D}}_q^{[0]}$, respectively.

2.2 Bayesian estimation

Hoff (2015) considered the conditionally conjugate prior framework for the multilinear model, assuming a Normal-Wishart prior for each mode. However, their framework assumes the existence of a dependence between the variance of innovations and that of the conditional mean parameters, for each dimension.

An independent Normal-Wishart prior framework can overcome this limitation by means of a proper prior assumption. Recall that given θ_{-i}, the MARX can be seen as a regression model with mode specific conditional mean and variance in its i-th dimension. By further assuming independence between $\mathscr{B}_i$ and Σ_i for both the dimensions, one can set independent prior distributions on the parameters of interest. Following the standard multivariate regression approach, we specify the following a priori assumptions for the mean parameters:

$$\pi(\beta_i) \sim \mathscr{N}(\underline{\beta_i}, \underline{\Omega_i}), \tag{14}$$

where $\beta_i = \operatorname{vec}(\mathscr{B}_i)$. As for the two covariance matrices, remember that they are not separately identifiable from the likelihood, only their product is. In particular, setting the prior degrees of freedom of Σ_1 and Σ_2, namely $\underline{\nu_1}$ and $\underline{\nu_2}$, is still a debated issue, given that their choice affects the total variation in the data, i.e. $\operatorname{tr}(\Sigma_1)\operatorname{tr}(\Sigma_2)$. Wang & West (2009) addressed this problem by imposing the hard constraint $\Sigma_{2,11} = 1$. We follow Hoff (2015), which proposes to add a level of dependency between Σ_1 and Σ_2 via an hyperparameter γ which reflects the total variation in the data:

$$\begin{aligned}\pi(\gamma) &\sim \mathscr{G}a(\underline{a}, \underline{b}), \\ \pi(\Sigma_i|\gamma) &\sim \mathscr{IW}(\gamma \underline{S_i}, \underline{\nu_i}),\end{aligned} \tag{15}$$

such that by setting $\underline{S_1} = \mathbf{I}_G/G$, $\underline{S_2} = \mathbf{I}_N/N$ and $\underline{\nu_1} = G+2$, $\underline{\nu_2} = N+2$ we have:

$$\begin{aligned}&\mathrm{E}[\operatorname{tr}(\operatorname{Cov}(\operatorname{vec}(\mathbf{E}_t)))] = \\ &\mathrm{E}[\operatorname{tr}(\Sigma_2 \otimes \Sigma_1)] = \\ &\mathrm{E}[\operatorname{tr}(\Sigma_1)\operatorname{tr}(\Sigma_2)] = \gamma^2.\end{aligned} \tag{16}$$

Therefore, the joint prior distribution can be summarized as:

$$\pi(\beta_1,\beta_2,\Sigma_1,\Sigma_2) = \pi(\beta_1)\pi(\beta_2)\pi(\Sigma_1|\gamma)\pi(\Sigma_2|\gamma)\pi(\gamma). \tag{17}$$

The independence structure among parameters does not allow to derive a closed form for the posterior distribution. Hence, we adopt a posterior simulator, the Gibbs Sampler, which is able to approximate the posterior joint distribution from the conditional posterior distribution of each parameter of interest. While the posterior distributions for β_1,β_2 are the same as in a standard VAR, those related to Σ_1,Σ_2 differ slightly due to the addition of the hyperparameter γ. As a consequence of the prior structure, the Gibbs sampler can be articulated as follows:

1. Draw Σ_1 from $\mathscr{IW}(\gamma\underline{S}_1+\hat{S}_1,\underline{\nu}_1+\mathscr{J}_1)$;
2. Draw β_1 from $\mathscr{N}(\bar{\beta}_1,\bar{\Omega}_1)$;
3. Draw Σ_2 from $\mathscr{IW}(\gamma\underline{S}_2+\hat{S}_2,\underline{\nu}_2+\mathscr{J}_2)$;
4. Draw β_2 from $\mathscr{N}(\bar{\beta}_2,\bar{\Omega}_2)$;
5. Draw γ from $\pi(\gamma|\Sigma_1,\Sigma_2)\sim\mathscr{G}a\Big(\underline{a}+\frac{1}{2}\big[\underline{\nu}_1 G+\underline{\nu}_2 N\big],\underline{b}+\frac{1}{2}\big[\mathrm{tr}(\underline{S}_1\Sigma_1^{-1})+\mathrm{tr}(\underline{S}_2\Sigma_2^{-1})\big]\Big)$

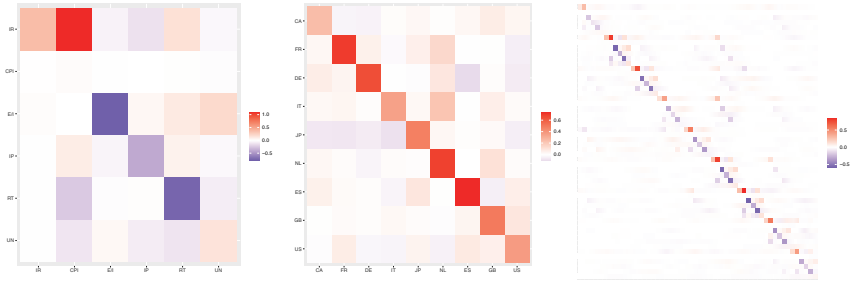
where $\bar{\Omega}_i=\big[\underline{\Omega}_i^{-1}+(\mathscr{X}_i\mathscr{X}_i)'\otimes\Sigma_i^{-1}\big]^{-1}$, $\bar{\beta}_i=\bar{\Omega}_i\big[\underline{\Omega}_i^{-1}\underline{\beta}_i+(\mathscr{X}_i\otimes\Sigma_i^{-1})\mathrm{vec}(\mathscr{Y}_i)\big]$ and $\hat{S}_i=(\mathscr{Y}_i-\hat{\mathscr{B}}_i\mathscr{X}_i)(\mathscr{Y}_i-\hat{\mathscr{B}}_i\mathscr{X}_i)'$ with $\hat{\mathscr{B}}_1$, $\hat{\mathscr{B}}_2$ being the conditional mean ML estimator.

3 Application

We now illustrate an empirical application of the proposed model. We study global interconnectedness of a high-dimensional panel of monthly macroeconomic indicators. Specifically, we consider $G=6$ monthly economic indicators, i.e. 10 year government Interest Rate (IR), Consumer Price Index (CPI), Export over Import (E/I), Industrial Production (IP), Retail Trade (RT) and Unemployment rate (U) for $N=9$ countries, namely Canada (CA), France (FR), Germany (DE), Italy (IT), Japan (JP), Netherlands (NL), Spain (ES), Great Britain (GB) and United States (US). The five European countries account for more than 70% of the total European Union GDP in 2021. Moreover, we include $K=3$ global indices: Agricultural Raw Material (ARM), Metals (M) and Crude Oil (CO) [1]. The analyzed period ranges from January 2000 to June 2019.

We fit a Bayesian MARX(2,1) whose posterior distribution of the parameters of interest are obtained with 3000 Monte Carlo iterations after 2000 discarded burn in ones. Figure 1 shows the estimated first order left and right coefficient matrices $\mathbf{A}_1$ and $\mathbf{B}_1$, measuring the first lag variable and country effects of the autoregressive dynamics respectively, and the reconstructed coefficient matrix $\mathbf{B}_1\otimes\mathbf{A}_1$.

[1] Data are retrieved from the OECD Database at https://stats.oecd.org/index.aspx?lang=en and from the IMF Primary Commodity Prices at https://www.imf.org/en/Research/commodity-prices



(a) **(b)** **(c)**

Fig. 1: Median of the posterior entries of the first order left coefficient matrix $\mathbf{A}_1$ (a), of the right one $\mathbf{B}_1$ (b), and of $\mathbf{B}_1 \otimes \mathbf{A}_1$ (c).

Overall, it is interesting to notice that the diagonal elements of the parameter matrices concur to a large portion of the system's autoregressive dynamics, as intuitively expected. This partly contrasts with Billio *et al.* (2022), whose symmetric parallel factor (PARAFAC) decomposition annihilates the relative importance of diagonal coefficients, thereby losing the structure of own autoregressive dynamics.

Attention should be paid to the sign of the coefficients in a MAR. $\mathbf{A}_1$ and $\mathbf{B}_1$ would yield the same kronecker product even if multiplied both by any real number and its reciprocal. To illustrate, if the generic element of $\mathbf{B}_1 \otimes \mathbf{A}_1$ is positive, we can only say that the two coefficients generating it have the same sign. However, due to the unidentifiability of the two, we cannot state whether these are both negative or positive (in case of same signs), or which is the positive one (in case of opposite signs).

For what concerns $\mathbf{A}_1$, a strong influence of own country past CPI changes on current ones in IR is detected. This is presumably due to the fact that an increase in inflation, normally related to economic growth, pushes the monetary authorities to raise the interest rate, in an attempt to curb the inflationary pressure as a countercyclical operation. On the other hand, the inflation rate does not seem to be driven by any of the other indicators used in this study, it turns out to be rather independent.

We also report out-sample forecast performances of the proposed ML and Bayesian estimators of the MAR model, as well as those of the same competing alternatives considered in the simulation study, relative to the performance of the stacked VAR estimator. Specifically, starting from the first month of 2000 to the end of the series (the last month of 2018), we fit the models and then derive the $H = 1, 3, 6$ step ahead predictions using data from January 2019 to June 2019. We then compute the logarithm of the ratio between the MFSE of each model and that of the stacked VAR, and collect the results in Table 1.

Table 1 shows that, overall, the Bayesian MAR overperforms all competing models in the real forecasting exercise. In particular, we see that both the ML and

	FH:1	FH:3	FH:6
MLE	0.381	-0.004	-0.006
Bayes	**-0.041**	**-0.252**	**-0.076**
CC	0.661	0.044	0.004
SSVS	1.111	0.231	0.062
SSSS	2.672	1.509	0.676
LASSO	0.447	0.001	-0.004

Table 1: Logarithm of the ratio between MSFE of each model and MSFE of VAR. A value below 0 corresponds to a better forecast accuracy with respect to the stacked VAR model.

Bayesian estimator of the MAR yield generally better forecasts if compared to the other competing alternatives, except for the ML estimator with respect to the standard VAR when considering $H = 1$.

4 Conclusion

We propose a generalization and Bayesian estimation of the autoregressive model for matrix-valued time series to higher order autoregressive structure and inclusion of vector-valued covariates. The model exploits the original matrix structure of data, fostering interpretability of multidimensional relationship structures, and yields a more parsimonious model representation, if compared to the standard VAR approach. In particular, its novel representation into compact forms provides a suitable procedure which overcomes the problem of iteratively estimating each single coefficient matrix in a separate way. Upon such general structure, we propose a fully Bayesian estimation procedure set up with independent Normal-Wishart prior.

MAR models may however still suffer from large dimensions, despite the number of parameters involved in the estimation is still crucially lower than that of the stacked VAR. This calls for the implementation of regularization and sparse and group-sparse estimation approaches in the future. For very large dimensional matrix time series, Wang et al. (2018) proposed a factor model in a bilinear form. Matrix autoregressive models can be used to model the factor matrix in that of Wang et al. (2018) to build a dynamic factor model in matrix form.

References

BAI, J., & NG, S. 2002. Determining the Number of Factors in Approximate Factor Models. *Econometrica*, **70**(1), 191–221.

BAŃBURA, M., GIANNONE, D., & REICHLIN, L. 2010. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, **25**(1), 71–92.

BILLIO, M., CASARIN, R., COSTOLA, M., & IACOPINI, M. 2021. A Matrix-Variate t Model for Networks. *Frontiers in Artificial Intelligence*, **4**, 49.

BILLIO, MONICA, CASARIN, ROBERTO, IACOPINI, MATTEO, & KAUFMANN, SYLVIA. 2022. Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 1–30.

BROWN, P.J., & GRIFFIN, J.E. 2010. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**(1), 171 – 188.

CHEN, R., H., XIAO, & YANG, D. 2021. Autoregressive models for matrix-valued time series. *Journal of Econometrics*, **222**(1, Part B), 539–560.

FORNI, M., HALLIN, M., LIPPI, M., & REICHLIN, L. 2005. The Generalized Dynamic Factor Model. *Journal of the American Statistical Association*, **100**(471), 830–840.

GEFANG, D. 2014. Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *International Journal of Forecasting*, **30**(1), 1–11.

HOFF, P. D. 2015. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, **9**(3), 1169–1193.

KOCK, A.B., & CALLOT, L. 2015. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, **186**(2), 325–344.

KOROBILIS, DIMITRIS. 2021. High-dimensional macroeconomic forecasting using message passing algorithms. *Journal of Business & Economic Statistics*, **39**(2), 493–504.

LOAN, C. F. VAN. 2000. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, **123**(1), 85–100. Numerical Analysis 2000. Vol. III: Linear Algebra.

LÜTKEPOHL, H. 2005. *New Introduction to Multiple Time Series Analysis*. Springer.

PARK, T., & CASELLA, G. 2008. The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

ROTHMAN, A. J., LEVINA, E., & ZHU, J. 2010. Sparse Multivariate Regression With Covariance Estimation. *Journal of Computational and Graphical Statistics*, **19**(4), 947–962. PMID: 24963268.

SCHNÜCKER, A.M. 2019 (Nov.). *Penalized Estimation of Panel Vector Autoregressive Models*. Econometric Institute Research Papers EI-2019-33. Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.

VAN LOAN, C. F., & PITSIANIS, N. 1993. *Approximation with Kronecker Products*. Dordrecht: Springer Netherlands. Pages 293–314.

WANG, D., LIU, X., & CHEN, R. 2019. Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, **208**(1), 231–248. Special Issue on Financial Engineering and Risk Management.

WANG, H., & WEST, M. 2009. Bayesian analysis of matrix normal graphical models. *Biometrika*, **96**(4), 821–834.

Advances in Classification and Data Analysis

Optimizing time slots in scientific meetings: a Latent Dirichlet allocation approach

Ottimizzazione degli slot nelle conferenze scientifiche: un approccio basato sulla Latent Dirichlet allocation

Luca Frigau

Abstract When participating in scientific conferences, often it happens that sessions with similar topics are scheduled at the same time, thus leading to having to choose which one to follow and giving up the others. Recently, to overcome this problem, an algorithm has been proposed that uses Latent Dirichlet allocation to optimize the allocation of sessions in slots. The results obtained on the Joint Statistical Meetings 2020 program, which concerned more than 40 parallel sessions, have been very interesting. In this paper, we investigate the actual adaptability and effectiveness of this algorithm also for medium-sized conference programs such as that of the SIS.
Abstract *Quando si partecipa a delle conference scientifiche, spesso capita che sessioni con argomenti simili siano programmate alla stessa ora, portando quindi a dover scegliere quale seguire e a rinunciare alle altre. Recentemente, per ovviare a questo problema è stato proposto un algoritmo che utilizza la Latent Dirichlet allocation per ottimizzare l'allocazione delle sessioni negli slot. I risultati ottenuti sul programma della Joint Statistical Meetings 2020, che prevedeva più di 40 sessioni in parallelo, sono stati molto interessanti. In questo lavoro investighiamo l'effettiva adattabilità ed efficacia di questo algoritmo anche per i programmi delle conferenze di media dimensione come quello della SIS.*

Key words: SIS, conference, LDA, topic modeling, optimization, parallel sessions.

1 Introduction

Organize a scientific meeting is very arduous, especially defining the scientific program. Despite among the members of the program committee there are usually several ones who have already had that role in other conference organizations and so

Luca Frigau
University of Cagliari, Viale S. Ignazio 17, 09123 Cagliari, Italy e-mail: frigau@unica.it

have expertise about it, each time it consists in a new challenge with never met problems.

In planning a scientific program it is almost always indispensable to provide parallel sessions for the success of the event. The main reasons are two: it is necessary that the number of days of the conference is not excessive to give everyone the opportunity to participate in the whole event; provide specialist talks at each time slot interesting for everyone. On the other hand, the main drawback of providing sessions in parallel is the possibility of overlapping between concurrent sessions, which involves giving up taking part in a session in which you are interested.

Normally, the conference sessions assignment to the time slots is performed manually by the program committee. This is an activity time-consuming, subjective, and where it is easy to make mistakes due to a large number of possible solutions. In fact, in addition to the constraints related to the availability of the rooms, it is possible that there are other constraints that require the same person must be present at other sessions besides the one in which he is a speaker, for example when he is assigned the role of the discussant. Consequently, the key task is to minimize overlapping content in the same time band among the contributed, solicited, and specialized sessions.

In order to overcome the above-mentioned issues, [4] proposed an automized flexible alternative strategy for organizing the Joint Statistical Meetings (JSM), which is the main conference of the American Statistical Association that brings together several thousand statisticians for a significant event, with usually more than 40 rooms available for parallel sessions. In particular, this method uses Latent Dirichlet allocation and optimal scheduling to minimize content conflicts among parallel sessions. Despite it having been developed on the JSM, that approach is generalizable to all conference.

In literature, it seems no other methods have been specifically proposed for this goal, even if other kinds of schedule optimization program have been published. For instance, [9] proposed a system for school schedule creation with an optimization led according to several methods and criteria. [6], instead, used a Genetic Algorithm to solve the problem of optimizing the scheduling of lecturing by managing rooms, lecturers, and times constraints. Moreover, [7] introduced a general solution for the School timetabling problem based on an adaptive approach with a primary aim to solve the issue of clashes in lectures and subjects, pertaining to teachers.

In this paper, we apply the method proposed by [4] to minimize overlapping between parallel sessions by demonstrating its usefulness also for the program scheduling of the medium-sized conferences in addition to those of high magnitude as shown in their work. In particular, we carried out it on a SIS conference, that is the main scientific meeting of the Italian Statistical Society. The remainder of the paper is organized as follows. Section 2 recalls the basics of the Latent Dirichlet allocation. In Section 3 the algorithm to minimize the overlapping between parallel sessions is illustrated. Section 4 reports the application of the method to a SIS conference, and in particular, describes the preparation and cleaning of the data, as well as the obtained results, showing a significant improvement over the 2014 program

in terms of total conflict. Finally, Section 5 ends the paper with some concluding remarks

2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a statistical topic model developed by [3] which has become widely used tool in textual analysis. The basic intuition behind LDA is that documents deal with multiple topics. It is easily described by its generative process, the imaginary process through which the model assumes the documents arose.

Let us define topics as distributions over a fixed vocabulary. For instance, the topic "finance" has words about finance with high probabilities and those about the other topics with zero-low probabilities. Moreover, the distributions of the topics are assumed to be defined *a priori* before any document has been generated and the order of words in a document does not matter, as LDA is a bag-of-words model. Then, for each document the words are generated in a two-stage process:

1. Randomly choose a distribution over topics.
2. For each word in the document:
   a. Randomly choose a topic from the distribution over topics in step 1.
   b. Randomly choose a word from the corresponding distribution over the vocabulary.

In the document, the topics are presented in a different proportion (step 1); the words of each document are drawn from one of the topics where the topic is chosen from the per-document distribution over topics.

LDA uses the Markov chain Monte Carlo (MCMC) to decode the generative process. Specifically, given a set of documents and a previously defined number of topics K, MCMC estimates the distributions corresponding to each topic as well as the mixture probabilities for any document on topics.

In Fig. 1, the dark node w represents words, which is the solely observed variable in the model. The light node z represents the topic, which hides inside the document.

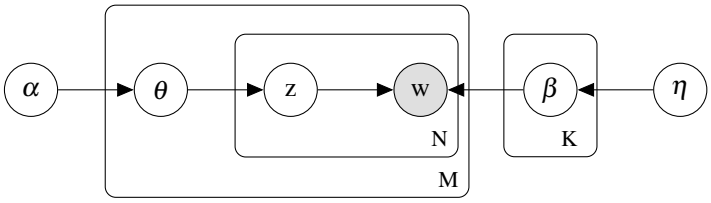


Fig. 1 Graphical model for LDA.

The topic distribution under each document is a Multinomial distribution $Mult(\theta)$ with its conjugate prior $Dir(\alpha)$. The word distribution under each topic is a Multinomial distribution $Mult(\beta)$ with its conjugate prior $Dir(\eta)$. For the n-th word in the certain document, first we select a topic z from per document-topic distribution $Mult(\theta)$, then select a word under this topic $w|z$ from per topic-word distribution $Mult(\beta)$. Here is the generative process:

1. Draw $\theta_m \sim \text{Dirichlet}(\alpha)$
2. For each topic $k \in \{1,...,K\}$
   - Draw $\beta_k \sim \text{Dirichlet}(\eta)$
3. For each word w_n in document m, $n \in \{1,...,N\}$
   - Draw topic $z_n \sim \text{Multinomial}(\theta_m)$
   - Draw word $w_n|z_n \sim \text{Multinomial}(\beta_k)$

Several authors have investigated the accuracy of LDA. If the number of topics is large, it is possible that in the topic distribution of each document is characterized by a strong dispersion of probabilities, which makes the identification of the main topics of the document less accurate. To overcome this problem, regularization is used. To regularize consists of zeroing out all the probabilities below a certain threshold (e.g. 10%) and then renormalizing the distribution of the probabilities to the other topics. The performance of LDA is characterized by high variability, nonetheless often the topics associated with the documents are convincing and reasonable. Empirically it has been noted that the results improve when the documents are long enough (at least one page) and the topics are clearly distinct.

An interesting problem concerns the definition of the "correct" number of topics. Several strategies can be used to solve that problem. They are based on assessing goodness-of-fit through already noted measures such as perplexity and variation of information distance or new distances created such as criterion curve [1]. Otherwise, the assessment is done through a subjective evaluation of the researchers by visualizing the plotted clustering results or checking the highest-probability words of the topics. However, that problem can be seen from another point of view, which considers a thematic structure that can be interpreted on many different scales. [10] found five significant topics in the subset of the 2012 political blog posts discussing the Trayvon Martin shooting. But [5] found crude arguments that they were clearly correct for the corpus of all political blog posts from 2012. The appropriate number of topics depends on whether you want fine resolution or coarse resolution or intermediate resolution in specificity of the topics.

3 Methodology

A session usually deals with more than one topic at the same time. Nevertheless for human beings, it is difficult to distinguish in a weighted way the different topics that compose it. In fact, they are led to consider only the prevailing one, without taking

into account the other ones. For instance, if a session concerns 60% clinical studies, 30% clustering and 10% time series, a human labels it only as clinical studies, not considering the remaining 40% of the information. Starting from this assumption, the subsequent minimization of the overlapping between topics within the same time band, in addition to computational problems linked to the limits of human beings, is strongly influenced by excessive approximation in the definition of the topics covered in the sessions.

The approach proposed by [4] identifies the distribution of general topics in the different sessions through an LDA, and then applies a greedy optimization strategy to minimize overlapping topics within the same time band. It would be preferable to consider the search for the global optimum on all possible combinations, but for computational reasons this solution is not feasible, especially for big-size meetings. The algorithm is made up of three phases:

1. Assign people with more than one role to sessions in different time bands.
2. Randomly assigns the remaining sessions.
3. Greedily optimizes the assignment.

Phases two and three are repeated ϕ times, and then the best optimum found is considered. Heuristically, it emerges smaller the number of sessions smaller ϕ is needed.

Let $\sigma = \{s_1, \ldots, s_N\}$ be the set of sessions that must be assigned to a band. And let $\boldsymbol{\Gamma}$ be the $N \times K$ matrix with entry equal to γ_{ij} being the extent to which session i participates in topic j. In order to zeroes out small topic weights, a regularization is applied. In particular, all γ_{ij} with a value lower than 0.05 are forced to zero and then the others are renormalized.

For two sessions s_i and s_j, their total variation distance is

$$\delta_{ij} = \frac{1}{2} \sum_{k=1}^{K} |\gamma_{ik} - \gamma_{jk}|, \tag{1}$$

so small values of δ_{ij} imply that the sessions have strongly overlapping content. To measure the topic overlap for an entire assignment of the sessions, we use

$$\rho = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \theta_{ij} \tag{2}$$

where $\theta_{ij} = 1$ if s_i and s_j are assigned to the same time band, and otherwise it is zero. Larger value of ρ better assignment is.

In the first phase, the algorithm takes into account the constraints existing, specifically the impossibility of assigning the same time band to two sessions in which the same person plays a role active in both. Through a method that minimizes the topic overlapping, these sessions are assigned to time bands in order to respect the constraints. In the second phase, the other sessions are randomly assigned to the remaining free time slots, respecting the number of parallel sessions allowed each time. The algorithm then computes ρ for that assignment. In the third phase, the

algorithm greedily reassigns the sessions to different time bands. Specifically, two time bands are chosen randomly, and within each one, a session is selected randomly respectively, so that the switch does not cause problems to the constraints. The two selected sessions are then swapped and the new ρ is calculated. If the new ρ is greater than the previous one then the exchange is maintained; otherwise the the swap is reset and a new swap is tried. The algorithm ends when 10,000 exchange attempts have been produced without obtaining a larger ρ.

4 Optimizing the SIS 2014 Schedule

In order to test the algorithm in a medium-size meeting, we consider the conference SIS held in Cagliari in 2014. The main reason for this choice is related to the availability of the data. In fact, since the author has been part of the local organizing committee of SIS 2014, he was able to collect easily all the input information needed for carrying out the algorithm, in particular the abstracts, keywords, titles, and speakers of the talks. A tentative to collect the data of the last SIS conference (e.g. SIS 2021) was made, but unfortunately, its abstracts were impossible to be scrapped from the Book of Abstracts due to technical problems.

In the SIS 2014 the talks were spread over three days, from June 11th, 2014 to June 13th, 2014. We can distinguish between the sessions that were plenarily conducted and those conducted in parallels. We focus on the latter, because obviously no overlapping problems can arise from the former. Specifically, three different kinds of sessions were scheduled in parallel: Contributed Paper Sessions (CP), Solicited Sessions (SL) and Specialized Sessions (SP). Table 1 illustrates the actual program scheduled by the committee. It emerges they scheduled in parallel sessions of the same type. In other words, in the same time bands we can find either all CP, or all SL, or all SP. This constraint splits the allocation problem of the sessions into three independent optimization problems, one for each kind of session.

Firstly, we collected the data of the SIS 2014. For each talk, we gathered the title, the abstract text, the keywords, and the speaker. The latter information was used to define the constraints whilst the other three were merged. In particular to boost the signal keywords and title were repeated three times. In that way, we obtained a single text for each talk. Successively, since the talks of the same sessions are considered inseparable, their corresponding texts were merged. The same was done for their speakers.

The next step consisted in removing the stop-words, which do not include important information but are usually considered noise. To define the stop-words we used two lists as references, specifically Lingua::StopWords (https:// metacpan.org/pod/Lingua::StopWords) and Stopwords ISO [2].

Then the words were stemmed, that is they were replaced by their corresponding token. This allows losing a little information in exchange for a reduction of dimensionality. In fact, nonetheless “estimate”, “estimates”, “estimation”, and “estimating” are different words and they can be considered as bearers of slightly different

Schedule	Room	Session ID	Session Title
June 11th 11:00 - 12:15	Aula A	CP-01	Demography
	Aula B	CP-02	Statistics in finance
	Aula Anfiteatro	CP-03	Statistics in medicine
	Aula 1	CP-04	Clustering methods: theory and applications
	Aula Arcari	CP-05	Functional data analysis
	Magna Econ	CP-06	Forensic statistics
June 11th 14:30 - 15:45	Aula A	SP-01	Recent advances in Biostatistics
	Aula B	SP-02	Clustering real time data streams
	Aula Arcari	SP-03	Bayesian nonparametrics: methods and applications
June 11th 16:30 - 17:45	Aula A	SP-04	Recent advances in time series analysis
	Aula B	SP-05	New challenges in survey sampling
	Aula Arcari	SP-06	Directional data
June 11th 17:45 - 19:00	Aula A	SL-01	Bayesian models for complex problems
	Aula B	SL-02	Geostatistics and environmental applications
	Aula Arcari	SL-03	Robust methods for the analysis of complex data
	Aula 11	SL-04	Statistics for environmental phenomena and their interactions
	Aula 12	SL-05	Mixture and latent variable models for causal inference and analysis of socio-economic data
June 12th 09:00 - 10:15	Aula Anfiteatro	SL-06	Equity and sustainability: theory and relationships
	Aula B	SL-07	Advances in Bayesian statistics
	Aula Arcari	SL-08	Statistical models for the analysis of energy markets
	Aula 11	SL-09	Recent developments in sampling theory
	Aula 12	SL-10	Functional data analysis
June 12th 11:45 - 13:00	Aula A	SP-07	Scoring Rules and Pseudo-likelihoods: connections and developments
	Aula B	SP-08	Quantile and M-quantile regression: random effects and regularization
	Aula Arcari	SP-09	Methodological Issues for constructing composite indicators
June 12th 14:30 - 15:45	Aula A	CP-07	Inequality measures in socio-economic phenomena
	Aula B	CP-08	Advances in statistical modelling
	Aula Anfiteatro	CP-09	Developments in Bayesian inference
	Aula 1	CP-10	Educational statistics
	Aula Arcari	CP-11	Sanitary statistics and epidemiology
	Magna Econ	CP-12	Survey methodology
June 13th 09:00 - 10:15	Aula Anfiteatro	SL-11	Extremes and dependent sequences
	Aula B	SL-12	Issues in ecological statistics
	Aula Arcari	SL-13	Computations with intractable likelihood
	Aula 11	SL-14	Geographical information in sampling and estimation
	Aula 12	SL-15	Clinical designs
June 13th 14:30 - 15:45	Aula A	CP-13	Statistical methods for the analysis of fertility and health
	Aula B	CP-14	Advances in compositional data analysis
	Aula Anfiteatro	CP-15	Spatial and spatio-temporal analysis
	Aula 1	CP-16	Environmental and poverty data analysis
	Aula Arcari	CP-17	Topics in regression models
	Magna Econ	CP-18	Bayesian methods and models
June 13th 16:30 - 17:45	Aula Anfiteatro	SL-16	Bayesian inference for high-dimensional data
	Aula B	SL-17	Use of Big Data for the production of statistical information
	Aula 11	SL-18	Measuring the Smart City
	Aula 12	SL-19	Forecasting economic and financial time series

Table 1 Program of the sessions scheduled in parallels of the SIS 2014.

information, their meaning is practically the same. Consequently, by stemming them into the same token the vocabulary is reduced and the model performs better. To stem the words we used the Snowball stemmer [8]. Moreover, the vocabulary was again reduced by removing any token that appeared fewer than five times and those that appeared in only a single session since these provide no information relevant to minimizing schedule conflicts. In the next step, we created the n-gram. We set $n = 5$ and than kept solely those considered as technical phrases and with an occurrence of at least six. To check the former condition we performed the standard binomial test and removed those with an observed proportion of cooccurrences that had exceeded that expected under independence with significance probability less than 0.005.

Finally, in order to increase the signal-to-noise ratio in the data, the words not conveying information about the scientific content of a session (e.g. "therefore", "however", "follows") were removed. To do that we removed all the tokens of a non-statistical text included in the sessions texts. The non-statistical text we used was the Trayvon Martin corpus, a collection of political blog posts from 2012 [10]. At the end of the data cleaning process, the number of words of the vocabulary was reduced to 289 tokens.

Successively, LDA was carried out on cleaned data, setting a number of topics equal to 16. The quality of LDA results can be assessed through the study of the distinctive words that characterize the topics. Here, the distinctively of the j-th token for the k-th topic is the posterior probability of the k-th topic given that the j-th token appears in that session's text, for a uniform prior over the topics. Highly distinctive tokens can only be associated with one topic, whilst those with smaller posterior probabilities can be referred to with more topics. Table 2 reports the top 5 distinctive tokens of the topics.

Before carrying out the session assignment algorithm, the regularization of the topic distributions with a threshold equal to 0.05 was performed. In particular, considering the constraint above mentioned that solely sessions of the same type are scheduled in parallel at the same time band, three different optimization processes were performed, respectively for CP, SL and SP. Considering the parameter ϕ, we set it equal to 5 since the number of sessions taken into account is not large. Table 3 reports the results in terms of ρ values. Since ρ is invariant to the order of the time bands, it is important to highlight that it would be possible to change the order of the time bands without modifying the value of ρ.

The maximum of ρ is a theoretical value occurring in the case with perfectly no overlapping. Even if in practice that is not reachable, it can be considered a top benchmark. A minimum benchmark, instead, can be defined as a random assignment of sessions to the slots. Consequently, we scheduled randomly the sessions 100 times (respecting the constraints) and considered the lowest ρ obtained among them. In Table 3 the column *Actual* reports the ρ computed on the actual scheduling defined by the program committee, whist the *Proposed* is the best one obtained among the five ones defined by the optimization algorithm. To facilitate comparison between the values, the global values have been normalized according to the following formula $(\rho - \min)/(\max - \min)$. If the normalized value of the actual scheduling is 0.20, using the assignment algorithm the value increases up to 0.39,

Topic 1	Prob	Topic 2	Prob	Topic 3	Prob
spatio_tempor	0.920	binomi	0.905	infer	0.970
space_tim	0.865	bivari	0.787	approxim_bayesian	0.840
gaussian_process	0.800	von_mise	0.784	variabl_select	0.830
characterist	0.683	confid_interv	0.752	nonparametr_estim	0.779
covari_function	0.665	negat_binomi	0.698	prior_variabl_select	0.779
Topic 4	Prob	Topic 5	Prob	Topic 6	Prob
matric	0.790	composit_indic	0.876	function_data	0.920
covari_matric	0.759	suffer	0.808	rainfal	0.836
depend_time	0.759	partial	0.789	function_data_analysi	0.777
empir_likelihood	0.759	socio_econom	0.718	classif	0.763
p_spline	0.759	social_econom	0.702	data_analysi	0.694
Topic 7	Prob	Topic 8	Prob	Topic 9	Prob
decomposit	0.925	clinic_trial	0.885	mobil_data	0.885
status	0.907	cohort	0.837	complex_survey	0.873
census	0.899	generalis	0.837	matrix	0.851
insight	0.889	spatial_balanc	0.837	official_statist	0.828
instabl	0.843	depend_random	0.811	analyt	0.812
Topic 10	Prob	Topic 11	Prob	Topic 12	Prob
spars	0.876	general_linear	0.862	multilevel_model	0.869
princip_compon	0.842	linear_model	0.837	finit_mixtur	0.847
data_applic	0.806	exact	0.820	cross_sect	0.833
distribut_function	0.781	general_linear_model	0.775	miss	0.833
princip_compon_analysi	0.781	linear_mix	0.775	cross_sect_data	0.796
Topic 13	Prob	Topic 14	Prob	Topic 15	Prob
efficienc	0.938	quantil	0.954	nonparametr	0.946
stochast_frontier	0.866	quantil_regress	0.892	bayesian_nonparametr	0.942
robust	0.824	hidden	0.833	nonparametr_model	0.850
frontier_model	0.802	hidden_markov_model	0.769	mixtur_model	0.805
stochast_frontier_model	0.802	markov_model	0.769	skew_norm	0.741
Topic 16	Prob				
cluster	0.967				
social_network	0.829				
trim	0.687				
fuzzi	0.661				
depend_structur	0.643				

Table 2 Top 5 distinctive stemmed words of the 16 topics defined by LDA.

approximately the double, showing an important reduction of topics overlapping. Finally, in Table 4 we report the best scheduling of the sessions proposed by the algorithm for the SIS 2014.

5 Conclusions

The scientific meeting schedule is a challenging task. In this paper, we applied the method proposed by [4] to minimize overlapping between parallel sessions by

	Minimum	Actual	Proposed	Maximum
CP	33.27	36.21	37.65	45.00
SP	7.96	9.48	10.00	12.00
SL	24.31	25.29	28.50	36.00
Global values	65.54	70.99	76.16	93.00
Normalized impact	0.00	0.20	0.39	1.00

Table 3 ρ values. The first three rows concern the three kinds of sessions evaluated separately. The fourth row shows the global values, which is the sum of the above rows. The fifth row reports the normalized impact of the assignment.

Time band	Rooms					
	1	2	3	4	5	6
1	CP-18	CP-11	CP-17	CP-01	CP-15	CP-05
2	CP-14	CP-10	CP-03	CP-07	CP-06	CP-09
3	CP-16	CP-02	CP-08	CP-12	CP-04	CP-13
4	SP-05	SP-04	SP-06			
5	SP-03	SP-07	SP-01			
6	SP-08	SP-09	SP-11			
7	SP-10	SP-12	SP-02			
8	SL-07	SL-12	SL-03	SL-10		
9	SL-14	SL-01	SL-19	SL-05	SL-17	
10	SL-08	SL-04	SL-06	SL-15	SL-16	
11	SL-02	SL-09	SL11	SL-18	SL-13	

Table 4 Best schedule of the sessions proposed by the algorithm.

demonstrating its usefulness also for the program scheduling of the medium-sized conferences in addition to those of high magnitude as shown in their work. For this purpose, we considered the SIS 2014 conference. In 2014, the program committee developed a schedule that improved over a random assignment by 0.20 (with respect to minimizing overlapping content). The assignment of the sessions to the time bands proposed by the algorithm improves significantly the same program up to a score of 0.40 over one.

References

1. Arun, R., Suresh, V., Veni Madhavan, C. E., & Murthy, N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Pacific-Asia conference on knowledge discovery and data mining (pp. 391-402). Springer, Berlin, Heidelberg.
2. Benoit, K., Muhr, D., & Watanabe, K. (2019). stopwords: Multilingual stopword lists. R package version, 1.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
4. Frigau, L., Wu, Q., Banks, D. (2021). Optimizing the JSM Program. Journal of the American Statistical Association, 1-10.
5. Henry, T. R., Banks, D., Owens-Oas, D., & Chai, C. (2019). Modeling community structure and topics in dynamic text networks. Journal of Classification, 36(2), 322-349.

6. Kristiadi, D., & Hartanto, R. (2019). Genetic Algorithm for lecturing schedule optimization. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(1), 83-94.
7. Nanda, A., Pai, M. P., & Gole, A. (2012). An algorithm to automatically generate schedule for school lectures using a heuristic approach. International journal of machine learning and computing, 2(4), 492.
8. Porter, M. F. (2001). Snowball: A language for stemming algorithms.
9. Pupeikienė, L., Mockus, J. (2005). School schedule optimisation program. Information Technology and Control, 34(2).
10. Soriano, J., Au, T., & Banks, D. (2013). Text mining in computational advertising. Statistical Analysis and Data Mining: The ASA Data Science Journal, 6(4), 273-285.

Clustering artists based on the energy distributions of their songs on Spotify via the Common Atoms Model

Clustering di artisti in base alla distribuzione dell'energia delle loro canzoni su Spotify con il Common Atom Model

Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira

Abstract Partially exchangeable datasets are characterized by observations grouped into known, heterogeneous units. The recently developed Common Atoms Model (CAM) is a Bayesian nonparametric technique suited for analyzing this type of data. CAM induces a two-layered clustering structure: one across observations and another across units. In particular, the units are clustered according to their distributional similarities. In this article, we illustrate the versatility of CAM with an application to an openly available Spotify dataset. The dataset contains quantitative audio features for a large number of songs grouped by artists. After describing the data preprocessing steps, we employ CAM to group the Spotify artists according to the distributions of the energy of their songs.

Abstract *Gli insiemi di dati parzialmente scambiabili sono caratterizzati da osservazioni raggruppate in unità note ed eterogenee. Il Common Atoms Model (CAM), recentemente sviluppato, è una tecnica bayesiana non parametrica adatta all'analisi di questo tipo di dati. CAM induce una struttura di clustering a due livelli: uno fra le osservazioni e un altro fra le unità. In particolare, le unità sono raggruppate insieme secondo le similarità delle loro distribuzioni. In questo articolo, illustriamo la versatilità del CAM con un'applicazione a un dataset Spotify disponibile online. Il dataset contiene un gran numero di misurazioni di caratteristiche audio di canzoni raggruppate per artisti. Dopo aver descritto le fasi di preprocessing dei dati, impieghiamo CAM per raggruppare gli artisti di Spotify secondo le distribuzioni dell'energia delle loro canzoni.*

Key words: Common Atoms Model, partially exchangeable data, nested data, Spotify dataset, Kaggle, energy

---

Francesco Denti
Università Cattolica del Sacro Cuore, Milan; e-mail: francesco.denti@unicatt.it

Federico Camerlenghi
University of Milan - Bicocca

Michele Guindani
University of California - Irvine, US

Antonietta Mira
Università della Svizzera italiana, Lugano and University of insubria, Como

1 Introduction

Spotify [7] is a streaming company that gained a lot of popularity in the last decade. The company defines itself as a *"digital music, podcast, and video service"* that grants access to millions of songs and other content from creators all over the world. Remarkably, the Spotify Web API provides users with a wide variety of quantitative measurements about artists, albums, and track data, as well as audio features. The availability of these data sparked research interest in the interpretation and modeling of music features [See, for example, 2, 4, 6]. In this paper, we will use audio features data to perform a distributional cluster analysis. More in detail, we are interested in illustrating the applicability of the Common Atom Model (CAM), recently introduced by [3]. The CAM is useful when data are divided into different groups, called *units*, and one is interested in recovering a unit-level clustering. Here, we test CAM on a modern dataset of songs by different artists characterized by continuous measurements. The article is structured as follows. In Section 1.1 we introduce the dataset that we used and describe the preprocessing pipeline we followed. Section 2 briefly reviews the CAM, while Section 3 summarizes the distributional clustering results. Finally, Section 4 concludes and delineates future research directions.

1.1 Data description and preprocessing

For our study, we consider an open-source Spotify dataset available from the Kaggle platform[1]. The original dataset contains more than 160,000 songs published between 1921 and 2020, authored by more than 1,500 authors. For each song, various audio features have been quantified by Spotify using scores between 0 and 1. These audio features provide a description of each song's *mood* (e.g., danceability, energy), *properties* (e.g., loudness, speechiness), and *context* (e.g., liveness, acousticness). One can find more details about these features in the documentation available on the *Spotify for developer* webpage [2]. As an example, in this paper, we focus our attention on the quantitative feature named *energy*. We consider the songs (observations) as exchangeable data points "within" each artist (unit). Our goal is to cluster the artists based on the distributional similarities of the energy score of their songs. To simplify the terminology, we will talk about "energy distribution" for each artist in the rest of the paper. Before briefly introducing our model, we describe the preprocessing steps followed to prepare the data for the analysis. First, we notice that 20% of the songs contained in the dataset have been authored by more than a single artist/band. To simplify the analysis, we assign each of these songs to a single *representative artist* that we identify as the first singer in the list of coauthors. From a simple exploratory analysis, we also notice that the majority of the artists have authored a small number of songs. A representation of the fre-

[1] https://www.kaggle.com/ektanegi/spotifydata-19212020

[2] https://developer.spotify.com/discover/

quencies of the number of released tracks per artist is reported in the top-right panel of Figure 1. More than 90% of the artists have authored less than 20 songs. The limited number of authored songs for a specific artist could make the estimation of the corresponding energy distributions challenging. To solve this issue, we focus our attention on the most productive artists: we include only the ones who authored more than 100 songs in the analysis. We also filter out the authors with more than 200 songs (0.3% of the artists) to simultaneously limit the computational cost and remove potential outliers. The barplot in the left panel of Figure 1 reports how the remaining 20,270 songs are partitioned across 154 artists. The highlighted bars indicated artists associated with more than 150 tracks. Then, we filter out energy levels identically equal to 0 and 1 – mostly associated with silent tracks or applause in live tracks. Finally, we map the energy index from $(0,1)$ to the real line via a logit transform. We will refer to the new variable of interest as logit-energy. As we can see from the bottom-right panel of Figure 1, the remaining artists present heterogeneous logit-energy distributions. As an illustration, we highlighted the distributions of the logit-energy for `AC/DC`, `Benny Goodman`, `Franz Schubert`, `Green Day`, and `The Doors` with different colors.

2 CAM for continuous data

In this section, we briefly review the CAM for nested data, introduced in [3]. Denote the logit-energy value for song i of artist j with $y_{i,j}$, where $i = 1,\ldots,n_j$ and $j = 1,\ldots,J$. Then, we indicate with G_j the distribution of the j-th experimental unit (artist). Under the partial exchangeability assumption of our data, we can write $y_{i,j}|\, G_1,\ldots,G_J \overset{ind.}{\sim} G_j$, independently across $i = 1,\ldots,n_j$ and $j = 1,\ldots,J$. These random variables take values over the real line $\mathbb{R}$, equipped with the Borel σ-field $\mathscr{B}$. The overarching goal is to induce a two-layer clustering across the observations (songs) and distributions (artists). Thus, the G_j's are assumed to be sampled from an almost surely discrete distribution Q over the space of probability distributions on $\mathscr{B}$, namely

$$G_1,\ldots,G_J|Q \overset{i.i.d.}{\sim} Q, \qquad\qquad Q = \sum_{k\geq 1} \pi_k\, \delta_{G_k^*}, \tag{1}$$

where $G_k^* = \sum_{l\geq 1} \omega_{l,k}\, \delta_{\theta_l}$, $k \geq 1$. Note that this model is a suitable modification of the nested Dirichlet process [5], which does not suffer from the degeneracy issue outlined by [1]. The G_k^*'s share the same set of atoms, $\theta_1, \theta_2, \ldots$, which are sampled from a non-atomic base measure H on $(\mathbb{R}, \mathscr{B})$. A stick–breaking representation is assumed for both the weights of the mixtures at the observational ($\omega_{l,k}$) and distributional (π_k) levels. Dealing with continuous data, it is better to convolute the discrete random measures with continuous parametric kernels $p(\cdot|\theta)$ –in our case, assumed to be Gaussian– obtaining:

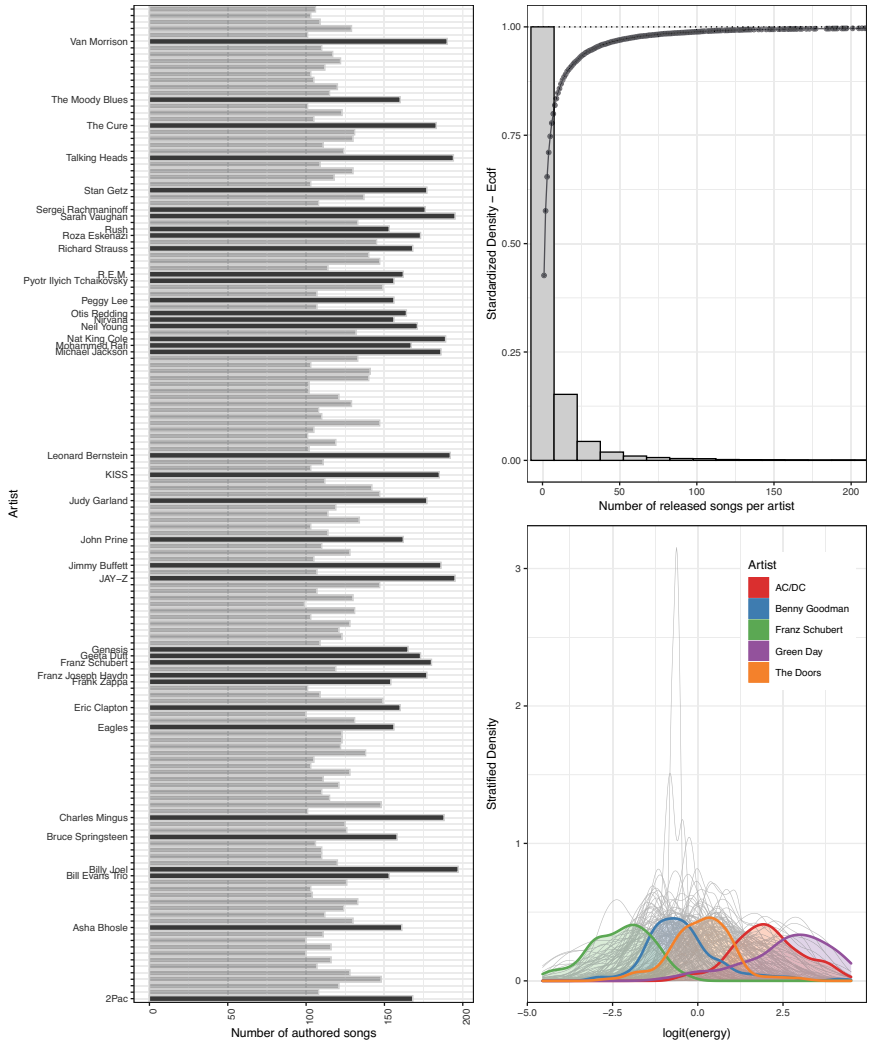


Fig. 1 Descriptive plots of the `Spotify` dataset. Left panel: barplots displaying the number of authored songs for each selected artist. Top right: histogram and empirical c.d.f. for the number of songs per artist (entire dataset). Bottom-right: density plots of the energy distributions stratified by selected artists.

$$
\begin{aligned}
(y_{i_1,1},\dots,y_{i_J,J})\,|\,f_1,\dots,f_J &\overset{ind.}{\sim} f_1\times\cdots\times f_J \qquad i_j=1,\dots,n_j,\ j=1,\dots,J,\\
f_j(\cdot) &= \int_\Theta p(\cdot|\theta)\,G_j(d\theta), \quad j=1,\dots,J.
\end{aligned}
\tag{2}
$$

The discrete nature of Q induces a clustering across the distributions, which is the quantity of interest in our analysis.

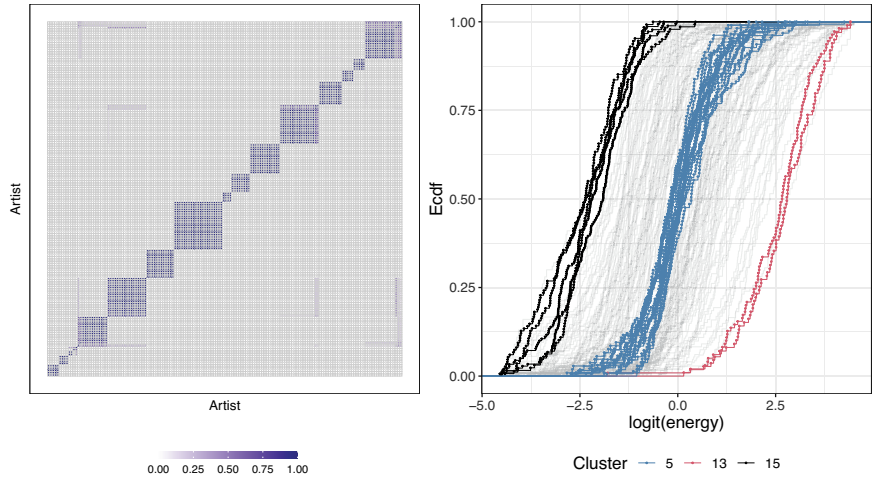


Fig. 2 Left panel: posterior co-clustering matrix between artists. Right panel: empirical cdfs for every logit-energy distribution considered. Three distributional clusters are highlighted.

3 Distributional clustering results

We run the nested slice sampler –a tailored algorithm developed to fit CAM– for 30,000 iterations, and we discard the first 20,000 as burn-in period. The concentration parameters for the outer and inner stick-breaking processes are fixed to 1. The left panel of Figure 2 displays the posterior co-clustering matrix across artists. Well-separated clusters are clearly visible. Estimating the best partition based on the minimization of the Variation of Information [8] leads to the detection of 16 groups. The right panel of Figure 2 reports all the empirical cdfs of the logit-energy distribution for each artist. To briefly illustrate the results of the distributional clustering, we highlight the functions assigned to three clusters: 5, 13, and 15. These indexes were chosen to exemplify distributional groups characterized by low, average, and high logit-energy, respectively. For example, cluster 13 contains two high logit-energy artists: `blink-182` and `Iron Maiden`. At the same time, cluster 15 contains mostly classical music composers, and it is characterized by low logit-energy values. A summary of these representative results is reported in Table 1.

4 Conclusion

We illustrated how the recently proposed CAM could be applied to music features data to estimate clusters of artists whose discographies share similar distributional characteristics. Our application highlights the versatility of this BNP method, espe-

Cluster id	# Assigned artists	Avg. logit-energy	Examples of members
5	17	0.07 (0.933)	`Eagles`, `Eric Clapton`, `Lana Del Rey`, ...
13	2	2.62 (0.948)	`blink-182`, `Iron Maiden`
15	5	-2.26 (0.956)	`Schubert`, `Ravel`, ...

Table 1 Summary of the characteristics of three representative distributional clusters obtained with CAM. For each selected cluster, the table contains its number of members, the average of logit-energy (and std. dev.), and a few examples of notable members.

cially given the complexity of the data and the large sample size. As a future direction, we aim to develop a variational inference version of the algorithm to scale up its application to even larger datasets. Once the most challenging computational issues are addressed, CAM could also be extended to multivariate settings, enabling the joint modeling of multiple music features.

References

[1] Federico Camerlenghi, David B. Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis*, 14(4), 2019.

[2] Christopher E. Jr.; Dawson, Steve; Mann, Edward; Roske, and Gauthier Vasseur. Spotify: You have a Hit! *SMU Data Science Review*, 5(3), 2021.

[3] Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira. A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, pages 1–12, 2021.

[4] Claire Howlin and Brendan Rooney. Patients choose music with high energy, danceability, and lyrics in analgesic music listening interventions. *Psychology of Music*, 49(4):931–944, 2021.

[5] Abel Rodríguez, David B. Dunson, and Alan E. Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1144, 2008.

[6] Mariangela Sciandra and Irene Carola Spera. A model-based approach to Spotify data analysis: a Beta GLMM. *Journal of Applied Statistics*, 49(1):214–229, 2022.

[7] Spotify. Spotify WebAPI. *Spotify USA INC.*, 2019.

[8] Sara Wade and Zoubin Ghahramani. Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

Hidden Markov models for four-way data

Modelli di Markov nascosti per dati a quattro vie

Salvatore D. Tomarchio and Antonio Punzo and Antonello Maruotti

Abstract Hidden Markov models (HMMs) constitute a powerful device for the modelization of heterogeneous longitudinal data. In this work, we discuss a family of HMMs for the analysis of four-way data. To introduce parsimony in the considered models, we use the eigen-decomposition of the components covariance matrices. The performances of our family of models are investigated on simulated data and comparisons with reference parsimonious models for three-way data, after data rearrangement in this form factor, are conducted.

Abstract *I modelli di Markov nascosti (HMMs) costituiscono un potente strumento per la modellizzazione di dati longitudinali eterogenei. In questo lavoro, discutiamo una famiglia di HMMs per l'analisi di dati a quattro vie. Per introdurre parsimonia nei modelli considerati, usiamo la decomposizione spettrale delle matrici di varianza e covarianza delle componenti del modello. Le prestazioni della nostra famiglia di modelli sono studiate su dati simulati e vengono condotti confronti con modelli parsimoniosi di riferimento per dati a tre vie, dopo il riarrangiamento dei dati in questo fattore di forma.*

Key words: Hidden Markov models, Model-based clustering

Salvatore D. Tomarchio
Dipartimento di Economia e Impresa, Università degli Studi di Catania, Catania, Italia e-mail: daniele.tomarchio@unict.it

Antonio Punzo
Dipartimento di Economia e Impresa, Università degli Studi di Catania, Catania, Italia e-mail: antonio.punzo@unict.it

Antonello Maruotti
Dipartimento GEPLI, Libera Università Maria Ss Assunta, Roma, Italia e-mail: a.maruotti@lumsa.it

1 Introduction

Hidden Markov models (HMMs) have been widely used in time series analysis, and they also provide a general-purpose setting for dealing with different application domains (see, e.g. [17] for a survey). A relatively recent field of application involves the use of HMMs for modeling longitudinal (or panel) data. Longitudinal data are generally characterized by serial dependence and heterogeneity in the sample units, that can be properly investigated and accounted for in an HMM framework [5]. Being dependent mixture models, HMMs allow the recover of the data structure by defining homogenous latent subgroups and, simultaneously, provide meaningful interpretation of the inferred partition. Furthermore, the way in which sample units move between the states can provide useful information along with the transition probabilities.

For univariate or multivariate longitudinal data, several HMMs have been proposed in the literature (see, e.g. [4, 8, 9]). However, in the recent years there has been an increased interest in the analysis of three-way data [3, 6, 10, 12, 13, 15, 16], where $P \times R$ matrices are observed on N sample units. Unfortunately, when the time T is indexed on either the rows or the columns of the matrices, the type of longitudinal data that can be analyzed in a three-way setting is reduced. Additionally, it is not possible for the sample units to move between the states over time as well as to fully understand the evolution of a certain behavior or phenomenon across time. To solve both issues, HMMs for three-way longitudinal data are herein discussed. The data are conveniently arranged in four-way arrays of dimension $P \times R \times N \times T$. However, such data structure can lead to overparameterization issues, especially because of the components covariance matrices. Therefore, we use the well-known eigen-decomposition of the components covariance matrices to address this problem [2, 10]. By using this approach, a family of 98 parsimonious HMMs, labeled MV-HMMs, is obtained and presented in Sect. 2.

In Sect. 3, we firstly assess the parameter recovery and model selection of our algorithm via simulated data. Moreover, over the same data, we investigate the differences between our models and a reference approach fitted on the rearranged data in a three-way structure. We highlight the drawbacks that such a procedure might cause and the better results obtained by considering the family of models herein discussed. Conclusions and final remarks, along with possible future extensions, are presented in Sect. 4.

2 Methodology

In the modelization of four-way arrays via HMMs, we assume the existence of the following two processes: an unobservable finite-state first-order Markov chain defined as $\{S_{it}; i = 1,\dots,N, t = 1,\dots,T\}$, with state space $\{1,\dots,K\}$ and being K the number of states, and an observed process defined as $\{\mathscr{X}_{it}; i = 1,\dots,N, t = 1,\dots,T\}$, where $\mathscr{X}_{it}$ denotes the $P \times R$ matrix for individual i at time t. We also assume that for

the state-dependent observation process $\{\mathscr{X}_{it}\}$ the conditional independence property holds, i.e.

$$f(\mathscr{X}_{it} = \mathbf{X}_{it} | \mathscr{X}_{i1} = \mathbf{X}_{i1}, \ldots, \mathscr{X}_{it-1} = \mathbf{X}_{it-1}, S_{i1} = s_{i1} \ldots, S_{it} = s_{it}) \\ = f(\mathscr{X}_{it} = \mathbf{X}_{it} | S_{it} = s_{it}),$$

where $f(\cdot)$ is the probability density function (pdf) of the matrix-variate normal distribution, i.e.

$$\phi(\mathbf{X}_{it} | S_{it} = k; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k) = \frac{\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_k^{-1}(\mathbf{X} - \mathbf{M}_k)\boldsymbol{\Psi}_k^{-1}(\mathbf{X} - \mathbf{M}_k)'\right]\right\}}{(2\pi)^{\frac{PR}{2}} |\boldsymbol{\Sigma}_k|^{\frac{R}{2}} |\boldsymbol{\Psi}_k|^{\frac{P}{2}}}, \quad (1)$$

where $\mathbf{M}_k$ is the $P \times R$ mean matrix, $\boldsymbol{\Sigma}_k$ is the $P \times P$ covariance matrix associated with the rows, $\boldsymbol{\Psi}_k$ is the $R \times R$ covariance matrix related to the columns and $\operatorname{tr}(\cdot)$ is the trace operator.

Other than the parameters of the state-dependent pdfs, we also have those related to the Markov chain. In particular, the parameters of the Markov chain are the initial probabilities $\pi_{ik} = \Pr(S_{i1} = k)$, $k = 1, \ldots, K$, and the transition probabilities

$$\pi_{ik|j} = \Pr(S_{it} = k | S_{it-1} = j), \;\; t = 2, \ldots, T \;\text{ and }\; j, k = 1, \ldots, K,$$

where k refers to the current state and j refers to the one previously visited. The initial probabilities are collected in the K-dimensional vector $\boldsymbol{\pi}$, while the transition probabilities are inserted in the $K \times K$ transition matrix $\boldsymbol{\Pi}$.

As mentioned in Sect. 1, to introduce parsimony in our MV-HMMs, we apply the eigen-decomposition to the covariance matrices of the state-dependent pdfs. We recall that a generic $Q \times Q$ component covariance matrix can be decomposed as

$$\boldsymbol{\Phi}_k = \lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k', \quad (2)$$

where $\lambda_k = |\boldsymbol{\Phi}_g|^{1/q}$, $\boldsymbol{\Gamma}_k$ is a $q \times q$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{\Phi}_g$, and $\boldsymbol{\Delta}_k$ is the scaled ($|\boldsymbol{\Delta}_k| = 1$) diagonal matrix of the eigenvalues of $\boldsymbol{\Phi}_k$. These elements correspond respectively to the volume, orientation and shape of the kth state. By constraining the three components in (2), the following 14 parsimonious structures are obtained: EII, VII, EEI, VEI, EVI, VVI, EEE, VEE, EVE, VVE, EEV, VEV, EVV, VVV, where "E" means equal, "V" stands for varying and "I" denotes the identity matrix.

It must be noted that we do not obtain 14 parsimonious structures for both $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Psi}_k$. Indeed, the following restriction $|\boldsymbol{\Psi}_k| = 1$ is imposed to avoid an identifiability issue. This makes the λ_k parameter unnecessary in the decomposition of $\boldsymbol{\Psi}_k$, and reduces from 14 to 7 the parsimonious structures for this covariance matrix: II, EI, VI, EE, VE, EV, VV. Therefore, we globally obtain a total of $14 \times 7 = 98$ parsimonious structures producing the family of MV-HMMs discussed in this paper.

To fit the models of our family, we use an expectation-conditional maximization (ECM) algorithm [7]. Useful insights for the implementation of our ECM algo-

rithm can be gained in [1, 2, 10, 14]. Our ECM algorithm is initialized by using the approach discussed in [12], where a generalization of the short-EM initialization strategy has been implemented.

3 Simulated analyses

In this section, we examine different aspects via simulated data. Considering the high number of models proposed, we will only focus on one of them for illustrative purposes. In detail, we consider the VVE-VE MV-HMM. We set $P = R = 2$, $N = 200$, $K = 3$ and generate data from the considered model having the following parameters $\boldsymbol{\pi} = (0.33, 0.33, 0.34)$,

$$\boldsymbol{\Pi} = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.05 & 0.70 & 0.25 \\ 0.00 & 0.15 & 0.85 \end{bmatrix}, \quad \mathbf{M}_1 = \begin{bmatrix} 2.00 & 3.00 \\ -1.00 & -1.00 \end{bmatrix}.$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.85 & 0.29 \\ 0.29 & 0.85 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.50 & 0.30 \\ 0.30 & 0.50 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.45 & 1.00 \\ 1.00 & 1.45 \end{bmatrix},$$

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 1.06 & 0.36 \\ 0.36 & 1.06 \end{bmatrix}, \boldsymbol{\Psi}_2 = \begin{bmatrix} 1.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}, \boldsymbol{\Psi}_3 = \begin{bmatrix} 1.45 & 1.05 \\ 1.05 & 1.45 \end{bmatrix}.$$

To obtain $\mathbf{M}_2$ and $\mathbf{M}_3$ we added a constant c to each element of $\mathbf{M}_1$. Specifically, we set $c = 4$ for obtaining $\mathbf{M}_2$ and $c = 8$ to get $\mathbf{M}_3$. We also consider three values for T, i.e. $T \in \{5, 10, 15\}$, and for each value of T we generate 50 datasets.

First of all, we fit over the simulated datasets the VVE-VE MV-HMM with $K = 3$ to evaluate the parameter recovery of our algorithm. Considering the high number of parameters that should be reported, we calculate the average among the mean square errors (MSEs) of the elements of each estimated parameter, over the $K = 3$ states and for each T, allowing us to summarize in a single number the MSE of each parameter. As we can see, the MSEs can be considered negligible for each

Table 1 Average MSEs of the parameter estimates for the VVE-VE MV-HMM. The average is computed among the MSEs of the elements of each estimated parameter, over the $K = 3$ states and 50 datasets for each T.

Parameter	$T = 5$	$T = 10$	$T = 15$
$\mathbf{M}$	0.0077	0.0045	0.0032
$\boldsymbol{\Sigma}$	0.0048	0.0029	0.0018
$\boldsymbol{\Psi}$	0.0036	0.0023	0.0018
$\boldsymbol{\pi}$	0.0020	0.0028	0.0018
$\boldsymbol{\Pi}$	0.0009	0.0007	0.0004

parameter and for each value of T. Additionally, it is interesting to note that their values generally become better with the increase of T.

We now investigate the capability of the Bayesian information criterion (BIC; [11]) in identifying the true parsimonious structure of the data. This is because we need to assess if the BIC, which is one of the most famous and used tools in model-based clustering, accurately works. Specifically, we fitted all the models in our family with $K = 3$ to the generated datasets. We report that regardless of the value of T, the BIC has always correctly identified the parsimonious structure of the data generating model.

A further analysis compares the performance of our models with those of an alternative reference approach that could be used if our models were not available. In detail, for each simulated dataset, we vectorize the $P \times R$ matrices of the statistical units into PR-dimensional vectors, thus obtaining a $PR \times N \times T$ array. Then, on these rearranged datasets, parsimonious multivariate normal HMMs (M-HMMs) are fitted. The results of such a comparison in terms of average BIC are reported in Table 2. Notice that, the BICs for M-HMMs refer to best fitting model over each dataset among the 14 available parsimonious models.

Table 2 Average BIC computed over 50 datasets for each T and competing model.

Model	$T = 5$	$T = 10$	$T = 15$
MV-HMMs	**5776.966**	**11336.02**	**16893.54**
M-HMMs	5828.779	11401.21	16958.32

We notice that the MV-HMMs always overwhelm the multivariate models. There are several reasons that lead to these differences. First of all, data vectorization increases the number of parameters of the models that can have negative effects on model selection. Indeed, data vectorization can cause an underestimation of the mixture order as well as increase the penalty term of information criteria [10, 13]. Secondly, the vectorization completely destroy the information contained in the component covariance matrices, since we would replace the row and covariance matrices with a unique (and higher dimensional) covariance matrix. Thus, other than severely increase the risk of overparameterization issues, this process reduces the interpretability and the fitting behavior of the obtained models.

4 Conclusions

In this paper, parsimonious hidden Markov models for four-way data have been discussed. Parsimony has been introduced via the eigen decomposition of the state covariance matrices, producing a family of 98 HMMs. By using simulated data, we have shown the capability of the estimation algorithm in recovering the parameters of the data generating model. When all the models in our family is fitted to the simu-

lated data, the BIC has proven to be able to detect the true parsimonious structure in the data. Furthermore, when compared to multivariate parsimonious hidden Markov models, our approach has provided better fitting results. Additionally, our models can attain the overparameterization issues and avoid the loss of information caused by the vectorization process.

There are different possibilities for further works. We could mention the extension of our models by using skewed or heavy tailed state-dependent probability density functions, or the inclusion of a set of covariates in a regression framework.

References

1. Baum, L. E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**(1), 164–171 (1970)
2. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**(5), 781–793 (1995)
3. Gallaugher, M.P.B., McNicholas P.D.: Finite mixtures of skewed matrix variate distributions. Pattern Recognit. **80**, 83–93 (2018)
4. Holzmann, H., Schwaiger, F.: Hidden Markov models with state-dependent mixtures: minimal representation, model testing and applications to clustering. Stat. Comput. **25**(6), 1185–1200 (2015)
5. Maruotti, A.: Mixed hidden markov models for longitudinal data: An overview. Int. Stat. Rev. **79**(3), 427–454 (2011)
6. Melnykov, V., Zhu, X.: Studying crime trends in the USA over the years 2000–2012. Adv. Data Anal. Classif. **13**(1), 325–341 (2019)
7. Meng, X.L., Van Dyk, D.: The EM algorithm-an old folk-song sung to a fast new tune. J. Royal Stat. Soc. B. **59**(3), 511–567 (1997)
8. Punzo, A., Maruotti, A.: Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. J. Comput. Graph. Stat. **25**(4), 1097–1098 (2016)
9. Punzo, A., Ingrassia, S., Maruotti, A.: Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. Stat. Med. **37**(19), 2797–2808 (2018)
10. Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)
11. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
12. Tomarchio, S.D., Punzo, A., Bagnato, L.: Two new matrix-variate distributions with application in model-based clustering. Comput. Stat. Data Anal. **152**, 107050 (2020)
13. Tomarchio, S.D., McNicholas, P.D., Punzo, A.: Matrix Normal Cluster-Weighted Models. J. Classif. **38**(3), 556–575 (2021)
14. Tomarchio, S.D., Punzo, A., Maruotti, A.: Parsimonious Hidden Markov Models for Matrix-Variate Longitudinal Data. arXiv:2107.04330 (2021)
15. Tomarchio, S.D., Gallaugher, M.P.B., Punzo, A., McNicholas, P.D.: Mixtures of Matrix-Variate Contaminated Normal Distributions. J. Comput. Gr. Stat. 1–22 (2022)
16. Tomarchio, S.D., Ingrassia, S., Melnykov, V.: Modelling students' career indicators via mixtures of parsimonious matrix-normal distributions. Aust N. Z. J. Stat. 1–16 (2022)
17. Zucchini, W., MacDonald, I.L.: Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC (2009)

Family demography

Does family of origin make a difference in occupational outcomes?

La famiglia d'origine fa la differenza in termini di risultati occupazionali?

Annalisa Busetta[1], Elena Fabrizi[2], Isabella Sulis[3], Giancarlo Ragozini[4]

Abstract

Disadvantages faced by parents adversely affect their children's chances of success in the labour market. We study the influence of intergenerational transmission of parental socio-economic background on the educational attainment and occupational outcome of children, also considering gender differences. To tackle such a complex system of relationships across the outcome variables (both exogenous and endogenous), we adopt a path analysis model. In particular, we study the intergenerational transmission of disadvantage using the innovative and rich AD-SILC database, which shows the evolution of occupational outcomes over eight years (measured by wages in 2011 and 2018). Our findings indicate that being raised in a single-parent family negatively affects men's education and wages. Furthermore, high levels of education of at least one parent positively affect children; this effect is especially evident for daughters who grew up with fathers with low education levels.

Abstract

Gli svantaggi affrontati dai genitori influiscono negativamente sulle possibilità di successo nel mercato del lavoro dei loro figli. In questo lavoro studiamo l'effetto della trasmissione intergenerazionale del background socio-economico della famiglia d'origine sull'istruzione e occupazione dei figli, tenendo in considerazione anche un'ottica di genere. Per tenere sotto controllo il complesso sistema di relazioni tra le variabili esogene ed endogene coinvolte, abbiamo adottato un modello di path analysis. In particolare attraverso il ricco e innovativo database AD-SILC è stato possibile studiare l'impatto della famiglia d'origine sull'evoluzione degli esiti occupazionali nel tempo (misurati attraverso il reddito al 2011 e al 2018). I risultati hanno mostrato che avere un genitore con un livello di istruzione post-secondario costituisce un vantaggio in termini di reddito per i figli, ma anche che avere una madre con un livello di istruzione post-secondario fa realmente la differenza per le donne che crescono in una famiglia in cui il padre ha un basso livello di istruzione.

Key words: Intergenerational transmission, social mobility, path models, gender studies, labour market

[1] Annalisa Busetta, University of Palermo. E-mail: annalisa.busetta@unipa.it
[2] Elena Fabrizi, University of Teramo. E-mail: efabrizi@unite.it
[3] Isabella Sulis, University of Cagliari. E-mail: isulis@unica.it
[4] Gaincarlo Ragozini, University of Naples Federico II. E-mail: giancarlo.ragozini@unina.it

1 Introduction

Disadvantages faced by parents adversely affect their children's chances of success. Previous studies have focused on the association between social origins and individuals' educational attainments, the association between individuals' education and their occupational outcomes over time and/or across countries [10, 12, 13] and the intergenerational transmission of socio-economic advantages/disadvantages.
The theory of education as the great equaliser assumes that education has the potential to balance out inequalities in society related to people's initial disadvantaged conditions. According to this framework, policies intended to remove obstacles to accessing higher educational levels ensure equal education opportunities despite differences in socio-economic origin and prevent inequalities in future employment opportunities and economic rewards [5]. Thus, education, especially in developed countries, is considered the main social elevator – able to activate processes of intergenerational social mobility [4].
To assess the complex system of hypothesised relationships across the exogenous and endogenous outcome variables involved, we start from the so-called social origin–education–destination (OED) triangle, which represents the basic processes underlying the intergenerational reproduction of inequality [5] The complex system of relationships across the (exogenous and endogenous) outcome variables involved in the OED model was estimated using a path analysis model. Using this model, we can evaluate the role that the family of origin's socio-economic conditions play in the intergenerational transmission of inequalities in the Italian context and, in particular, in occupational outcomes, hereafter measured by individual wage. To assess the mediating role of education in the intergenerational reproduction of inequalities, we estimate both the direct and indirect effects of a) social origin on educational attainment (*educational inequality*), b) educational attainment on individual work history (*occupational returns to education*) and c) social origin on individual career over and above individuals' differences in achieved education (*social background on occupation*).
We study the influence of the socio-economic conditions of the family of origin on educational and occupational success at different points across individuals' work history [8]. To this aim, we leverage the innovative and rich Admistrative Data-Statistics on Income and Living Conditions (AD-SILC) database, which provides a unique opportunity to analyse the intergenerational transmission of disadvantages and its long-terms effects by showing occupational outcomes at two points in time.

In addition to parental socio-economic characteristics, this paper considers the role of family disruption in individual's educational and occupational outcomes. Studies on the intergenerational consequences of family disruption have suggested that individuals who spend part of their childhood in one-parent families tend to marry and have children early and experience nonmarital childbearing and separation or divorce [11]. Moreover, they experience short-term decline in physical and

psychological well-being and longer-term reductions in educational achievement and economic security [3].
Finally, to examine gender differences in the effect of family background on occupational success, we stratify the models by gender.

2 Data

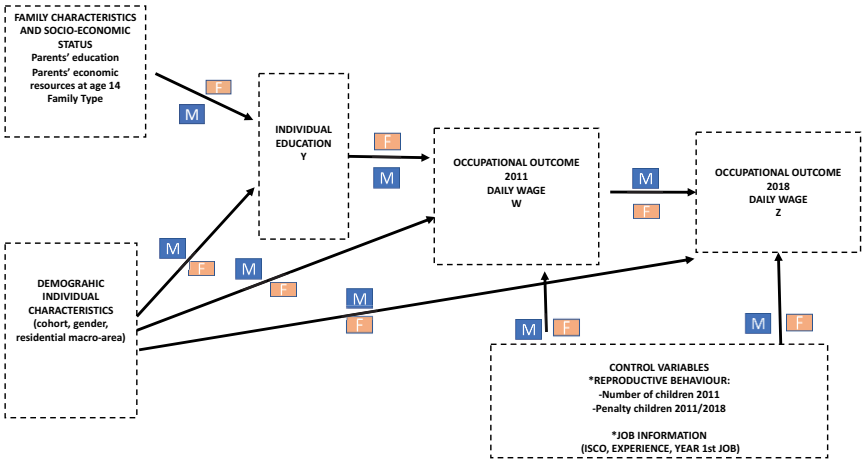
We use the AD-SILC[1] database, which is constructed by matching longitudinal information from administrative archives held by the National Institute of Social Security (INPS) with survey data collected by the National Institute of Statistics (ISTAT). In our database, we have information on the socio-economic conditions of the interviewees (from the 2011 IT-SILC survey, i.e. the Italian component of the European SILC survey) and their individual working career histories (collected in the administrative archives) from 2011 to 2018. Thanks to the 2011 IT-SILC special module on the intergenerational transmission of disadvantage, we also have the information on the interviewee's parents when the individual was 14 years old (i.e. education, occupation and difficulties in making ends meet). To exclude individuals who have not yet completed their education, the youngest cohort (25–29 years old) was not considered in the analysis. Hence, we concentrate on only four cohorts of offspring aged 30–34, 35–39, 40–44 and 45–49 years old, respectively, in 2011.
Using the path analysis model (depicted in Figure 1), we can simultaneously consider the effect of family socio-economic conditions on educational achievement (Y) and on occupational outcomes in 2011 (W) and 2018 (Z). In particular, the model specifies three concatenated linear regression models for the three outcome variables Y, Z and W, which are linked in chronological order inside the path analysis model; educational attainment Y influences occupational outcome W observed in 2011, which, in turn, influences occupational outcome in 2018, namely Z. The occupational outcomes W and Z are measured by the gross hourly earnings (including personal income taxes and social contributions) divided by the worked days. The income of the offspring is referred to as daily wages.

The main independent variables are those regarding the social origin (e.g. socio-economic status of both parents). They are measured by the education of the two parents and the ability of the family to make ends meet, both when the individual is 14 years old. One advantage of using parental education as a proxy for parental income is that education is likely to be a more permanent feature than current wages, while being highly correlated with wages in most countries [6].

[1] The AD-SILC database was built by the Italian Department of the Treasury for a European Union-funded research project entitled 'Modernising the social protection system in Italy' (Mospi) in response to the 'Call for proposals on social innovation and national reforms' and 'Access to social protection and national reform support' (Vp/2018/0103).

Specifically, three types of variables are involved in the specified regression framework. The exogenous set of predictors relates to the following categories: i) the socio-economic status of the family of origin when the respondent was 14 years old (parents' level of education,[1] parents' economic resources and type of family, i.e. single-parent versus two-parent families); ii) individuals' characteristics (e.g. cohort of birth, gender, education and geographical area); iii) control variables, such as reproductive behaviour[2] (measure by the number of children in 2011 and in the time span 2011–2018) and parental leave for both men and women (measured in days); and iv) information related to the individual's job in 2011 (e.g. year of the first job, cumulative time of employment, job qualification encoded by ISCO level). The four sets of predictors have direct and indirect effects on the three outcome variables (i.e. educational achievement and occupational outcome in 2011 and 2018).

Figure 1. *Path analysis hybrid model for intergenerational reproduction of inequality from a gender perspective*



Source: Authors' elaborations

Two endogenous mediator outcome variables are specified – namely, the individual ISCED level of education (expressed as the average number of years required to reach the corresponding level of education) and the daily wage in 2011. The causal relationship between these variables allows us to capture the short- and medium-term returns, respectively, of education on occupational outcomes. Finally, the two

[1] Because including both parents significantly changes the results of the trend of educational inequalities in opportunities over cohorts [2], we opt to include parental education with a 'full interaction' coding.
[2] Although we are conscious that parental and individual characteristics also influence reproductive behaviour, in the path analysis model, we included reproductive behaviour variables only as control variables.

mediator variables (ISCED and daily wage 2011) have both direct and indirect effects that explain differences in the daily wage in 2018. The latter equation captures the long-term effect of education and the 2011 wage on long-term occupational outcome (daily wage in 2018).

3 Preliminary Results

Using the path analysis model [1, 7, 9] depicted in Figure 1, we can test specific hypotheses on the direct and indirect effects of family socio-economic status on individual educational achievement and the short- and long-term intergenerational economic returns of exogenous and mediator variables on daily wage monitored in 2011 and 2018.

As a result, a hybrid model is specified, where the effect of the family's socio-economic conditions (parents' education and economic resources) on occupational outcomes is fully mediated by the individual's educational level, whereas the individual's demographic characteristics and reproductive behaviour before and after 2011 have direct effects on both occupational outcomes. The hybrid model is coherent with the theory of education as the great equaliser. It seems that once two individuals achieve the same level of education, they become equal and have almost the same chances of success in the labour market, even though they differ in a number of other important characteristics.

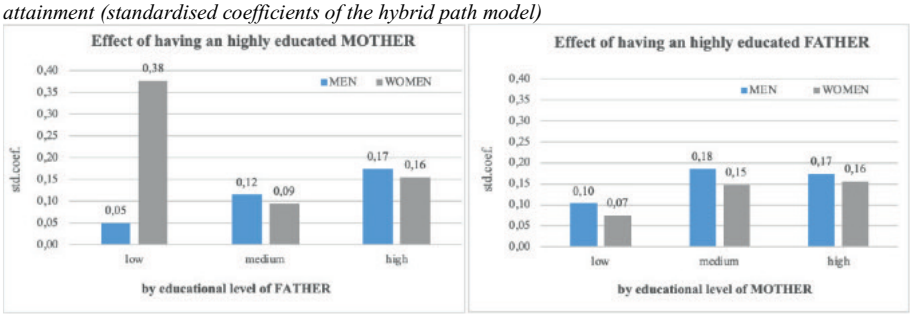
We assess the presence of gender bias in the effect of socio-economic status on both indicators of occupational outcomes by estimating two separate models for men and women. The comparison between the direct and indirect effects of family socio-economic conditions for the two genders highlights gender bias in the intergenerational transmission of inequalities. The full results are reported in Table A.

The first equation on individual educational level (outcome variable Y) suggests the presence of a gender bias in the transmission of family educational capital. The positive association between individual educational achievement and parents' education is generally stronger for men. It is worth noting that the highest positive effect on individual education is registered for women who have a mother with a high level of education and a father with a low level (Figure 2). In contrast, the magnitude of the effect shrinks for men when the father has a high level of education.

The family's economic conditions when the individual is 14 years old (measured by the ability to make ends meet) also play a more significant role in educational achievement for women than for men. As expected, younger cohorts are better educated, and individuals from southern regions are, on average, penalised in comparison to those from other regions (see Table A).

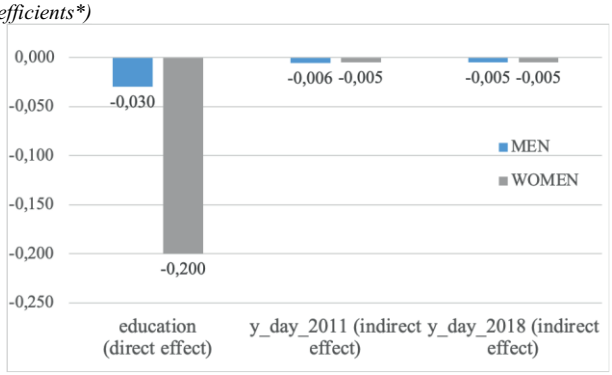
When other factors are equal, family disruption seems to have a small but significant direct effect on men's educational levels (see Figure 3).

Figure 2. *Direct effect of educational level of mother and father on women's and men's educational attainment (standardised coefficients of the hybrid path model)*



*All effects are significant, with p value < 0.05
Source: Authors' elaborations on AD-SILC data

Figure 3. *Penalising effect of growing up in a single-parent family on the three dependent variables (standardised coefficients*)*



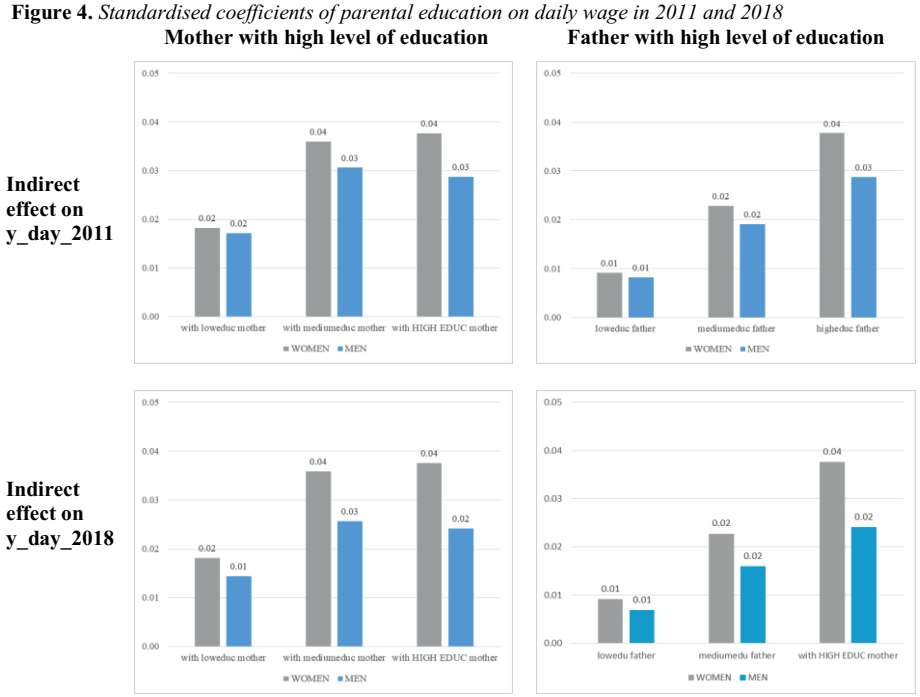
* The men coefficient is significant, with p value < 0.05. The women coefficients are not significant.
Source: Authors' elaborations on AD-SILC data

The individual level of education is the variable with the strongest direct effect on occupational outcome in 2011 (outcome variable W), with higher returns for women.

The education of the parents and their ability to make ends meet both significantly affect interviewee's occupational outcomes (daily wage in 2011 and 2018). These latter effects are larger in magnitude with respect to the effects of the interviewee educational level for both men and women. Parental socio-economic conditions also significantly affect job qualification (measured using the classification of occupations in five categories using the ISCO levels). The indirect effect of parents' education on occupational outcome in 2011 is stronger for men than women, as well as for individuals from highly educated families; in contrast, the findings reveal a gender bias in favour of women in the indirect positive effect of the family's economic conditions on individuals' daily wages in 2011 (see Figure 4). However, the former effect is stronger than the latter, suggesting the persistence of gender inequality mechanisms in the intergenerational transmission of economic and educational capital, with a clear bias in favour of men in the occupational attainment indicators. This result also holds true for women from highly educated families.

When other factors are equal, being raised in a single-parent family slightly indirectly influenced the average wage of men in both 2011 and 2018 (see Figure 3).

Figure 4. *Standardised coefficients of parental education on daily wage in 2011 and 2018*



*All effects are significant, with p value < 0.05
Source: Authors' elaborations on AD-SILC data

Finally, both models on the economic returns in 2011 and 2018 control for the reproductive behaviour of individuals. The models show disadvantage in the labour market for women that are mothers: the number of children is negatively associated with women's average wages but positively associated with men's. The opposite role played by the number of children in the men and women models deserves consideration. On the one hand, it is possible that men with a high number of children or a growing family decide to invest more in their careers to guarantee a higher income. On the other hand, men in stable and well-paid positions in the labour market – or with expectations of a rapid career in the coming years – may have the economic possibility to realise their fertility intentions or decide to have a greater number of children.

What is clear from our model is that having children penalises only women's occupational outcomes. Moreover, this penalisation becomes greater the closer the parental leave is to 2011 and 2018 and the longer it endures. The number of children had before 2011 also has a negative indirect impact on the long-term indicators of occupational outcomes in 2018.

Table A. *Hybrid Path Model for intergenerational reproduction of inequality in a gender perspective*

Equation on educational achievement (Y)

PREDICTORS		DIRECT EFFECT		INDIRECT EFFECT		INDIRECT EFFECT	
		ISCED 2011 (Y)		**daily wage 2011 (Z)**		**daily wage 2018 (W)**	
		beta	**r**	**beta**	**r**	**beta**	**r**
AREA NORTH	F	.	.				
(fer. Centre)	M	.	.				
AREA SOUTH	F	-0.26	-0.04				
	M	-0.50	-0.07				
Cohort 2	F	.	.				
(age at 2011 40-44)	M	0.34	0.05				
Cohort 3	F	0.66	0.09				
(age at 2011 35-39)	M	0.55	0.07				
Cohort 4	F	0.88	0.11				
(age at 2011 30-34)	M	0.55	0.07				
Single Parents	F	.	.	.	.	.	.
	M	-0.46	-0.03	-1.38	-0.01	-1.24	0.00
Loweduc father and mediumeduc mother	F	1.52	0.10	4.73	0.03	4.92	0.03
	M	1.59	0.10	4.76	0.02	4.26	0.01
Loweduc father and higheduc mother	F	2.69	0.04	8.36	0.01	8.69	0.01
	M	3.61	0.05	10.80	0.01	9.67	0.01
Mediumeduc father and loweduc mother	F	1.46	0.13	4.53	0.03	4.71	0.03
	M	1.86	0.16	5.56	0.03	4.98	0.02
Both medium education	F	2.53	0.24	7.87	0.06	8.18	0.06
	M	2.73	0.24	8.16	0.04	7.31	0.03
Mediumeduc father and higheduc mother	F	3.35	0.09	10.42	0.02	10.83	0.02
	M	3.64	0.12	10.89	0.02	9.76	0.02
Higheduc father and loweduc mother	F	2.84	0.07	8.82	0.02	9.17	0.02
	M	3.76	0.10	11.24	0.02	10.07	0.01
Highedu father and mediumeduc mother	F	3.08	0.15	9.56	0.04	9.94	0.04
	M	3.84	0.18	11.75	0.03	10.29	0.03
Both high education	F	3.36	0.16	10.46	0.04	10.87	0.04
	M	3.93	0.17	11.75	0.03	10.53	0.02
Endsmeets with some difficulty	F	1.12	0.17	3.47	0.04	3.61	0.04
(ref. Endsmeets with difficulty)	M	0.75	0.11	2.25	0.02	2.01	0.02
Endsmeets fairly easy	F	1.67	0.27	5.20	0.06	5.41	0.06
	M	1.23	0.19	3.67	0.03	3.29	0.03
Endsmeets easily	F	1.83	0.23	5.68	0.06	5.91	0.06
	M	1.27	0.15	3.81	0.02	3.42	0.02

Only coefficients with p-value<0.05 are reported. "." p-value >0.05
Source: Authors' elaborations on AD-SILC data

Table A. *continue*

Occupational outcome at 2011 (Z) and at 2018 (W)

PREDICTORS		DIRECT EFFECT				INDIRECT EFFECT			
		daily wage 2011 (Z)		**daily wage 2018 (W)**		**daily wage 2011 (Z)**		**daily wage 2018 (W)**	
		beta	**r**	**beta**	**r**	**beta**	**r**	**beta**	**r**
y_day_2011	F			0.70	0.67				
	M			0.68	0.64				
isced_anni	F	3.11	0.24	1.07	0.08			2.16	0.16
	M	2.99	0.17	0.63	0.03			2.04	0.11
AREA NORTH	F	5.40	0.07	**.**	.	.	.	3.48	0.04
(fer. Centre)	M	10.24	0.09	4.27	0.04	.	.	7.26	0.06
AREA SOUTH	F	-4.81	-0.05	**.**	.	-0.79	-0.01	-4.17	-0.04
	M	-13.82	-0.11	-3.73	-0.03	-1.50	-0.01	-10.79	-0.08
COHORT 2	F	.	.	1.67	0.02	.	.	**.**	**.**
(age at 2011 40-44)	M	.	.	3.84	0.03	1.01	0.01	**.**	**.**
COHORT 3	F	.	.	.	.	2.06	0.02	**.**	**.**
(age at 2011 35-39)	M	.	.	4.61	0.03	1.63	0.01	**.**	**.**
COHORT 4	F	.	.	.	.	2.74	0.03	**.**	**.**
(age at 2011 30-34)	M	.	.	5.83	0.04	1.65	0.01	**.**	**.**
job experience at 2011	F	0.13	0.23					0.09	0.15
	M	0.19	0.21					0.13	0.13
ISCO 2	F	.	.	.	.			.	.
	M	-6.07	-0.05	.	.			-4.15	-0.03
ISCO 3	F	7.94	0.10	.	.			5.53	0.07
	M	5.19	0.04	.	.			3.55	0.02
ISCO 4	F	20.10	0.19	.	.			13.98	0.13
	M	14.58	0.10	5.39	0.03			9.97	0.06
ISCO 5	F	17.77	0.19	.	.			12.36	0.12
	M	20.92	0.14	.	.			14.30	0.09
Year of the 1st job	F	-0.40	-0.08	-0.16	-0.03			-0.27	-0.05
	M	-0.76	-0.10	.	.			-0.52	-0.06
Penalty for maternity leave	F	-0.88	-0.07	.	.			-0.61	-0.05
	M	.	.	0.003	0.02			**.**	0.01
NUMBER OF children (2011 or 2018)	F	-3.08	-0.08	.	.			-2.14	-0.05
	M	4.51	0.08	.	.			3.08	0.05
dur_cum2018	F			0.08	0.28				
	M			0.11	0.22				

Only coefficients with p-value<0.05 are reported. "." p-value >0.05
Source: Authors' elaborations on AD-SILC data

References

1. Acock, A. C. (2013). Discovering structural equation modeling using Stata, Texas, Stata Press Books.
2. Ballarino, G., Meraviglia, C., & Panichella, N. (2021). Both parents matter. Family-based educational inequality in Italy over the second half of the 20th century. *Research in Social Stratification and Mobility*, 73, 100597.
3. Bernardi, F., & Comolli, C. L. (2019). Parental separation and children's educational attainment: Heterogeneity and rare and common educational outcomes. *Journal of Family Research*, 31(1), 3-26.
4. Bernardi F. Plavgo I. (2019). Education as an equalizer for human development? UNDP Human Development Report BACKGROUND PAPER NO. 4-2019
5. Bernardi, G. Ballarino (2016) (a cura di), Education, Occupation and Social Origin. A Comparative Analysis of the Transmission of Socio-Economic Inequalities, Londra: Elgar, pp. 255-282.
6. Causa, O., & Johansson, Å. (2011). Intergenerational social mobility in OECD countries. *OECD Journal: Economic Studies*, 2010(1), 1-44.
7. Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York, Routledge.
8. Hornstra, M., & Maas, I. (2021). Does the impact of the family increase or decrease over the life course? Sibling similarities in occupational status across different career points. *Research in Social Stratification and Mobility*, 75, 100643.
9. Kline, R. B. (2015). Principles and practice of structural equation modeling. New York, Guilford publications.
10. Kogan, I., Noelke, C., & Gebel, M. (Eds.). (2011). *Making the transition: Education and labor market entry in Central and Eastern Europe*. Stanford University Press.
11. McLanahan, S., & Bumpass, L. (1988). Intergenerational consequences of family disruption. *American journal of Sociology*, *94*(1), 130-152.
12. Shavit, Y., & Blossfeld, H. P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Social Inequality Series*. Westview Press, 5500 Central Avenue, Boulder, CO 80301-2847.
13. Shavit, Y., & Muller, W. (1998). *From School to Work. A Comparative Study of Educational Qualifications and Occupational Destinations*. Oxford University Press, 2001 Evans Road, Cary, NC 27513.

Is there a cultural driver pushing Italian low fertility?

C'è un fattore culturale che spinge la bassa fecondità italiana?

Francesca Luppi, Alessandro Rosina and Maria Rita Testa[1]

Abstract This study analysis the possible role of cultural factors in explaining, together with job and economic uncertainties, the Italian low fertility. Using data from the 2020-2021 panel survey of the Toniolo Institute's Youth Report, through a series of random-effects logit models, we analysed the combined impact of employment uncertainty and attitudes towards work and family on individual's motivation to have at least one child or two children, in a representative sample of young Italians aged between 18 and 34 years. Employment uncertainty weighs in determining the motivation for parenting only for those who see their work an important dimension of self-fulfilment rather than a mean to achieve other ends in life.

Abstract *Questo studio analizza il possibile ruolo di fattori culturali nel concorrere, insieme alle incertezze lavorative ed economiche, a spiegare la bassa fecondità italiana. Usando i dati dell'indagine panel 2020-2021 del Rapporto Giovani dell'Istituto Toniolo, attraverso una serie di modelli logit ad effetti casuali, abbiamo analizzato l'impatto combinato dell'incertezza occupazionale e degli atteggiamenti nei confronti del lavoro e della famiglia sulla motivazione ad avere almeno un figlio o due figli, in un campione rappresentativo di giovani italiani di età compresa fra i 18 e i 34 anni. L'incertezza occupazionale pesa nel determinare la motivazione alla genitorialità solo per coloro che vedono nel lavoro soprattutto una dimensione*

[1] Francesca Luppi, Università Cattolica del Sacro Cuore; email: francesca.luppi1@unicatt.it
Alessandro Rosina, Università Cattolica del Sacro Cuore; email: alessandro.rosina@unicatt.it
Maria Rita Testa, LUISS; email: mtesta@luiss.it

importante di autorealizzazione piuttosto che un mezzo per realizzare altri fini nella vita.

Key words: fertility motivation, low fertility, culture, uncertainty, Italy, young people

1 Introduction

Since the second half of the 1970s, the fertility rate in Italy is below the replacement level - i.e., lower than 2.1 children per woman. During the 1990s, the country reached the lowest-low fertility ever seen in Europe: in the 1995 the total fertility rate was 1.18 children per woman. After a short period of fertility recuperation during the first decade of the 2000s, the Italian fertility rate started to decline again as a consequence of the 2008 Great Recession. During the last 5 years, Italy, together with Spain, has the sad record of showing the lowest total fertility rate among European countries. The Covid-19 crisis has even more deepened the fertility gap of these two countries. This low fertility trend has been largely explained as due to postponement of the transition to motherhood (i.e., increasing mother's age at first birth) and the gradual affirmation of the single-child family model. In the Italian cultural, welfare and social system, where the family represents the foundation of the society and the individual's well-being, having at least one child in life has represented for a long time a "moral obligation". The fact that having children is highly valued in the Italian society is mirrored by the statistics showing how, on average, Italians declare to desire a two-children family in their life (Régnier-Loilier et al. 2011, Mencarini & Vignoli 2018). The difficulties to reach the desired fertility has been attributed to the increased economic uncertainty and labour market vulnerability of the young generations, as a results of the economic globalization first, the Great Recession and, finally, the Covid-19 crisis (e.g., Vignoli and Comolli 2021).

However, while the increase of the incidence of childless people has been often treated and studied as an involuntary outcome due to external/contingent obstacles to the realization of fertility desires, in recent times the Italian low fertility has been fostered by an increasing relative number of childless women, which declare not to desire to have children in their life (ISTAT 2019, Sobotka 2017). In the literature, they have been labelled as "childfree", to highlight their absence of desire to become mothers, against the dominant value of motherhood as women's predestination (Tanturri & Mencarini 2008). This evidence suggests that some cultural drivers might be increasingly playing a role to explain low fertility behaviours in the Italian context (Luppi 2022).

Therefore, we argue that, although the reasons for the Italian low fertility were often searched among structural (e.g., dysfunctional labour market, inadequate family policies, etc.) and conjunctural (e.g., economic recessions) factors, the spread

of zero- or single-child family models might become a new normality, which is socially accepted and even desirable. Consistently, because answers about abstract ideal family size are likely to reflect societal norms instead of individual's motivation (Goldstein, Lutz, & Testa 2003; Sobotka & Beaujouan 2014), we think that the fertility desires is not a fully reliable indicator when aiming to explore cultural changes pushing individual's fertility at (in this case) lower levels.

The aim of our study is to explore whether and to what extent low fertility behaviours in Italy derived not only by the contingent situation of uncertainty (not just economic), which increases the perceived risk associated with the decision to have a child, but also by a cultural driver, which might explain the limited relevance of parenthood in designing young-adult Italians' identities. If the uncertainty-explanation holds, it would mean that young people with more vulnerable positions in the labour market (i.e., those without a tertiary degree, unemployed individuals and those holding unstable occupations) are at higher risk of losing motivation to have a(nother) child. However, we claim that losing motivation in childbearing is linked to a higher acceptance of having other important sources of self-realization in life (such as work), while family is just one of them. When economic uncertainty increases, those who feel to have other important dimensions of self-realization in life besides becoming a parent (or having one more child) might be at higher risk of abandoning their childbearing motivation, especially if they see these goals as incompatible or increasing the relative opportunity-costs of having children.

2 Data and sample

To explore the role of cultural factors which might lead to low fertility behaviours (i.e., not having children or having only one child), we exploit data from the Rapporto Giovani panel survey of Istituto Toniolo, conducted on a quota sample of Italian young people (aged 18 to 34) in November 20201 (first wave; 7000 cases) and December 2021 (second wave). These data are particularly useful to our purpose as, alongside the traditional questions about the number of desired and expected children, the survey includes one specific question aimed at investigating the intrinsic motivation to parenthood, i.e., the value that individuals give to becoming a parent as a necessary experience to feel fully realized in their life. Next to that, other questions explore the level of traditionalism in family values; the meaning in life given to the work sphere; the reasons for not intending to have a child in the short term; the expected positive and negative impacts of having a child soon; and how increased risks in several life spheres would lead people to temporarily renounce to have children. Some of these questions are taken from other national and international surveys (e.g., Italian Multipurpose Survey; European Values Study),

[1] For another study on the same data on fertility intentions and motivations see Bonanomi, A., Luppi, F., & Rosina, A. (2021). Il futuro tenuto a distanza: progetti di vita in sospeso. In (VV.AA.) La condizione giovanile in Italia - Rapporto Giovani 2020. Il Mulino, Bologna.

thus allowing interesting cross-national comparisons; other questions are instead unique of this survey, thus importantly contributing to the understanding of low-fertility in Italy.

After dropping cases for attrition in the second wave and missing values in the variables considered for the analyses, the final sample is made by 1594 individuals (3188 observation), distributed across age classes as follow: 332 in the 18-24 group (310 weighted cases); 500 in the 25-29 group (542 weighted cases); 762 in the 30-34 group (741 weighted cases). The analyses focus on:

- childless individuals and those with only one child, excluding who has an ongoing pregnancy (1506 cases), as they are those at risk of adopting low-fertility behaviours;
- oldest individuals (i.e., 30-34 years old), even though also the youngest individuals will be considered. This choice is led by some considerations. First, this is the age group in which - on average - women have children in Italy. Secondly, after the age of 30 most young people have already acquired a first position, albeit precarious, in the labour market, while career prospects are better outlined with respect to the younger cohorts. Finally, even though a family growth is still possible, women in this age class have a limited time frame available to plan a birth.

Our dependent variable (operationalized as in the "Models" section) is derived by the question "With which of the following statements do you recognize yourself the most?" with four possible alternative answers: [1] I think that I would have a fulfilled life even without children; [2] I think I would have a fulfilled life only with one child; [3] I think I would have a fulfilled life only with two children; [4] I think I would have a fulfilled life only with a large family (at least three children).

To exploit the panel structure of the data, we first need to explore whether childbearing motivation changes for the same individual over the two waves of the survey. For the 30-34 age group, considering women and men together, table 1 shows the distribution of the answers to the previous question and reports the proportion of individuals who, in the 2021, confirmed (in the light-grey cells) or moved (downwards in the dark-grey cells, upwards in the white cells) from the answer given in the 2020. Most individuals confirm the same answer reported one year before, although the responses' consistency decreases with the increase of the family size: consistency goes from 77% in the case of 0-child to 45% in the case of 3-children or more. If they change opinion, they mainly revise downwards their motivation to parenthood.

Table 1: Variation in the motivation to have children in life among young people aged 30-34, between 2020 and 2021 (Sample size: 686 cases)

		2021				
		0 children	***1 child***	***2 children***	***3+ children***	***Total***
2020	***0 children***	77.27	12.74	6.24	3.76	100
	1 child	25.54	56.19	14.91	3.36	100
	2 children	12.9	33.81	41.50	11.78	100
	3+ children	6.72	19.95	28.50	44.83	100

3 Models and variables

Random effects logit models will be used in the empirical analysis (Hausman tested). Female and male samples are pooled in the main analyses, but gender differences are also tested.

In a first set of models the dependent variable is dichotomous, contrasting the option "I think that I would have a fulfilled life even without children" versus all remaining options (see the range in the above section). In a second set of models, we will consider a dichotomous variable contrasting the option "I think I would have a fulfilled life only with one child" versus the other two options including families with more children. The choice to have two different sets of models lies in the distinctive features characterizing the childless life trajectories from all other childbearing life trajectories.

Explanatory variables will include sociodemographic and economic characteristics of individuals (like education, employment, partnership status, financial situation, etc.): variables related to socio-economic status are especially relevant as low education and vulnerability in the labour market are associated with higher economic risks and uncertainty. Because of the small number of individuals declaring to have reached only a primary education, education has been included as a dummy, contrasting tertiary degree with lower education attainments. Occupational status has been included in four categories: unemployed, inactive, holding a stable occupation (i.e., professionals, managers, self-employed workers and permanent employees), and holding an unstable occupation (i.e., fixed-term employees, temporary workers, seasonal workers, project workers, etc.). This classification mirrors different kinds of vulnerability and uncertainty in the labour market.

The culture-related covariates are mainly represented by attitudes towards work and family. Attitudes towards work are asked through the following question: "Which of the following statements best reflects your idea of work (answer in general terms, not according to any work you are doing)? For me, work is above all ..." with the following alternatives (single answer allowed):

1. A mean to providing income
2. A space of personal commitment
3. A source of fatigue and stress
4. A dimension of self-realization
5. A way to face the future
6. A tool for building a family
7. A source of success
8. A source of social prestige

The variable has been dichotomized, taking value 1 in case work is mainly seen as a mean for achieving other things il life (i.e., answers n. 1, 3, 5 and 6), and value 0 in case work is mainly seen as a dimension of self-realized (i.e., answers n. 2, 4, 7 and 8).

Attitudes towards family are measured through the question: “How much do you agree with these statements?” on a scale ranging from 1 (totally disagree) to 4 (totally agree), referring to the following items:

1. Marriage is an outdated institution
2. A couple can live together even without planning to get married
3. A woman can have a child alone even if she doesn't want to have a stable relationship with a man
4. When children are around 18-20 years old, they should leave the parental home
5. It is right that a couple with an unhappy marriage divorces even if they have children
6. If the parents separate / divorce it is better that the children stay with the mother rather than with the father
7. When parents need care, it is natural that daughters more than sons take care of them
8. Being a housewife allows a woman to feel self-fulfilled as holding a paid job

After providing the same polarity to each item (higher values on traditional attitudes: item n. 4, 6, 7, and 8; lower values on progressive attitudes: item n.1, 2, 3, and 5), we performed a factor analysis to get confirmation of the existence of only one factor behind this scale. Then, an additive index has been calculated by summing the scores on each item. The final variable is a dichotomization of this index, taking value 1 in case the individual’s score falls above the median of the index distribution (traditional attitude) and value 0 in the other case (progressive attitude).

Besides the above-mentioned covariates, an additional set of variables considering expected short-term impacts (or risks) of childbearing on other life domains (e.g., free time, couple relationship, life satisfaction, certainties in life, etc.) will be included in the analyses to encompass other possible conflicting priorities.

4 Results

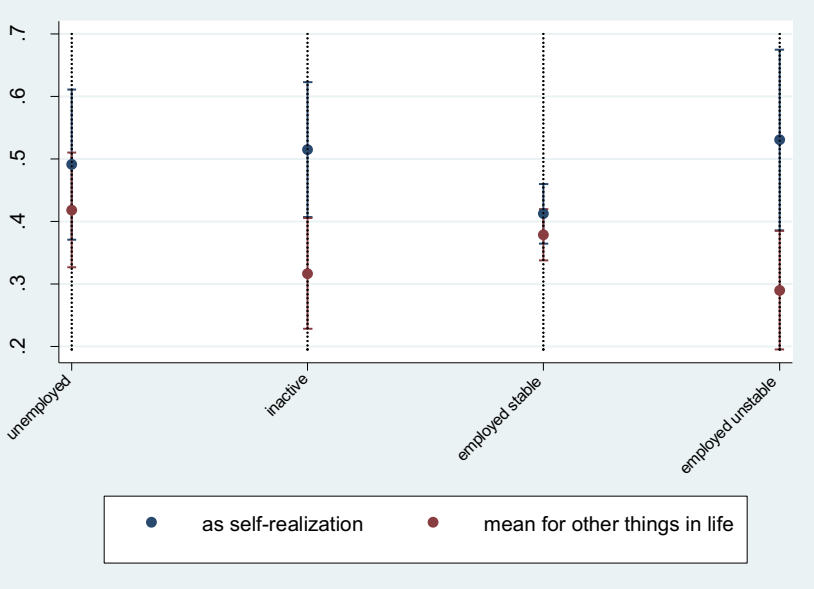
Preliminary results from random effects logit models show that the probability to reduce - over the survey period - the motivation towards having children (and in particular for those declaring they would feel fully realized in their life without children) is associated with holding a tertiary degree and not having a partner, while the occupational status is not significantly related with changes in fertility motivation (see table 2).

Table 2: Results from random effects logit models on the motivation of not having children in life for childless people (Reference categories: men, holding secondary or primary education, not partnered, unemployed. Respondents aged 25-34. Sample size: 978)

	Coef.	S. E.	p-value	Coef.	S. E.	p-value	Coef.	S.E.	p-value
woman	0.749	0.289	0.010	0.694	0.290	0.017	0.472	0.275	0.086
tertiary education	0.662	0.190	0.000	0.673	0.191	0.000	0.684	0.207	0.001
partnered	-0.506	0.240	0.035	-0.504	0.240	0.036	-0.449	0.245	0.067
Occupational status:									
inactive	-0.249	0.403	0.536	-0.281	0.404	0.486	-0.330	0.413	0.424
employed stable	-0.384	0.341	0.260	-0.421	0.342	0.218	-0.324	0.345	0.348
employed unstable	0.320	0.461	0.487	0.291	0.462	0.528	0.495	0.473	0.296
Attitudes towards work: work is a mean for other things in life				-0.382	0.206	0.064			
Attitudes towards family: traditional family attitudes							-1.581	0.245	0.000
Constant	-0.776	0.378	0.040	-0.514	0.402	0.201	0.175	0.388	0.651

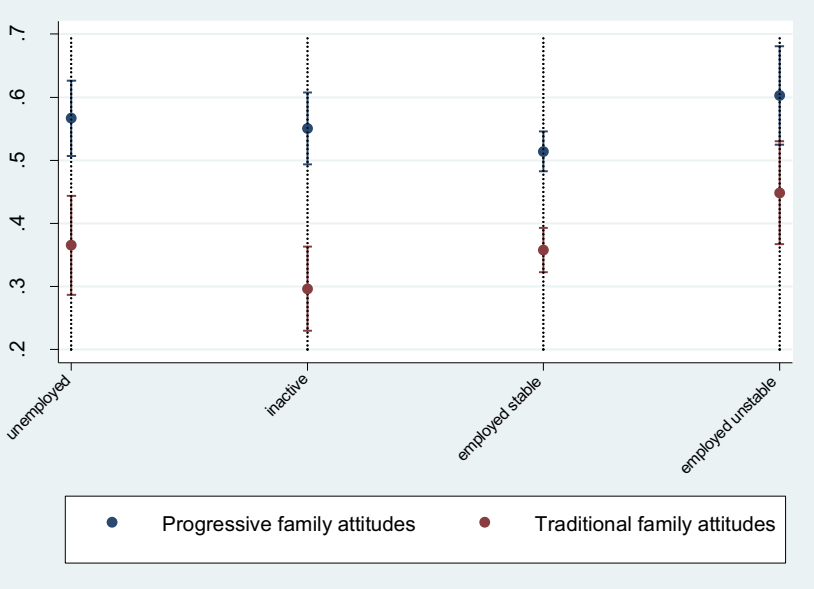
However, when we consider the mediation role of the cultural dimension things slightly change. In particular, for those considering their work as an important dimension of self-realization, becoming inactive or holding an unstable job contract increases the probability of reducing the relevance of having a child as a source of life meaning, compared to those considering their job as a mean for having other things in life and to those holding a stable job. This relationship appears especially for men (see figure 1).

Figure 1: Predicted probabilities of losing importance in having children in life for childless men perceiving their work as a way for self-realization or as mean for other things in life (from random effects logit model, c.i. at 83.55%. Reference categories: men, holding secondary or primary education, not partnered, unemployed. Respondents aged 25-34. Sample size: 413)



Meanwhile, those reporting more traditional attitudes towards family have lower chances to reduce their motivation about having a child if compared with those showing more progressive attitudes (see figure 2). However, this difference disappears among those holding an unstable job, who report similar chances of revising downward their childbearing motivation, independently by their attitudes towards family ideal.

Figure 2: Predicted probabilities of losing importance in having children in life for childless women and men (pooled) holding traditional or progressive attitudes towards family (from random effects logit model, c.i. at 83.55%. Reference categories: men, holding secondary or primary education, not partnered, unemployed. Respondents aged 25-34. Sample size: 978)



5 Conclusion

Uncertainty, and especially the economic one, has been recently appointed as one of the main causes of the delayed and reduced Italian fertility (Vignoli et al. 2021; Guetto et al. 2022), challenging the widespread ideal of the two-children family. We argue that motivation to childbearing is vulnerable to uncertainties, but this vulnerability depends also on how individuals value their self-realization as parent and against competing life spheres (work primarily, but also couple relationship, leisure time etc.). In Italy work-family difficult reconciliation is usually seen as the main obstacle to progression to higher fertility.

Our preliminary analyses focus exactly on how motivation to childbearing reacts to job (and economic) vulnerabilities, depending on individuals' values regarding their self-realization in work and family. We found that experiencing labour market vulnerabilities is not always associated with lower childbearing motivation: this is observable among those (especially men) perceiving their work as an important sphere of self-realization. Those holding progressive attitudes towards family ideal are more prone to reduce their motivation towards having children compared to those with a more traditional family ideal, even though having an unstable job

eliminates this gap. Despite results are very preliminary, they suggest that labour market vulnerabilities do not equally weight on young Italians' motivation to childbearing, and a cultural driver is playing a role.

Further analyses will go in the direction of disentangling the role of the cultural drivers, first trying to better explore the previous findings through expanding the models by including possible omitted variables and running some heterogeneity analyses. For example, results on education suggest that holding tertiary education is not enhancing higher childbearing motivation, by potentially reducing occupational vulnerability and economic uncertainty; in this sense, it might be a proxy of a cultural driver relatively reducing the importance of parenthood for self-realization, but further investigation is needed. Additionally, we will consider other possible competing life spheres (such as leisure time and couple relationship) that might be associated with lower (declining) childbearing motivation.

Our data are timely and original at the same time. They provide detailed information on young adults' desire, intentions, and motivation to childbearing. Additionally, they include unique information also on attitudes and values related with family, work, leisure time and couple relationship, even exploring the perceived risks childbearing brings about the possibility to fully enjoy them. In other words, we can detect the presence of possible competing life priorities as opposed to (or not fully compatible with) parenting. As already mentioned, some of these questions are unique of this survey, thus importantly contributing and deepening the understanding of the low-fertility mechanisms.

Moreover, having both women and men in the sample will allow us to better understand the potentially gendered reproductive decision-making process of young couples in the Italian context of persistent low fertility. In a context where traditional gender values are still widespread, the gender-gap is high in terms of labour market and income opportunities-costs of childbearing, the work-family reconciliation lays more on women's than men's shoulders, also the way in which fertility competes with other goals in life (and work first) is very gendered.

Finally, our data have been collected during the Covid-19 crisis, at the beginning of the second (November 2020) and the third wave (November 2021). Even though our aim is not to assess the impact of the crisis on fertility motivations, we cannot avoid considering that the contingency of the moment had a great impact on the perceived uncertainty in many life spheres (not just economic) and on the societal structure, potentially impacting also on preferences and priorities. This might have stressed even more the conflict between life goals and the relative costs associated with reaching them.

6 References

1. Comolli, C. L., & Vignoli, D.: Spreading uncertainty, shrinking birth rates: a natural experiment for Italy. Eur. Sociol. Rev. (2021), 37(4), 555-570.
2. Goldstein, J., Lutz W., and Testa M.R.: "The emergence of sub-replacement family size ideals in Europe." Popul. Res. Policy Rev. 22.5 (2003): 479-496.
3. ISTAT: Natalità e fecondità nella popolazione residente (2018) Available via https://www.istat.it/it/files/2019/11/Report_natalit%C3%A0_anno2018_def.pdf
4. Guetto, R., Bazzani, G., & Vignoli, D.: Narratives of the future and fertility decision-making in uncertain times. An application to the COVID-19 pandemic. VYPR (2022) 20, 1-38.
5. Luppi F.: Le ragioni della bassa fecondità italiana: fra cambiamento culturale, incertezza economica e rigidità istituzionali, Rivista di Politica Economica (2022), 2: 57-80.
6. Mencarini L., & Vignoli D.: Genitori cercasi: l'Italia nella trappola demografica, EGEA spa (2018)
7. Régnier-Loilier A., Vignoli D., & Dutreuilh C.: Fertility Intentions and Obstacles to their Realization in France and Italy, Pop. (2011), 66 (2), pp. 361-389.
8. Sobotka T.: Childlessness in Europe: Reconstructing Long-Term Trends Among Women Born in 1900–1972, in Kreyenfeld M., Konietzka D. (eds.) Childlessness in Europe: Contexts, Causes, and Consequences, Demog. Res. Monographs (2017)
9. Sobotka, T., & Beaujouan, E.: Two Is best? The persistence of a two-child family ideal in Europe. Popul. Dev. Rev. (2014), 40(3), 391-419.
10. Tanturri, M. L., & Mencarini, L.: Childless or childfree? Paths to voluntary childlessness in Italy. Popul. Dev. Rev. (2008) 34(1), 51-77.
11. Vignoli, D., Minello, A., Bazzani, G., Matera, C., & Rapallini, C.: Economic Uncertainty and Fertility Intentions: The Causal Effect of Narratives of the Future (No. 2021_05). Università degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni" G. Parenti" Available via https://local.disia.unifi.it/wp_disia/2021/wp_disia_2021_05.pdf

Unpaid family work and the subjective well-being of Italian women during lockdown

Il lavoro domestico e di cura e il benessere soggettivo delle Italiane durante il lockdown

Marina Zannella, Erica Aloé, Marcella Corsi and Alessandra de Rose

Abstract This article is based on data from a web survey conducted in Italy, from May to June 2020, aimed at exploring how the confinement measures taken against the spread of COVID-19 affected family life and time use for paid and unpaid work. In addition to information on time use before/during/after confinement, respondents were also asked to report changes in their feelings associated with different activities. Our data show that during lockdown, women spent significantly more time on unpaid family work, while men only slightly increased their contribution to domestic and care work. The lack of rebalancing is reflected in women's subjective well-being: they reported more stress and fatigue associated with unpaid work. Instead, most mothers reported a greater sense of purpose (i.e., feeling more useful to others) in relation to childcare.

Abstract *Questo articolo utilizza i dati di un'indagine on-line condotta tra maggio e giugno 2020 per studiare gli effetti delle misure di lockdown sui tempi di vita ed il benessere degli italiani. I dati mostrano che durante il confinamento le ore giornaliere dedicate al lavoro non retribuito sono aumentate significativamente per le donne che vivevano in coppia, soprattutto per le madri, mentre le italiane hanno riportato solo un leggero incremento del tempo dedicato dai loro partner al lavoro domestico e di cura. L'assenza di riequilibrio nella distribuzione del carico di lavoro familiare all'interno delle coppie ha avuto ricadute sul benessere soggettivo delle donne che hanno riportato livelli più elevati di stress e stanchezza in associazione al lavoro familiare; tuttavia, solamente in relazione alla cura dei figli,*

[1] Marina Zannella, Sapienza University of Rome; email: marina.zannella@uniroma1.it
Erica Aloè, Sapienza University of Rome; email: erica.aloe@uniroma1.it
Marcella Corsi, Sapienza University of Rome; email: marcella.corsi@uniroma1.it
Alessandra De Rose, Sapienza University of Rome; email: alessandra.derose@uniroma1.it

la maggioranza delle madri intervistate ha sperimentato anche un maggior senso di utilità.

Key words: Covid-19; Lockdown; Time Use; Gender; Couples; Children; Subjective Well-being

1 Introduction

At the end of February 2020, Italy reported the largest COVID-19 outbreak outside of China (Chen et al., 2020). Thus, in March 2020, Italy was the first European country to impose a nationwide lockdown followed, later, by social distancing measures. The lockdown lasted 69 days. Moreover, schools have been closed nationwide for in-person activities until the end of the school year (June), a relatively longer period compared to most OECD countries where schools began to re-open in April and May (OECD, 2020).

The pandemic generated several compounding crises harming the economy and the well-being of people in addition to health. It has soon been evident that the consequences of these crises were not gender-neutral but were disproportionally attributed to women. Women are serving on the frontlines against COVID-19, and the impact of the crisis on women is stark. Women face compounding burdens: they continue to do most of the unpaid care work in households, face higher risks of economic insecurity, and face increased risks of violence, exploitation, abuse, or harassment during times of crises and quarantine compared to men (OECD 2020). Women, in Europe, are also more likely than men to work in occupations – such as health, care, education and hospitality – that are more exposed to the risk of being infected by contagious diseases spread by respiratory or close-contact route (Lewandowsky et al., 2021). Moreover, women continue to bear the burden of family care and to do most of the unpaid family work increased by stay-at-home recommendations, quarantine, lockdown periods and school closures.

Thus, the global pandemic caused by COVID-19 and the consequent lockdown did not represent only a danger in economic terms, but also a threat to the process towards gender equality (Bahn et al., 2020, Kabeer et al., 2021). Under the confinement measures there was an unprecedented increase in the demand for household production and the associated input of unpaid labour, a gendered economic phenomenon. Several phenomena affected the use of time at household level, including: closure of schools, with pupils having to bring forward school programs at home; suspension of non-necessary activities, affecting formal and informal sectors; introduction of remote work where it was possible; introduction of various limitations to people mobility. In this context, the unavailability of paid services (such as laundries, restaurants, baby-sitters, care givers, etc.) as well as the impossibility to benefit from informal care (e.g., by grandparents) contributed to the creation of additional unpaid work within households. This “extra” work fell disproportionally onto women, exacerbating the already existing inequalities in the

gender division of unpaid work (Raile et al., 2020). In particular, the shift to remote-work and the unavailability of formal and informal care disproportionally affected women's paid and unpaid work (Craig and Churchill, 2021).

Andrew et al. (2020) show, by using survey data collected in the UK, that during the pandemic women bore the brunt of the increased time needed for household chores and childcare. Findings from the study highlight that mothers who stopped working in the labour market did far more domestic work than fathers in the equivalent situation. These results seem to suggest that asymmetries in the gender allocation of the extra-amount of domestic work created by the pandemic cannot be explained as a sole effect of gender differences in employment and earnings, but mostly depend on social norms regulating gender roles as well as expectations on motherhood. Similar pandemic time-use surveys provide supporting results (see for example Farre, et.al. 2020 for Spain; Ilkkaracan and Memiş 2021 for Turkey).

Regarding Italy, Del Boca and colleagues (2020) used survey data collected in April 2020 on women living in dual-earner heterosexual couples to show that most of the additional housework and childcare associated with COVID-19 felt on women, while childcare activities were more equally shared within the couple than housework activities. Mangiavacchi et al. (2021) confirm that Italian households experienced a greater involvement of fathers in childcare during the lockdown. Their study also highlighted that men whose partners continued to work at their usual workplace spent more time on housework than before. Additionally, analysis of satisfaction with work–life balance shows that working women with children aged 0–5 years are those who found balancing work and family more difficult during COVID-19. The work–life balance was especially difficult to achieve for those with partners who continued to work outside the home during the emergency. From the perspective of paid work, using data from the Italian Labour Force Survey for the years 2019 and 2020, Brini and colleagues (2021) found no evidence of retraditionalization of gender roles in paid work among couples in Italy with dependent children. On the contrary, the authors found that the pandemic reduced time spent in paid work (and earnings) more for fathers than for mothers.

Other international studies, reviewed by Seedat and Rondon (2021), have documented a greater rise in psychological distress in women than in men during the lockdown. The higher risk of depressive and anxiety symptoms among women may be partially explained by the disproportionate burden of work that fell onto them.

Based on this background, this paper explores how the lockdown measures adopted in contrast to the diffusion of COVID-19 affected Italian women's use of time for unpaid family work. The assessment is based on real-time survey that collected more than 1,000 observations of persons aged 18 years or older living in Italy. The questionnaire was administrated on-line to the respondents immediately after the confinement period and, in addition to information about the use of time for paid and unpaid work, it included a set of well-being questions. This paper aims at describing changes in the couple's division of unpaid care and domestic work as well as in the levels of stress and fatigue experienced by women in association with these activities during the lockdown. In particular, the paper concentrates on the weight of increased care burdens due to lockdown measures and highlights the

different impact that such measures had on women that lived with children below 18 years old compared to other women.

2 Survey

To create the survey, we used the instrument developed by Donehower (2020) as a base. We translated the original survey from English into Italian, and we adapted it to the purposes of our study adjusting some of the queries and adding new questions. The final survey -structured in multiple choice questions- consisted of nine sections: household composition, health status, paid work (own and partner's), unpaid care work (active and passive), unpaid domestic work, informal help to/from other households, division of unpaid care work and unpaid domestic work within the household, feelings, socio-demographic information. Questions were asked to the respondent in relation to three different moments: before the pandemic, during the first lockdown and in the moment when they responded to the survey (that is immediately after the end of the lockdown). Responses were collected from May 22 and June 12, 2020. The lockdown in Italy ended on May 19th, 2020 and was followed by a so called 'phase two', which still implied several restrictions, including school closures.
The online questionnaire was open to anyone who was at least 18 years old while completing the survey and resided in Italy. The survey collected 1,008 observations (reduced to 979 when the dataset was cleaned from missing and invalid responses). In our analysis, we focus on women representing the great majority of the respondents (81%). Data collection was conducted anonymously and participation in the survey was voluntary. The survey was promoted through the institutional website of Sapienza and the main social networks, as well as through the mailing lists of scientific associations and professional contacts. The dataset is mainly composed by women with a high level of education (three fourth of them have a level of education higher than college). This is mirrored by a high reported employment rate among them. Therefore, data were post-stratified to ensure consistency with the main socio-demographic characteristics of the Italian population (i.e., age, education, geographical area of residence).

3 Results

The survey gathered detailed information about the time devoted by each respondent to unpaid work in the household. In this context, it becomes relevant to observe the differences between women that live with a minor children and other women. In fact, the answers that we collected highlight that, during the lockdown, but also after it, women faced an increase in the amount of time that they devoted to unpaid domestic work. The magnitude of this increase was higher for mothers of

minor children aged less than 18 years compared to women with no children or with adult children (Figure 1). According to our data, time for domestic chores increased from 2.4 hours per day to almost 3.7 hours per day for women with no dependent children, while it increased from 3.1 hours per day to almost 5 hours for women with minor children. For what concerns time devoted to childcare, Figure 2 shows that for women with small children (under 5 years old) during the lockdown childcare time became similar to a full-time job, more than 7 hours per day. Women with older children devoted to childcare less time than women with smaller children -around 6 hours per day with children between 6 and 10 years old and around 4 hours per day with children between 11 and 17 years old. It is relevant to notice that after the end of the lockdown the time devoted to childcare decreased only slightly and this was caused by the fact that all schools in Italy remained closed until September.

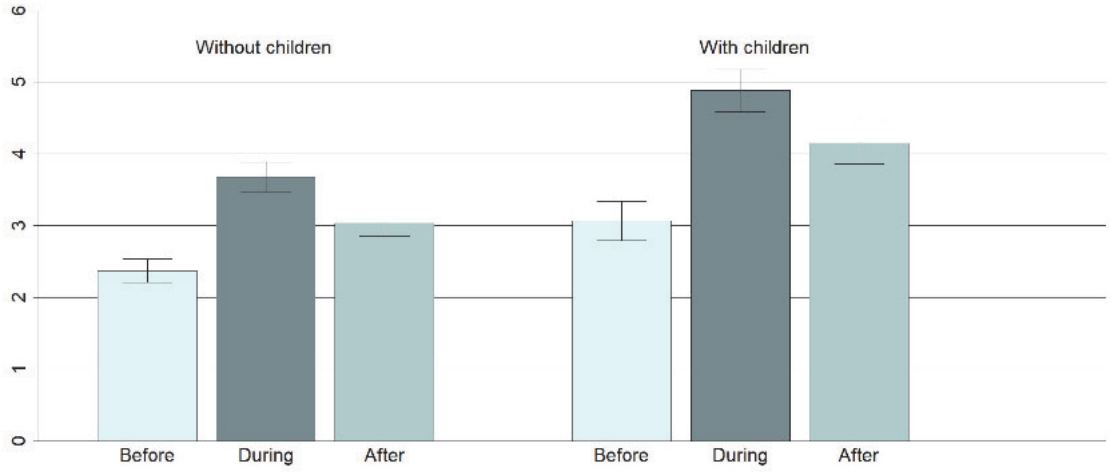


Figure 1: Women's *average daily hours of unpaid domestic work.*

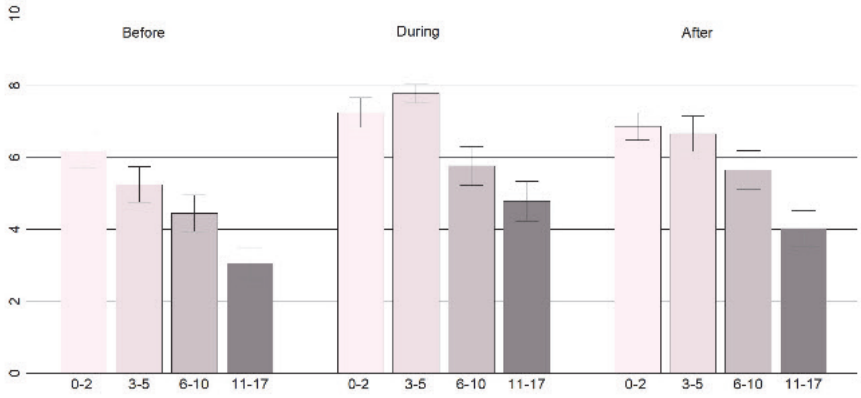


Figure 2: *Mothers' average daily hours of childcare according to the age of the youngest child.*

The survey asked each respondent to report the approximate share of the total household's unpaid care and domestic work performed by the partner if present. The results revealed that, before the pandemic, the male partner's share of unpaid care and domestic work was around 26% for men with no dependent children and 28% for fathers (see Figure 3). During the lockdown this share increased to almost 28% for men without young children, and 31% for men with children under 18 years old. However, soon after the end of the lockdown the male partner's share of unpaid care and domestic work lowered compared to its pre-pandemic level (25 and 26, respectively).

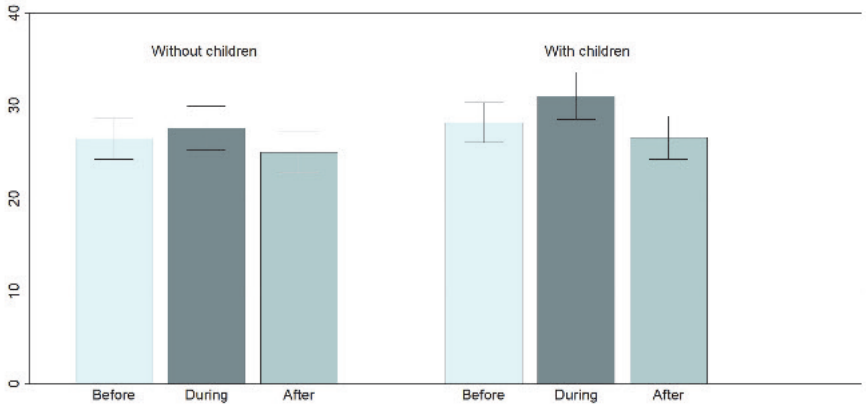


Figure 3: *Partner's average share of unpaid care and domestic work.*

Around 41% of women reported to feel more tired about domestic work during the lockdown, while 35.5% reported more stress.; the corresponding percentages increases to 50.4% and 39.3% among mothers of minor children. A closer look on mothers in Figure 4 reveals that about 46% of women with young children reported to be more tired and stressed doing childcare; however, about 59% reported more sense of purpose (i.e., feeling more useful to others) associated to childcare giving, 44.4% reported to feel more contented and

40.9% felt happier. Among women who reported more stress associated to childcare, responses were concentrated on the response modality indicating more intense changes ("much more"), while the opposite is true for positive feelings (meaningfulness and happiness) for which women reported to have experiences more moderate changes. The results of the changes towards more positive feelings associated with childcare seem to suggest that, despite the fatigue and stress due to the additional unpaid work, in the first phase of the health emergency most mothers positively valued the increased time available to spend with their children. The situation may have changed in the later stages of the health emergency due to the prolonged closure of schools.

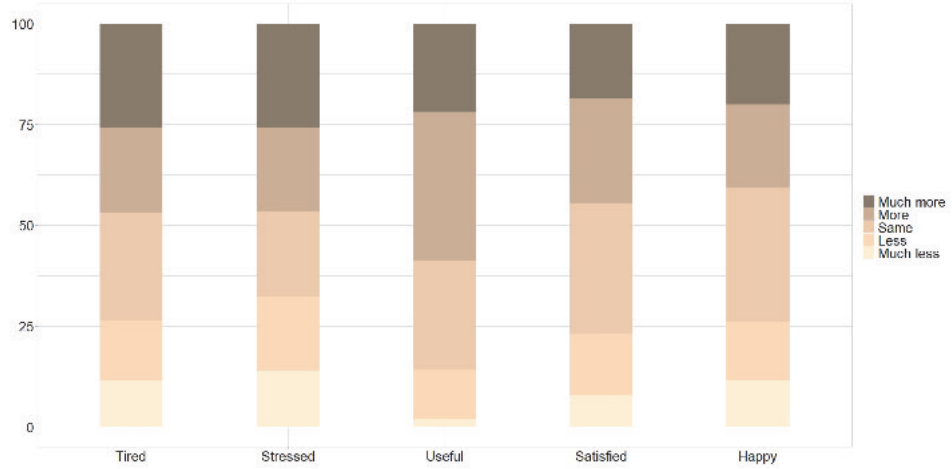


Figure 4: *Did you feel more or less ... than usual while spending time on childcare during the lockdown?*

4 Concluding remarks

Our data show that in Italy women became time poorer during the first phase of the pandemic: women were required to provide more unpaid care and domestic work (in particular, those with young children). Women reported that their partners only slightly increased their share of unpaid care and domestic work during the lockdown and that they returned to their pre-lockdown share soon after. This change in the use of time during the pandemic does not seem to suggest that a real and stable change in the division of unpaid work has been triggered, so to achieve a rebalancing of roles, parental and non-parental. The lack of rebalancing shows its effects, in our investigation, also on the subjective well-being experienced during phase 1 of the emergency. Women, especially those with minor children, reported to feel more stress and tiredness in association to paid and unpaid work activities while, only in relation to childcare, most women highlighted to have experienced a greater sense of purpose. To conclude, our results suggest that lockdown and social distancing measures introduced to contrast the pandemic have exacerbated the pre-existing

gender inequalities in the quantity and in the nature of unpaid family work (Zannella and De Rose 2020; 2021).

References

Andrew, A., Cattan, S., Costa Dias, M., Farquharson, C., Kraftman, L. Krutikova, S., Phimister, A., Sevilla. A. The gendered division of paid and domestic work under lockdown. IZA Discussion Paper 13500. Bonn, Germany: IZA Institute of Labor Economics (2020)

Bahn, K., Cohen, J., van der Meulen Rodgers, Y. A feminist perspective on COVID-19 and the value of care work globally. Gender Work Organization (2020) doi: https://doi.org/10.1111/gwao.12459

Brini, E., Lenko, M., Scherer, S., & Vitali, A. Retraditionalisation? Work patterns of families with children during the pandemic in Italy. *Demographic Research* (2021) doi: 10.4054/DemRes.2021.45.31

Chen, J., Lu, H., Melino, G., Boccia, S., Piacentini, M., Ricciardi, W., Wang, Y., Shi, Y., & Zhu, T. COVID-19 Infection: The China and Italy Perspectives. Cell Death & Disease (2020) doi:10.1038/s41419-020-2603-0

Connelly, R., Kimmel, J. If you're happy and you know it: How do mothers and fathers in the US really feel about caring for their children? Feminist Economics (2015) doi :https://doi.org/10.1080/13545701.2014.970210

Craig, L. Does father care mean fathers share? A comparison of how mothers and fathers in intact families spend time with children. Gender & Society (2006) https://doi.org/10.1177/0891243205285212

Craig, L., Churchill, B. Working Caring at Home: Gender Differences in the Effects of Covid-19 on Paid and Unpaid Labor in Australia. Feminist Economics. (2021) doi: https://doi.org/10.1080/13545701.2020.1831039

Craig, L., Powell, A. Non-standard work schedules, work-family balance and the gendered division of childcare. Work, Employment and Society (2011) doi: https://doi.org/10.1177/0950017011398894

Del Boca, D., Oggero, N., Profeta, P., & Rossi, M. Women's and men's work, housework and childcare, before and during COVID-19. Review of Economics of the Household, (2020) doi: https://doi.org/10.1007/s11150-020-09502-1

Donehower, G. Counting Women's Work: Unpaid care work and Covid19 (2020) https://www.countingwomenswork.org/news/unpaid-care-work-and-covid19-take-the-survey

Farre, L., Y. Fawaz, L. Gonzalez and Graves, L. How the Covid-19 Lockdown affected gender inequality in paid and unpaid work in Spain? IZA Discussion Paper No. 13434 (2020)

Kabeer, N., Razavi, S., van der Meulen Rodgers, Y. Feminist Economic Perspectives on the COVID-19 Pandemic. Feminist Economics (2021) doi: https://doi.org/10.1080/13545701.2021.1876906

Ilkkaracan, I. & Memiş, E. Transformations in the Gender Gaps in Paid and Unpaid Work During the COVID-19 Pandemic: Findings from Turkey. Feminist Economics (2021) doi: 10.1080/13545701.2020.1849764

Lewandowsky, P., Lipowska, K., Magda, I. The Gender Dimension of Occupational Exposure to Contagion in Europe. Feminist Economics (2021) doi: https://doi.org/10.1080/13545701.2021.1880016

Mangiavacchi, L., Piccoli, L., Pieroni, L. Fathers matter: Intrahousehold responsibilities and children's wellbeing during the COVID-19 lockdown in Italy. Economics & Human Biology (2021) doi: 10.1016/j.ehb.2021.101016

Musick, K., Meier, A., Flood, S. How parents fare: Mothers' and fathers' subjective wellbeing in time with children. American Sociological Review (2016) doi: https://doi.org/10.1177/0003122416663917

OECD. OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis. OECD Publishing: Paris (2020).

Raile, A.N.W., Raile, E.D., Parker, D.C.W., Shanahan, E.A., Haines, P. Women and the weight of a pandemic: A survey of four Western US states early in the Coronavirus outbreak. Gender Work Organization (2020) doi; https://doi.org/10.1111/gwao.12590

Seedat, S., Rondon, M. Women's Wellbeing and the Burden of Unpaid Work. BMJ (2021) doi: https://doi.org/10.1136/bmj.n1972

Zannella, M., & De Rose, A. Gender differences in the subjective perception of parenting time. RIEDS - Rivista Italiana di Economia, Demografia e Statistica - Italian Review of Economics, Demography and Statistics (2020) http://www.sieds.it/wp-content/uploads/2020/12/Volume-LXXIV-N.-2-Aprile-Giugno-2020.pdf

Zannella, M., & De Rose, A. Fathers' and mothers' enjoyment of childcare: the role of multitasking. Vienna Yearbook of Population Research (2021) doi: http://dx.doi.org/10.1553/populationyearbook2021.res3.1

New Frontiers in the theory of composite indicators

Methodological PLS-PM Framework for Model Based Composite Indicators

PLS-PM per indicatori compositi basati su modelli

Cataldo Rosanna

Abstract Today, Composite indicators (CIs) have been widely accepted as a tool for assessing and ranking countries and institutions in terms of environmental performance, sustainability, and other complex concepts that are not directly measurable. The proliferation of the production of composite indicators by all the major international organizations is a clear symptom of this political importance and operational relevance in the decision-making process. Consequently, the way these indicators are constructed and used appears to be a very important research question from both a theoretical and operational point of view. The work focuses on building a system of composite indicators through to Structural Equation Modeling, specifically with the use of Partial Least Squares-Path Modeling. The aim is to show the role key that the Partial Least Squares Path Modeling has in the estimation process of the composite indicators.

Abstract *Oggi, gli Indicatori Compositi sono stati ampiamente accettati come strumenti per valutare e classificare paesi e istituzioni in termini di prestazioni ambientali, sostenibilità e altri concetti complessi che non sono direttamente misurabili. Il proliferare della produzione di indicatori compositi da parte di tutte le maggiori organizzazioni internazionali è un chiaro sintomo di questa importanza politica e rilevanza operativa nel processo decisionale. Di conseguenza, il modo in cui questi indicatori sono costruiti e utilizzati sembra essere una questione di ricerca molto importante sia dal punto di vista teorico che operativo. Il lavoro si concentra sulla costruzione di un sistema di indicatori compositi attraverso i modelli ad equazione strutturale, in modo particolare attraverso i modelli Partial Least Squares Path Modeling. L'obiettivo è mostrare il ruolo chiave che la tali modelli hanno nel processo di stima degli indicatori compositi.*

Key words: Composite Indicators, Partial Least Squares Path Modeling

---

Cataldo Rosanna
University of Naples Federico II, e-mail: rosanna.cataldo2@unina.it

1 Motivations of method

Composite indicators (also referred to as Synthetic Indices) are popular tools for assessing the performance of nations on many social and economic complex phenomena that don't seem to be directly measurable and not uniquely defined, like human development, sustainability, innovation and competitiveness. In line with Saisana and Tarantola [32], a Composite Indicator (CI) is defined as *a mathematical combination of single indicators that represent different dimensions of a concept whose description is the objective of the analysis*. CIs are very useful so as to handle those phenomena that may not be observed directly. As is understood, building a composite indicator may be a delicate task and stuffed with pitfalls: from the issues regarding the availability of informations and also the choice of individual indicators, to their treatment in order to match (normalization) and aggregate them (weighting and aggregation). No universal method exists for composite indicators construction. Generally, three different approaches are available for its construction [26]. The primary approach is *Theory Based approach* in keeping with which CIs are computed by simple formulas that combine some observable variables. This approach requires strong knowledge or assumptions about the phenomena under study considering usually a well-defined set of variables. In contrast, a *Data Driven approach* overcomes the shortage of information, putting into the method of building a CI many observed variables, proxies of the concept to be measured. These two approaches have some limitations with respect to the quantity of EIs used, to the selection of the system of weightings used to aggregate the EIs and to the absence of any relationship between the EIs and the CIs. Mid-way between Theory Based and Data Driven CIs, the *Model Based approach* allows you to require into consideration some a priori information about the context of the phenomena by considering the relationship between the target or output CI and other representing inputs and outcomes of the system under study in terms of a path diagram. In order to compute a Model Based CI, taking into consideration all a prior information, a relevant role is played by the Structural Equation Modeling (SEM) methodology, particularly Partial Least Squares Path Modeling (PLS-PM) that's a statistical approach for modeling complex multivariable relationships among observed and latent variables. According to this methodology, it is possible to define a CI as a multidimensional Latent Variable (LV) not measurable directly and related to its single indicators or Manifest Variables (MVs) by either a reflective or formative relationship or by both (this defines the measurement or outer model). Each CI is related to other CIs, in a systemic vision, by linear regression equations specifying the so called Structural Model (or Inner Model). As a result a Systemic CI or a System of CIs is obtained, where the word *systemic* derives from the definition of system given by von Bertalanffy [41], in keeping with which *a system is a set of elements in interaction*, not just an aggregation of EIs but a set of indicators related to each other by mutual relationships, expressed through functional links and, summarized in a specific model. The basic idea is that the complexity inside a system can be studied by taking into account the whole set of causal relationships among latent concepts (LVs), each measured by several observed indicators usually defined as MVs.

PLS-PM represents a very important breakthrough with respect to traditional aggregation methods, like a PCA or a simple arithmetic mean of the original indicators. Instead of taking the unweighted sum of the indicators (or unit-weights for all the indicators), PLS-PM assigns weights to the initial variables taking under consideration the network of relationships between the constructs and the variance and covariance structure within and between the blocks of variables. Moreover, PLS-PM provides components with specific proprieties in order to enhance interpretation of the composites and the relationships among them. Specifically, depending on the chosen estimation options, PLS-PM provides components that are as much correlated as possible to each other while explaining the variances of their own set of variables. The choice of using the PLS-PM is particularly useful for several reasons [26]. This approach has as its main advantages its applicability to small sample, the ability to estimate quite complex models (with many latent and observable variables) and less restrictive requirements concerning normality and variable and error distributions [21]. Furthermore, PLS-PM approach provides the possibility of working with missing data and in the presence of multi-collinearity. Another advantage of this approach, as compared to other multivariate techniques, is that it examines simultaneous a series of dependence relationship, using a single statistical approach to test the full scope of projected relations [18]. Furthermore, this approach provides researchers with much more flexibility as it enables using both formative and reflective measurement models, providing a more nuanced testing of theoretical concepts [17]. Finally, according to Tenenhaus et al. [37] the PLS-PM approach should be used in order to not only reduce the number of dimensions but find relations between Composite Indicators and their blocks.

2 Method and its main variants

The PLS-PM is a multivariate statistical technique first introduced by Wold in the late 1960s [37]. PLS-PM is made up of two elements, the measurement model (also called the outer model), which describes the relationships between each construct (LVs) and its associated observed variables (also often called indicators, items or MVs), and the structural model (also called the inner model), which describes the casual-predictive relationships between the constructs (Fig. 1). A LV is called endogenous if it is supposed to depend on other LVs and exogenous one otherwise. The structural model can be written as:

$$\xi_j = \sum_{(q:\xi_q \to \xi_j)} \beta_{qj}\xi_q + \zeta_j \tag{1}$$

where ξ_j is an endogeneous LV, β_{qj} is the path coefficient linking the exogenous $q-th$ LV to the $j-th$ endogenous one expressing the impact on the endogenous LV ξ_j of the connected exogenous LVs, and ζ_j is the error in the inner relationship.

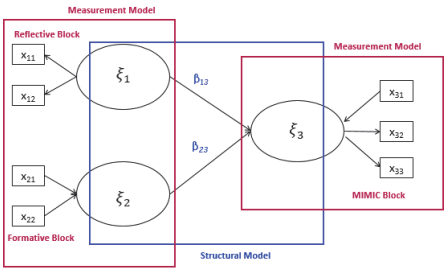


Fig. 1 PLS-PM: structural model and measurement model

The measurement model formulation depends on the direction of the relationships between the LVs and the corresponding MVs [12]; [40]. The kind of measurement depends on the construct conceptualization, the aim of the research and the role of the construct in the model [33]. Hair et al. [13] in their book provide guidelines for choosing the appropriate measurement specification. There are three types of measurement model that relate the MVs to their LVs:

- Reflective model (or outwards directed model). Each manifest variable reflects the corresponding latent variable. In this case, it assumes that the block of manifest variables related to a latent variable measures a unique underlying concept and the indicators linked to the same latent variable should covary: changes in one indicator imply changes in the others. In this case the relation between each manifest variable (MV) and the corresponding LV is made explicit through the following equation:

$$x_{pq} = \lambda_{pq}\xi_{pq} + \varepsilon_q \tag{2}$$

  where ξ_{pq} is the exogenous LV, and λ_{pq} is the simple regression coefficient between the MV and the LV, the so called *loading*.

- Formative model (or inwards directed model). The LV is supposed to be generated by its own MVs, i.e each manifest variable or every set of manifest variables represents a different level of the underlying latent concept. Thus the measurement model could be expressed as:

$$\xi_q = \sum_{p=1}^{P_q} \omega_{pq}x_{pq} + \delta_q \tag{3}$$

  where ω_{pq} is the coefficient linking each MV to the corresponding LV and δ_q is the error that represents the part of the LV not explained by the block of MVs.

- MIMIC model (a mixture of the two previous models), which is a combination of the reflective and formative ways.

The PLS-PM approach consists of an iterative algorithm that computes the estimation of the LVs, measured by a set of MVs, and the relationships between them,

by means of an interdependent system of equations based on multiple and simple regression. The idea is to determine the scores of LVs through a process, that, iteratively, computes first an outer and then an inner estimation [37]. The extraction of CI scores represents a key characteristic of the PLS-PM method. In the system of LV built with PLS-PM, you can obtain the scores for each LV, exogenous or endogenous, and for each indicator you can make a ranking among units. Moreover, PLS-PM provides information on the relative importance of constructs in explaining other constructs in the structural model. Information on the importance of constructs is relevant for drawing conclusions. For this reason, a Decision Matrix is considered a valuable decision making tool. In recent years, researchers have proposed valid approaches to solve problems related to the role that composite indicators have within that system. In building a CI, we are interested in (i) including elementary indicators on a non numerical scale, (ordinal and nominal data); (ii) including some kind of CI relationship (logical, hierarchical, temporal or spatial); (iii) defining the roles of the EIs (MVs) as mediator and moderator variables; and (iv) defining the roles of the CIs (LVs) in the inner model (mediator and moderator LVs). For this reason, many improvements, in order to extend the classic algorithm of PLS-PM to the treatment of particular data, have been made, in particular to non-metric data, mediator and moderator data and hierarchical data. Furthermore, several clustering techniques have been developed in PLS-PM to look for latent classes.

Non Numerical Models

PLS-PM is a technique devised to handle quantitative variables. However, in practice categorical indicators could be used to measure complex concepts as well. When we study complex phenomena in various research disciplines, some EIs are not on a numerical scale (nominal and ordinal variables). This kind of MV can play several different roles in PLS-PM, in particular it can have an active role in the analysis. An active categorical variable directly participates in the construction of the system of CIs. In other words, it is a categorical indicator impacting on a CI jointly with other indicators. In order to deal with this type of variable, the existing literature provides new algorithms to quantify and use the MVs for the estimation of an SEM, according to the PLS-PM algorithm. One of these is Partial Alternating Least Squares Optimal Scaling-Path Modeling (PALSOS-PM) [29] and another is called the Non-Metric PLS Path Modeling algorithm [31].

Modeling heterogeneity in PLS-PM

Another important topic in PLS-PM is the mediation and moderation effect. A significant mediator variable or moderator variable may to some extent absorb a cause-effect relationship. Examining these variables enables researchers to better understand the relationships between dependent and predictor constructs. Mediation and moderation are two important topics in the context of PLS-SEM. The mediation

function of a third variable represents the generative mechanism through which the focal independent variable is able to influence the dependent variable of interest. The moderator function of the third variable splits up a focal independent variable into sub-groups that establish its domains of maximal effectiveness with regard to a given dependent variable.

Higher-Order Constructs in PLS-PM

In Wold's original design of the PLS-PM [43] it was expected that each construct would be necessarily connected to a set of observed variables. On this basis, Lohmöller [27] proposed a procedure to treat hierarchical constructs, the so-called hierarchical component model. The hierarchical constructs are multidimensional constructs that involve more than one dimension. PLS-PM allows for the conceptualization of a hierarchical model, through the use of two main approaches existing in the literature: the Repeated Indicators Approach [27] and the Two Step Approach [39]. Different approaches have been developed and proposed in the literature: the Repeated Indicator Approach [27]; the Two Step Approach [39]; the Mixed Two Step Approach [5]; [10]; and the PLS Components Regression Approach [5]; [10].

3 Main bibliographic references

PLS-PM has been mainly developed by Wold, who was the first to formalize in his original article [45] the idea of Partial Least Squares as part of the analysis into principal components, introducing the NILES (Non Linear Iterative Least Squares) algorithm; subsequently this algorithm and its extension to the analysis of canonical correlations and to specific situations with multiple blocks of variables took the name of NIPALS (Non Linear Iterative Partial Least Squares) [44]. The first presentation of PLS Path Modeling was published by Wold in 1979, and the PLS-PM algorithm is described in two Wold's publications [43]; [42]. Two very important developments of the PLS approach to Structural Equation Models are by Chin [8] and Tenenhaus et al. [37]. In recent years, the number of published articles and books on the PLS-PM [14] increased significantly [15]. Several books have been published which are considered valuable manuals for researchers who want to explore these approaches [13]; [25]; [11]; [8]. Many article illustrate and analyzed how PS-PM can be used in many different applications [24]; [2] and other works propose methodological extensions to the basic PLS-PM approach [1]; [5]; [19], or make a critical analysis and review of some aspects of the PLS-PM [23]; [16]; [15]; [9]; [26].

4 Main application fields of PLS-PM

Today several researchers agree that some socio-economic phenomena cannot be measured by a single descriptive indicator and that, instead, they must be delineated by different dimensions, every measure a precise side of the phenomenon. Nowadays, phenomena such as Sustainable Development [2], Poverty [34], Social Inequality [6], Quality of Life [4], etc., require, so as to be measured, that the combination of various dimensions are unit thought of along because the proxy of the phenomenon [26]. In the literature, there are many works that propose PLS-PM as a method for learning of these phenomena, underling its advantages with respect to alternative simple and acknowledged approaches. As an example, recently, Cataldo et al. [2] have been proposed PLS-PM for studying existing Sustainable Development Goals (SDGs) indicators and have been demonstrated how it could help you to define the framework for SDGs indicators in order to provide a better measure of this complex multidimensional social phenomenon. There have been many and varied approaches in literature, such as simple arithmetic mean of the original indicators, Multidimensional Data Analysis (MDA) approaches, like Factorial Analysis (FA) or Principal Component Analysis (PCA) to measure sustainable development, based on simple aggregation techniques and easily calculated, however in their work the authors underlined how the choice of using the PLS-PM can facilitate researchers to identify critical indicators and to construct a ranking of countries. In addition, this approach, as compared to other multivariate techniques, examines simultaneous a series of dependence relationships, using a single statistical approach to test the full scope of projected relations [18].
PLS-PM is often used across different management disciplines, including organization research [35] and strategic management [17]. Hair et al. [17] review the applications of PLS-SEM and make some recommendations on how to improve the use of the method. According to Nitzl [30] PLS-SEM provides a useful tool for management accounting research due to the high degree of flexibility it offers for the interplay between theory and data [9], which seems urgently necessary given the current state of research in management accounting, especially with regard to developing a more holistic map of causes and effects [28].

5 Open issues in PLS-PM approach

The goal of this work has been to present the PLS-PM approach as a methodological framework that can be useful for creating a system of CIs in terms of decision-making, highlighting its potentiality. To date, there are several open issues, both from the methodological point of view of the PLS-PM model and from the point of view of new research areas. From the point of methodological point of view of the PLS-PM model, there are several aspects that need to be studied. Some are listed below:

- Several indices are used in order to evaluate partially the model but there is not yet an index for its global evaluation. Tenenhaus et al. [37] proposed a Goodness-of-Fit (GoF) as an index for validating the PLS model globally. Over the years, several researchers have criticized the usefulness of GoF both conceptually and empirically, arguing that GoF is not suitable for model validation [20]. There is a lack of adequate global and complete evaluation measures capable of evaluating the goodness of the PLS-PM model.
- Moreover, it would also be interesting to look further into the issue of considering different methods of estimation in place of the Ordinary Least Squares, inside the PLS-PM algorithm. Further research will be undertaken to find out if we can use a Weighted Least Squares method, namely a variant of the Ordinary Least Squares method, optimizing the weighted fitting criterion to find the parameter estimates that allow the weights to determine the contribution of each indicator to the final Composite Indicator estimates. The aim is to find an internal optimization in the algorithm, which allows us to have indicators weighted according to their importance and their predictive power within the model.

Instead, from the point of view of new research areas in PLS-PM approach, some aspects would to be considered in the model.

- Many phenomena need to be studied also considering qualitative information that help researchers to understand if there are differences among the units related to the analyzed issue, be they countries, regions or individuals. Therefore a frontier problem in the PLS-PM approach is the treatment and inclusion of this type of indicators in the model.
- Today, several researchers uses textual analysis and different approaches have been developed and proposed for the treatment of these types of data. Still open research field is the analysis of mixed data (official data, administrative data, networking data and social data), using some indicators extracted from different sources.
- To entirely study a phenomenon it is necessary to know its past and follow this phenomenon over time. The main objective of many researches is the study of the evolution over time of a phenomenon. Current treatments of longitudinal data are rather ad hoc and do not truly take time variant effects into account [22]. A preliminary work has studied how PLS-PM can be used to implement and to analyze the longitudinal data [3]. A still open issue concerns the analysis of time series, completed through the implementation of a longitudinal PLS-PM model.

These are just a few aspects that need to be analyzed. The goal is to make the PLS-PM a very useful tool that is easy to apply in various fields. Regardless of the advancements of the PLS-PM approach and its some open issues, the reliability of a model and its applicability largely depends on the quality of the data. The overall quality of the composite indicator depends on several aspects, related primarly to the quality of elementary data and to their availability. The lack of elementary indicators leads to the construction of incomplete indicators and consequently any method used to synthesize these indicators will be compromised. In fact, Stiglitz

et al. [36] in their *report on the measurement of economic performance and social progress* said that "what we measure affects what we do; and if our measurements are flawed, decisions may be distorted", emphasizing how data quality is important in making decisions.

References

1. Becker, J.M., Klein, K. and Wetzels, M.: Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. Long range planning, Elsevier, **45**,5–6 (2012)
2. Cataldo, R., Crocetta, C., Grassia, M.G. Lauro, N.C., Marino, M., and Voytsekhovska, V.: Methodological PLS-PM framework for SDGs system. Social Indicators Research, Springer, **156** (2), 701–723. (2021)
3. Cataldo R., Crocetta, C., Grassia, M.G. and Marino, M.: Longitudinal data analysis using PLS-PM approach. Pearson. (2020)
4. Cataldo, R., Corbisiero, F., Delle Cave, L., Grassia, M.G., Marino, M. and Zavarrone, E.: The Quality of Life in the Historic Centre of Naples: the use of PLS-PM Models to measure the Well-Being of the Citizens of Naples. Italian Studies on Quality of Life, Springer, 111–125. (2019)
5. Cataldo, R., Grassia, M.G., Lauro, N.C., and Marino, M.: Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. Quality & Quantity, Springer, **51** (2), 657–674. (2017)
6. Cherchye, L. Moesen, W. and Van Puyenbroeck, T.: Legitimately diverse, yet comparable: on synthesizing social inclusion performance in the EU. JCMS: Journal of Common Market Studies, Wiley Online Library, **42** (5), 919–955. (2004)
7. Chin, W.W. and Newsted, P.R.: Structural equation modeling analysis with small samples using partial least squares. Statistical strategies for small sample research, **1** (1), 307–341. (1999)
8. Chin, W.W.: The partial least squares approach to structural equation modeling. Modern methods for business research, Mahwah, NJ, **295** (2), 295–336. (1998)
9. Chin W.W.:Issues and opinion on structural equation modelling, Management Information. Systems quarterly, **22**(1), 1–8. (1998)
10. Crocetta, C., Antonucci, L., Cataldo, R., Galasso, R., Grassia, M.G., Lauro, C.N. and Marino, M.: Higher-order PLS-PM approach for different types of constructs. Social Indicators Research, Springer, **154** (2), 725–754. (2021)
11. Esposito Vinzi, V., Chin, W.W., Henseler, J. and Wang, H.: Handbook of partial least squares: Concepts, methods and applications. Heidelberg, Dordrecht, London, New York: Springer. (2010)
12. Fornell, C. and Bookstein, F.L.:Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. Journal of Marketing research, SAGE Publications Sage CA: Los Angeles, CA, **19** (4), 440–452. (1982)
13. Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications (2021)
14. Hair Jr, J.F., Risher, J.J., Sarstedt, M. and Ringle, C.M.: When to use and how to report the results of PLS-SEM. European business review, Emerald Publishing Limited. (2019)
15. Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M. and Thiele, K.O.: Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods. Journal of the academy of marketing science, Springer, **45** (5), 616–632. (2017)
16. Hair, J.F., Sarstedt, M., Ringle, C.M., and Mena, J.A.: An assessment of the use of partial least squares structural equation modeling in marketing research. Journal of the academy of marketing science, Springer, **40** (3), 414–433. (2012)

17. Hair, J.F. Sarstedt, M., Pieper, T.M. and Ringle, C.M.: The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. Long range planning, Elsevier, **45** (5-6), (320–340). (2012)
18. Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E.: Multivariate data analysis. Upper Saddle River, NJ: Pearson Prentice Hall. (2009)
19. Henseler, J.: Composite-based structural equation modeling: Analyzing latent and emergent variables. New York: Guilford Press. (2020)
20. Henseler, J. and Sarstedt, M.: Goodness-of-fit indices for partial least squares path modeling. Computational statistics, Springer, **28** (2), 565–580. (2013)
21. Henseler, J., Ringle, C.M. and Sinkovics, R.R.: The use of partial least squares path modeling in international marketing. New challenges to international marketing, Emerald Group Publishing Limited. (2009)
22. Hwang, H., Sarstedt, M., Cheah, J.H. and Ringle, C.M.: A concept analysis of methodological research on composite-based structural equation modeling: bridging PLSPM and GSCA. Behaviormetrika, Springer, **47** (1), 219–241. (2020)
23. Jarvis, C.B., MacKenzie, S.B. and Podsakoff, P.M.: A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of consumer research, Oxford University Press, **30**(2), 199–218. (2003)
24. Latan, H.: PLS path modeling in hospitality and tourism research: the golden age and days of future past. Applying partial least squares in tourism and hospitality research, Emerald Publishing Limited. (2018)
25. Latan, H., & Noonan, R. (Eds.): Partial least squares structural equation modeling: Basic concepts, methodological issues and applications. Berlin/Heidelberg: Springer. (2017)
26. Lauro, C. N., Grassia, M. G., and Cataldo, R.: Model based composite indicators: New developments in partial least squares-path modeling for the building of different types of composite indicators. Social Indicators Research, Springer, **135**(2), 421–455. (2018)
27. Lohmöller, J-B.: Latent variable path modeling with partial least squares. Springer Science & Business Media. (2013)
28. Luft, J. and Shields, M.D.: Subjectivity in developing and validating causal explanations in positivist accounting research. Accounting, Organizations and Society, Elsevier, **39** (7), 550–558. (2014)
29. Nappo, D.: Sem with ordinal manifest variables. An alternating least squares approach. Phd diss., University of Naples Federico II. (2009)
30. Nitzl, C.: The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. Journal of Accounting Literature, Elsevier, **37**, 19–35. (2016)
31. Russolillo, G.: Non-metric partial least squares. Electronic Journal of Statistics, Institute of Mathematical Statistics and Bernoulli Society, **6**, 1641–1669. (2012)
32. Saisana, M. and Tarantola, S.: State-of-the-art report on current methodologies and practices for composite indicator development. Citeseer. (2002)
33. Sarstedt, M., Ringle, C.M., Smith, D., Reams, R., Hair Jr, J.F.: Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers. Journal of family business strategy, Elsevier, **5** (1), 105–115. (2014)
34. Sen, A.: Poverty and famines: an essay on entitlement and deprivation. Oxford university press. (1982)
35. Sosik, J.J. Kahai, S.S. and Piovoso, M.J.: Silver bullet or voodoo statistics? A primer for using the partial least squares data analytic technique in group and organization research. Group & Organization Management, Sage Publications Sage CA: Los Angeles, CA, **34** (1), 5–36. (2009)
36. Stiglitz, J.E., Sen, A: and Fitoussi, J-P.: Report by the commission on the measurement of economic performance and social progress, Citeseer. (2009)
37. Tenenhaus, M., Vinzi, Esposito, V., Chatelin, Y.-M. and Lauro, C.N..: PLS path modeling. Computational statistics & data analysis, Elsevier, **4** (1), 159–205. (2005)

38. Trinchera, L., Russolillo, G. and Lauro, C.N.: Using categorical variables in PLS Path Modeling to build system of composite indicators. Statistica Applicata, **20**, (3-4), 309–330. (2008)
39. Wilson, B.: Using PLS to investigate interaction effects between higher order branding constructs. Handbook of partial least squares, Springer, 621–652. (2010)
40. Vinzi, Esposito, V., Trinchera, L. and Amato, S.: PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. Handbook of partial least squares, Springer, 47–82. (2010)
41. Von Bertalanffy, L.: General system theory: foundations, development, applications. George Braziller. Inc., New York. (1968)
42. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York, 581–591. (1985)
43. Wold, H.: Soft modeling: the basic design and some extensions. Systems under indirect observation, **2**, 343. (1982)
44. Wold, H.: Path models with latent variables: The NIPALS approach. Quantitative sociology, Elsevier, 307–357. (1975)
45. Wold, H.: Estimation of principal components and related models by iterative least squares. Multivariate analysis, Academic Press, 391–420. (1966)

Open issues in composite indicators construction

Problematiche aperte nella costruzione di indicatori compositi

Leonardo Salvatore Alaimo

Abstract Composite indicators are useful to represent in a easy-to-read way a complex phenomenon. Over the years, their use has significantly, both among academics and policy makers. At the same time, issues related to their use have emerged. They constitute open questions in the debate on the subject and frontiers for the research. In this paper, we aim to briefly present the state of the art on this topic and illustrate the main issues and the directions the literature has taken to address them. The latter constitute potential topics of interest also for those who want to undertake the study of composite indicators for the first time.

Abstract *Gli indicatori compositi sono utili per rappresentare in modo semplice e immediatamente comprensibile un fenomeno complesso. Nel corso degli anni, il loro uso è aumentato significativamente, sia tra gli accademici che tra i decisori pubblici. Allo stesso tempo, sono emerse questioni relative al loro uso, che costituiscono argomenti aperti nel dibattito e nuove frontiere per la ricerca. In questo articolo, ci proponiamo di presentare brevemente lo stato dell'arte su questo tema e di illustrare le questioni principali e le direzioni che la letteratura ha preso per affrontarle. Queste ultime costituiscono potenziali argomenti di interesse anche per chi voglia intraprendere per la prima volta lo studio degli indicatori compositi.*

Key words: Multi-indicators systems, Synthesis of statistical indicators, Composite indicators

1 Introduction

As Karl Pearson stated *if you haven't measured something, you really don't know very much about it*. Measurement allows the production of scientific knowledge

Leonardo Salvatore Alaimo
Department of Social Sciences and Economics, Sapienza University of Rome, e-mail: leonardo.alaimo@uniroma1.it

about reality. Indeed, it develops as a *dialogue between logic and evidence*, it is the result of a complex interaction between theory and observations represented and realized by measurement. This interaction is necessary and unavoidable [1]. Dealing with phenomena defining reality (wellbeing, poverty, quality of life, development, and so on) requires an approach capable of grasping their complex and multidimensional nature. They are *complex adaptive systems*, i.e. open systems made up of numerous elements interacting with each other, in a linear and a non-linear way, that constitute a unique and organic entity capable of evolving and adapting to the environment [2, 3, 4]. They are multidimensional and their different elements are linked together in a non-linear way. They evolve over time, modifying both their dimensions and the links between them. Consequently, their measurement needs to consider different aspects. They are not directly observable, but they derive theoretically from observations. Almost all measures in social sciences are developed by means of a *defining process*, namely achieved as a consequence of a definition confirmed through the relationship observed between observations and the concept to be measured. The measurement process in social sciences is associated with the construction of systems of indicators. It is necessary to use a variety of elementary indicators and a criterion for summarising the information they contain. In statistics, an elementary indicator refers to indirect measures of phenomena that cannot be measured directly. In this perspective, an indicator is not simply raw statistical information, but represents a measure organically linked to a conceptual model aimed at describing different aspects of reality. They are not simply collections of measures. Indicators within a system are interconnected and new properties typical of the system and not of its constituent elements emerge from these interconnections. Therefore, a system of indicators allows the measurement of a complex concept that would not otherwise be measurable by taking into account the indicators individually. They play a key role in describing and understanding socio-economic phenomena. The complex nature of systems of indicators requires approaches allowing more concise views in order to analyse and understand them. The guiding concept is *synthesis*. The synthesis of indicators' systems has become a main issue in the literature. A variety of statistical methods useful for this purpose have been defined and used. From a technical perspective, these methods can be classified into two different approaches: the aggregative-compensative [5] and the non-aggregative [6, 7, 8, 9]. In this paper, we focus on the first one, the dominant framework in literature. Despite its success, the aggregative-compensative approach has been criticised and a series of conceptual and methodological issues have been posed. These questions are still open and inflame the debate in the literature on this topic. In this paper, we focus on some of them and how they constitute frontiers for the research in the composite indicators' field. Why should we continue to work and research on composite indices? We will try to answer this question.

2 Composite indicators: some conceptual and methodological research questions

A system of indicators is a three-way data array of type "same objects × same indicators × time occasions", which can be algebraically formalised as [10]:

$$\mathbf{X} \equiv \{x_{ijt} : i = 1,\ldots,N;\ j = 1,\ldots,J;\ t = 1,\ldots,T\} \tag{1}$$

where the indices i, j and t stand for the units, the indicators and the times, respectively and x_{ijt} is the value of the j-th indicator observed for the i-th unit at time t. These data structures are characterised by a great complexity and require the use of specific statistical tools allowing a more concise view. Given $\mathbf{X} \equiv \{x_{ijt}\}$, the objective of the synthesis, generally, is to obtain a bi-dimensional data matrix:

$$\mathbf{V} \equiv \left\{v_{it} : i = 1,\ldots,N; t = 1,\ldots,T\right\} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1T} \\ v_{21} & v_{22} & \cdots & v_{2T} \\ \vdots & \ddots & \ddots & \vdots \\ v_{N1} & v_{N2} & \cdots & v_{NT} \end{pmatrix} \tag{2}$$

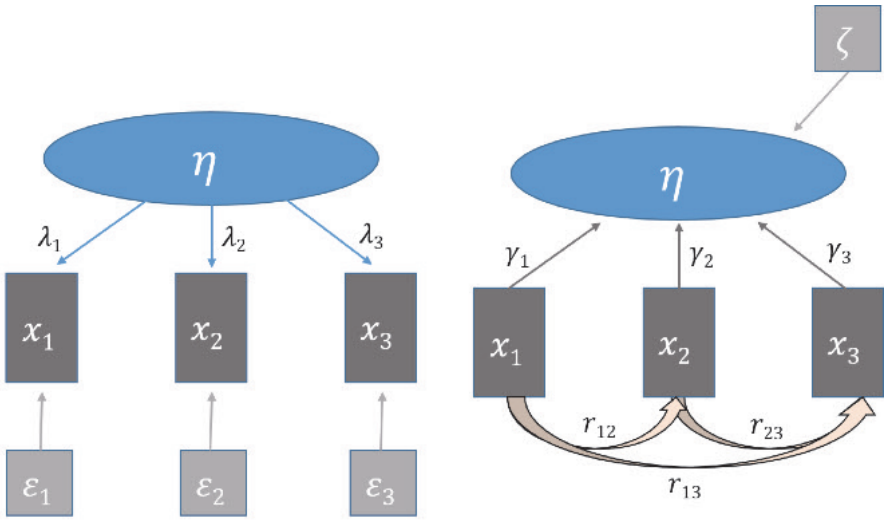
where v_{it} is the synthetic value of the unit ith at the time t-th. In the aggregative-compensative approach, the synthesis of $\mathbf{X}$ is performed by means of a mathematical function that combines the (previously normalised) basic indicators. In other words, it consists of the mathematical combination (or aggregation) of the set of indicators, obtained by applying specific methodologies [11] known as composite indicators (CIs). Over the years, these methodologies have been widely used in literature and by various international organisations and institutions for measuring and evaluating a great variety of socio-economic phenomena. The main purpose of their importance and success is to be informative. It is easier for the public to understand a synthetic indicator (one single measure) than many elementary indicators.

One of the main critical points is the treatment of multidimensional systems of ordinal data [7]. Ordinal indicators cannot be synthesised by using an aggregative method, suitable only for cardinal data. In fact, ordinal scores cannot be treated as numbers. Despite this, we often see their transformation into numerical scores, by more or less sophisticated scaling tools, in order to make possible their synthesis by aggregative procedures. These procedures may lead to controversial and incorrect results and pose delicate methodological and conceptual questions. This has led researchers to identify methods that can deal with non-cardinal indicator systems.

Another focal issue in composites construction is how to treat subjectivity. It is involved in any phase of composites' construction. Subjectivity is not negative per se, but it becomes so when it turns into *arbitrariness*. The first step in any synthesis is the definition of the phenomenon we want to measure and the subsequent identification of the theoretical framework and the relevant variables. The concept must always refer to a theoretical framework that gives it meaning. No meaning can be attributed without subjectivity. The role of the subject in knowledge production is clear. Fundamental attention must be given to the analysis of the *measurement*

model, referring to the relationship between concepts and indicators. The debate on measurement models is part of the literature on the evaluation of latent variables, which has a long tradition in social science [12]. Latent variables are phenomena of theoretical interest which cannot be directly observed and have to be assessed by manifest measures which are observable. Two different conceptual approaches can be identified: *reflective* and *formative* [13, 14, 15, 16, 17].

Fig. 1: Measurement models: reflective (left); formative (right).



The reflective measurement models have a long tradition in social sciences (in particular, in psychometric research) and are based on classical test theory, according to which measures are effects of an underlying latent construct [18]. Therefore, causality is from the construct to the measures and, consequently, a change in the latent variable causes variation in all measures simultaneously (all indicators must be positively correlated). In a formative model, indicators are causes of the construct rather than its effects (like in the reflective one) and they determine the latent variable giving it its meaning [13, 19]. Accordingly, indicators are not interchangeable: omitting an indicator is omitting part of the construct [20]. Thus, the choice of indicators determines what we want to measure.

The literature about the difference between reflective and formative models is rich and the debate on this issue continues. We would like to point out that the choice between the two models does not depend on the researcher, but exclusively on the nature and direction of relationships between constructs and measures. Different methods of normalisation, weighting and aggregation exist and can be used, leading to different results and interpretations. Of course, the choice of methods is also subjective, although it must be guided by knowledge of the phenomenon and

based on clear assumptions so as not to be arbitrary. Each method has strengths and weaknesses. Different choices lead to different syntheses that often give a different interpretation of the phenomena studied. These considerations lead to two research questions. The first is whether, given a system of indicators, there can be a method that is better than the others, i.e. that is able to represent the phenomenon better than the others. The second question, strictly linked to the previous one, is whether it is possible to define a criterion for choosing such a method.

As highlighted in equation 2, the synthesis aims at obtaining for each unit of the original system a synthetic measure that is representative of its original profile (i.e., the combination in the basic indicators) at a specific time t-th. Such a measure gives an easy-to-read information about the phenomenon. Switching from multi-dimensional to uni-dimensional necessarily determines a loss of information, justified by the need to have a synthetic view of the measured phenomenon. In many cases, this loss of information is excessive. Synthesising a complex phenomenon into a single number can be not straightforward and lead to misleading results and conclusions which increase if the indicator is poorly defined and constructed. This can lead researchers and/or policy-makers to give an over-simplistic interpretation of a phenomenon. This aspect has been investigated in literature and has prompted researchers to question whether the synthesis of a multi-indicators system must necessarily be a single number assigned to each statistical unit at a specific time.

3 Frontiers of the research

The questions presented in the previous section constitute challenges for researchers and the answer to them might be a reason to approach the study of composite indicators.

The impossibility of synthesising indicator systems in which non-cardinal indicators are also present is intrinsically linked to the nature of composite indicators, which are obtained through the mathematical combination of elementary indicators. For this reason, over the years the research has focused on finding methods suitable for dealing with systems of indicators at different scaling levels. In this way, the so-called non-aggregative approach gradually became widespread: the synthetic indicator is obtained without any aggregation of the basic indicators. Among the different methodologies belonging to this approach (for instance, the Social choices theory [21, 22, 23] or the Multi-criteria Analysis [24, 25, 26]), the Partially Ordered Set (poset) Theory [27, 28, 29] has become a reference. The spread of these new methods was facilitated by the concomitant spread of increasingly powerful computer tools, which made their computation possible. Undoubtedly, research is moving towards the identification of methods that do not depend on and can, consequently, be used regardless of the scale of the elementary indicators. Subjectivity is an ineradicable element, but it must never become arbitrariness. Research has also focused on the management of the various subjective choices involved in the composites' construction. As regards the definition of the phenomenon to be measured

and the choice of the elementary indicators, one way to avoid arbitrariness is to stand *on the shoulders of giants* [30], i.e. always rely on a careful analysis of the literature and what others have done before. This does not translate into a kind of immobility, into the impossibility of departing from what has been done in the past. On the contrary, research in the field of indicators is highly dynamic as it is linked to societal evolution. For instance, if we wanted to construct an indicator measuring deprivation, we could not disregard the work on this subject by Townsend [31]. It is clear that the deprivation nowadays is not the same as it was in Townsend's reference and that the concept must therefore be adjusted. However, Townsend's work would be the starting point. Research in the field of indicators' synthesis is alive and evolving; phenomena change in different contexts (spatial, temporal, cultural). In this perspective, the researcher plays a decisive role and subjectivity becomes the lens through which he or she observes the world in an unique way. As mentioned above, the methodological choices are also subjective (obviously, based on the knowledge of the phenomenon) and different methods lead to different results. The research therefore focused on identifying *the best method* for the synthesis. There is no absolute method that is preferable to all others. However, a criterion for choosing a method would be useful. We often deal with CIs obtained by the most different methods, often chosen arbitrarily by the researchers. This makes the choices questionable. But, even if the choices are agreeable, it remains to be seen how much a method is a "good and valid choice". In this perspective, different authors [32, 33] have suggested robustness as a selection criterion: among the different choices and methods, we must select those which guarantee greater robustness of rankings, assessed by means of uncertainty analysis (how uncertainty in the input factors propagates through the structure of the composite index and affects the results) and sensitivity analysis (how much each individual source of uncertainty contributes to the output variance). However, this approach leaves a question open: why should a more robust method better represent a phenomenon? In particular, the idea that preferring the method which, by excluding and including individual indicators and setting different decision rules to construct the composite index, leaves the rankings obtained most unchanged is highly questionable. It could be argued that such an approach does not take the measurement model into account. Indeed, in a reflective model, the exclusion of an indicator does not affect the latent variable that is being measured. On the contrary, in a formative model excluding or including an indicator changes, even strongly, the measured latent variable. Let's take an example. Suppose we want to measure human development using UNDP's framework [34], considering three dimensions and four indicators: a long and healthy life assessed by life expectancy at birth; knowledge measured by means of mean years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age; a decent standard of living measured by gross national income per capita. If we remove the economic variable, we expect a significant change in the ranking obtained which will be different from that obtained by excluding life expectancy. Consequently, why, among the various methods, should we choose the one that leaves the rankings obtained by excluding an indicator more unchanged? As easily understood, the debate on this issue is very lively. Recently, an approach

linking the choice of method to the nature and structure of the indicators' system has been proposed. It is quite obvious that a good synthetic measure should give a good fit of the distributive assumptions on data. In other words, a composite can be considered "good" if it is able to give a good representation of the distributional form (or multiple forms) assumed on the system of indicators.

The last question addressed in this work is if the synthesis must necessarily be a single number or, more precisely, whether a single number is capable of accounting for the complexity of the observed phenomenon. In literature, we can find arguments in favour of the composites and against them. Some scholars [35] criticised the choice of constructing a single composite index, suggesting that it would be a better choice to use a dashboard, because it allows to avoid an arbitrary choice of the functional form and the weighting scheme and to observe a phenomenon from multiple points of view. In this perspective, the synthesis is an informative patrimony capable of describing the observed reality. Other researchers highlighted that a synthetic measure can be an object, a map or an image. There is a large amount of literature on the use of metaphoric images for the representation and synthesis of socio-economic phenomena [36, 37]. Another approach is to use intervals of composites rather than individual measures [38, 39, 40, 41]. The proposed intervals, although different one another, all respond to the idea of identifying a range of values within which the synthetic measure is included.

4 Conclusions

Composites indicators are a tool for measuring and understanding phenomena. They have become the focus of attention of researchers and policy makers for their ease of reading and usefulness for decision making and evaluation. Over the years, their use has increased, as well as the areas in which they have been applied. At the same time, the debate in the literature has become increasingly animated, focusing on problems and new areas of application and frontiers of research. In this paper, we have presented some of them, which are, of course, only examples that, although relevant, do not do justice to the enormous academic debate and production on composite indicator topic. This testifies to the liveliness of research in this field, the possibility of exploring new or established themes from new perspectives. Undoubtedly, we can consider this an adequate answer to why we must continue to study composite indicators.

References

[1] Leonardo S. Alaimo. Complexity of Social Phenomena: Measurements, Analysis, Representations and Synthesis. *Unpublished Doctoral Dissertation, University of Rome" La Sapienza", Rome, Italy*, 2020.

[2] Mitchell M. Waldrop. *Complexity: The Emerging Science at the Edge of Order and Chaos*. New York: Simon and Schuster, 1992.

[3] Leonardo S. Alaimo. Complexity and knowledge. In F. Maggino, editor, *Encyclopedia of Quality of Life and Well-being Research*, pages 1–2. Cham: Springer, 2021. doi: 10.1007/978-3-319-69909-7_104658-1.

[4] Leonardo S. Alaimo. Complex systems and complex adaptive systems. In F. Maggino, editor, *Encyclopedia of Quality of Life and Well-being Research*, pages 1–3. Cham: Springer, 2021. doi: 10.1007/978-3-319-69909-7_104659-1.

[5] OECD. Handbook on Constructing Composite Indicators. Methodology and User Guide, 2008.

[6] Rainer Bruggemann and Ganapati P Patil. *Ranking and prioritization for multi-indicator systems: Introduction to partial order applications*. Dordrecht: Springer Science & Business Media, 2011.

[7] Marco Fattore. Synthesis of Indicators: The Non-aggregative Approach. In F. Maggino, editor, *Complexity in Society: From Indicators Construction to their Synthesis*, pages 193–212. Cham: Springer, 2017.

[8] Leonardo S. Alaimo, Alberto Arcagni, Marco Fattore, and Filomena Maggino. Synthesis of multi-indicator system over time: A poset-based approach. *Social Indicators Research*, pages 1–23, 2020. doi: 10.1007/s11205-020-02398-5.

[9] Filomena Maggino, Rainer Bruggemann, and Leonardo S. Alaimo. Indicators in the framework of partial order. In R. Bruggemann, L. Carlsen, T. Beycan, C. Suter, and F. Maggino, editors, *Measuring and Understanding Complex Phenomena: Indicators and their Analysis in Different Scientific Fields*, pages 17–29. Cham: Springer International Publishing, 2021.

[10] P. D'Urso. Dissimilarity measures for time trajectories. *Stat. Methods Appl.*, 9(1-3):53–83, 2000.

[11] Michela Nardo, Michaela Saisana, Andrea Saltelli, and Stefano Tarantola. Tools for composite indicators building. *European Commission, Ispra*, 15(1):19–20, 2005.

[12] Otis D. Duncan. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation, 1984.

[13] Hubert M. Blalock. *Causal Inferences in Nonexperimental Research*. N.C.: University of North Carolina Press, 1964.

[14] Kenneth A. Bollen. *Structural Equations with Latent Variables*. New York: Wiley, 1989.

[15] Adamantios Diamantopoulos and Heidi M. Winklhofer. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, 38(2):269–277, 2001.

[16] Adamantios Diamantopoulos and Judy A Siguaw. Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British journal of management*, 17(4):263–282, 2006.

[17] Adamantios Diamantopoulos, Petra Riefler, and Katharina P. Roth. Advancing Formative Measurement Models. *Journal of Business Research*, 61(12):1203–1218, 2008.

[18] Kenneth A. Bollen and Richard Lennox. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, 110(2):305, 1991.

[19] Hubert M. Blalock. The Measurement Problem: A Gap between the Languages of Theory and Research. In F. Kerlinger, editor, *Methodology in Social Research*, pages 5–27. New York: McGraw-Hill, 1968.

[20] Kenneth A. Bollen. Multiple Indicators: Internal consistency or No Necessary relationship? *Quality and Quantity*, 18(4):377–385, 1984.

[21] Amartya Sen. Social choice theory: A re-examination. *Econometrica: Journal of the Econometric Society*, pages 53–89, 1977.

[22] Amartya Sen. Social Choice Theory. *Handbook of mathematical economics*, 3:1073–1181, 1986.

[23] Iain McLean. The Borda and Condorcet principles: Three Medieval applications. *Social Choice and Welfare*, 7(2):99–108, 1990.

[24] Phil Macoun and Ravi Prabhu. *Guidelines for applying multi-criteria analysis to the assessment of criteria and indicators*, volume 9. CIFOR, 1999.

[25] Peter Nijkamp and Ad van Delft. *Multi-criteria analysis and regional decision-making*, volume 8. Springer Science & Business Media, 1977.

[26] Constantin Zopounidis and Panos M Pardalos. *Handbook of multicriteria analysis*, volume 103. Springer Science & Business Media, 2010.

[27] Joseph Neggers and Hee S. Kim. *Basic Posets*. Singapore: World Scientific Publishing, 1998.

[28] Hilary A. Priestley. Ordered Sets and Complete Lattices. In R. Backhouse, R. Crole, and J. Gibbon, editors, *Algebraic and Co-algebraic Methods in the Mathematics of Program Construction. International Summer School and Workshop, Oxford, April 10-14, 2000, revised lectures*, pages 21–78. Dordrecht: Springer, 2002.

[29] Berndt Schröder. *Ordered Set. An Introduction*. Boston: Birkäuser, 2002.

[30] Robert K. Merton. *On the Shoulders of Giants: A Shandean Postscript*. San Diego (Calif.): Harcourt Brace Jovanivich, 1985.

[31] Peter Townsend. Deprivation. *Journal of Social Policy*, 16(2):125–146, 1987.

[32] Michael Freudenber. *Composite Indicators of Country Performance*. Paris: OECD Publishing, 2003.

[33] Matteo Mazziotta and Adriano Pareto. Synthesis of Indicators: The Composite Indicators Approach. In F. Maggino, editor, *Complexity in Society: From Indicators Construction to their Synthesis*, pages 159–191. Cham: Springer, 2017.

[34] UNDP. *Human Development Report 2020: The Next Frontier Human Development and the Anthropocene*. New York, NY: UNDP, 2020.

[35] Ed Diener and Eunkook Suh. Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, 40(1):189–216, 1997.

[36] Edward R. Tufte. *The Visual Display of Quantitative Information*, volume 2. Cheshire, Connecticut: Graphics Press, 2001.

[37] Manuel Lima. *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press, 2013.

[38] Luis Dıaz-Balteiro and Carlos Romero. In search of a natural systems sustainability index. *Ecological Economics*, 49(3):401–405, 2004.

[39] Francisco J Blancas, Rafael Caballero, Mercedes González, Macarena Lozano-Oyola, and Fátima Pérez. Goal programming synthetic indicators: An application for sustainable tourism in andalusian coastal counties. *Ecological Economics*, 69(11):2158–2172, 2010.

[40] Matteo Mazziotta and Adriano Pareto. Composite indices construction: The performance interval approach. *Social Indicators Research*, pages 1–11, 2020.

[41] Emiliano Seri, Leonardo S. Alaimo, and Vittoria C. Malpassuti. BoD – min range: A Robustness Analysis Method for Composite Indicators. In A. Pollice, N. Salvati, and F. Schirripa Spagnolo, editors, *Book of Short Papers SIS 2020*, pages 1154–1159. Milano: Pearson, 2020.

The posetic approach to the construction of socio-economic indicators: open issues and research opportunities

L'approccio "poset" alla costruzione di indicatori socio-economici: problemi aperti e opportunità di ricerca

Marco Fattore

Abstract The paper introduces and motivates the "posetic approach" to the construction of socio-economic indicators, providing an overview of its recent developments and identifying open issues and research opportunities. While the use of order theory paves the way to new developments in data analysis particularly, but not exclusively, when dealing with ordinal data systems, it also poses interesting and non-trivial conceptual, methodological and computational problems which challenge researchers. To illustrate and discuss them, in view of raising interest towards the topic, is the main aim of the paper.

Abstract *L'articolo presenta e motiva l'approccio "posetico" alla costruzione di indicatori socio-economici, fornendo una panormaica dei suoi più recenti sviluppi e identificando criticità a opportunità di ricerca. Benche l'utilizzo della teoria delle relazioni d'ordine apra a nuovi sviluppi nell'analisi dei dati multidimensionali particolarmente, ma non in via esclusiva, nel caso ordinale, esso pone ai ricercatori anche interessanti e non banali problemi, di tipo concettuale, metodologico e algoritmico. Illustrarli e discuterli, per stimoalre l'interesse verso il tema, è il primcipale obiettivo dell'articolo.*

Key words: Order structures, Multi-indicator systems, Partial order algorithms, Synthetic indicators

1 Introduction

This short paper aims at motivating the so-called "posetic approach" to the construction of socio-economic indicators, sketching its recent developments and mostly discussing the open issues and the related research opportunities. As better explained in

Marco Fattore
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 - Milano e-mail: marco.fattore@unimib.it

the following sections, the attempt to systematically use partially and quasi-ordered structures in data analysis, particularly in the analysis of multi-indicator systems, is quite recent; it reflects both a progressive shift of paradigm towards a "soft modeling" and structural approach to the statistical measurement and evaluation of complex phenomena and the increasing availability of software resources, which are of key importance given the combinatorial nature of the algorithms, required to implement posetic tools in real data analysis. In particular, the use of order theory in indicator construction enhances what has come to be called the "non-aggregative approach" to synthesis, which overcomes many drawbacks of composite indicators and allows the consistent treatment of ordinal attribute systems, not requiring the aggregation of variable scores. While this opens new opportunities for indicator construction, at the same time it poses various kinds of conceptual, methodological and computational problems. To discuss them is the main aim of the paper, which is organized as follows: Section 2 introduces the main motivations of the posetic approach; Section 3 provides a short literature review; Section 4 illustrates some essentials of the approach through a simple example; Section 5 discusses the main open issues and the research opportunities, from methodological, applicative and computational points of view; Section 6 concludes.

2 Why the posetic approach to indicator construction

Partially/quasi-ordered sets and order theory play a prominent role in the analysis of multi-indicator systems and in the construction of synthetic indicators, since they provide the natural data structures for this kind of statistical problems, together with the mathematical tools needed to address them, properly [7, 8]. Many topics in applied statistics and multi-criteria decision making, e.g. the evaluation of multidimensional traits like deprivation or well-being, involve complex systems of *ordinal* indicators that cannot be synthesized with metric tools. The indicators simply order statistical units in terms of *greater than* / *lower than* and since they are likely to order units in conflicting ways, just a partially or quasi-ordered set results. Indicator construction then requires the extraction of information from *partially/quasi-ordered structures*, with the aid of the mathematics of *order theory* [5, 17]. In the posetic approach to indicator construction, statistical units get scored, and possibly ranked, not based on some aggregation of the input variables but, informally stated, based on their relational position in the network of multidimensional comparisons they are embedded in, through the order relation set on them. In other words, information is extracted from the *relational structure* of the data (i.e. the partial/quasi order relation) and not just from the *unstructured set* of input variables. Avoiding aggregative-compensative procedures makes it possible to treat multi-indicator systems of ordinal variables in a sound and consistent way, but it must be noticed that the role of order theory in indicator construction goes beyond the ordinal case. Socio-economic statistics often deals with multi-faceted and multi-shaped phenomena, where the impossibility of comparing all of the respective manifestations is

intrinsic to their essential complexity. Even if they are described through numerical variables that, at least technically, can be treated with usual aggregative tools (e.g. averages and other kinds of means), it may well be meaningless to do so. In such cases, the posetic approach overcomes the drawbacks of composite indicators, particularly their compensative nature which hides the nuances and facets of the traits under assessment and often makes their interpretation difficult. Once partially and quasi-ordered sets are acknowledged as proper structures for data analysis, and in particular for synthetic indicator construction, the development of a comprehensive set of algorithms and procedures for *doing statistics* becomes mandatory. This is why many research opportunities open.

3 A short literature review

From the early attempts in the '80s of the last century, the literature on the posetic approach to indicator construction and data analysis has been growing, particularly in the last 20 years. Most of the developments are related to the analysis of multi-indicator systems, particularly (but not exclusively) when ordinal attributes are involved. Main problems here are the construction of rankings, the evaluation of multidimensional and multi-faceted traits, the prioritization of tasks or interventions... often in the socio-economic and environmental field, although the range of applications is growing. To help readers addressing the topic, here we provide a synthetic list of some of the main references about order theory and posetic tools, according to different subtopics.

- *Mathematics of order relations*. Statistical applications of order theory rely on a wide and sound body of mathematical concepts and results that must be, at least to some extent, acquired by any serious "posetic" practitioner. The standard reference on order theory, with an emphasis on lattice theory, is the book by Davey and Priestley [5], which provides a comprehensive view of many fundamental topics, with also an eye to other application fields, like computer science. The book by Schroeder [17] is another important reference, more focused on posets and combinatorial aspects. Many other resources do exist, both on the general mathematical foundations of order relation theory and focusing on more specific topics. Detailed references can be found in cited texts.
- *Statistical methodology and application areas*. As to indicator construction, the main application areas of posetic tools are in evaluation studies in the socio-economic (e.g. [9, 1, 12]) and environmental ones [3], where multi-indicator systems, often of an ordinal nature, are one of the most typical data structures. It is virtually impossible to provide a list of the hundreds of methodological and applied papers existing in literature, on evaluation, ranking, prioritization, indicator construction, sensitivity analysis and many other topics. A good overview of tools and applications, with many references to other resources, can be however found in the books edited by Brueggemann and Patil [3] and by Fattore and Brueggemann [10]. Other references will be given when discussing open issues.

- *Computational aspects*. Computational aspects are of outermost importance in applications of order theory to data analysis, due to the complex combinatorial nature of many posetic tools. To our knowledge, however, there is no comprehensive resource providing a collection of posetic algorithms, which must be retrieved in single papers spread across the literature. Among the main computational issues, maybe the most relevant one pertains to the computations of so-called *mutual ranking probability* (see Section 4), which are involved in many posetic applications. Here, the main reference is [6], while others will be given in Section 5.

4 A small paradigmatic example

Just to give the flavor of how posetic tools can be used in indicator construction, let us consider the following toy example. Table 1 reports the position of some Belgium regions, on the three pillars of competitiveness, according to the European Regional Competitiveness Index (for details on the index, see *https://ec.europa.eu/regional_policy/en/information/maps/regional_competitiveness/*). As it can be seen, while some regions dominate others, being in better positions on each pillar, other pairs of regions cannot be compared, having "conflicting" positions, on different pillars. As a result, the seven Belgium regions here considered get represented as a *partially ordered set*, or a *poset* for short, as depicted in the left panel of Figure 1. In the diagram, a generic region A dominates a generic region B if and only if a descending sequence of edges exists, linking the former to the latter; if no such path exists between two nodes, then the corresponding regions are said to be *incomparable*. As it can be seen, among the selected regions, none dominates all of the others (i.e. there is no *maximum*, but two *maximal* units) and none is dominated by all of the others (i.e. there is no *minimum*, but two *minimal* elements). Some regions can indeed be ordered (e.g. $BE34 < BE33 < BE22 < BE21$ and $BE32 < BE25 < BE21$ provide two so-called *chains*), while others cannot (e.g. $BE22$ and $BE32$ are *incomparable*, written $BE22||BE32$, and form a so-called two-element *antichain*). Given the poset of Figure 1, typical questions are whether or not it is possible to order the seven regions, in a complete ranking and how to achieve this. To this goal, some "competitiveness score" should be computed, to linearly order the regions accordingly. But dealing with data pertaining to in-homogeneous dimensions, we do not want to aggregate them and must search for a different scoring strategy. From a posetic point of view, the ranking information we look for is comprised in the structure of the input partial order and from it must be extracted. As formally explained in [13], a way to do this is to perform the following steps: (i) construct all of the possible rankings of the seven Belgium regions, that can be formed without conflicting with the dominances of the input poset (these are called *linear extensions*), (ii) compute how frequently one region dominates another in such a set of rankings and (iii) use these frequencies to assess the dominance degrees of each region, over the others. In practice, the dominance frequencies are turned into so-called *mutual*

ranking probabilities and arranged into the *mutual ranking probability matrix M*, reported in Table 2. To get the final dominance scores, mutual ranking probabilities must be synthesized, e.g. by computing the first eigenvector of $M^T M$ [13] (i.e. performing a 1-dimensional Singular Value Decomposition of M), getting the results reported in the last column of Table 2 (the final ranking is depicted in the right panel of Figure 1). As it can be noticed, two regions share the same score, since the input poset is invariant under exchanging them in the Hasse diagram (i.e., since they share equivalent positions in it). Despite its simplicity, the example shows in action the "essence" of the posetic approach: information extraction from the structure of the partial order relation (here, through mutual ranking probabilities) and no variable aggregation.

Table 1 Positions of seven Belgium regions, on the three JRC competitiveness pillars (lower positions mean higher competitiveness.

Region	Code	Basic	Efficiency	Innovation
Antwerp	BE21	1	5	4
Linburg	BE22	2	7	6
Oost-Vlaanderen	BE23	7	1	5
West-Vlaanderen	BE25	8	6	9
Hainaut	BE32	9	11	10
Liege	BE33	6	9	9
Luxembourg	BE34	11	10	11

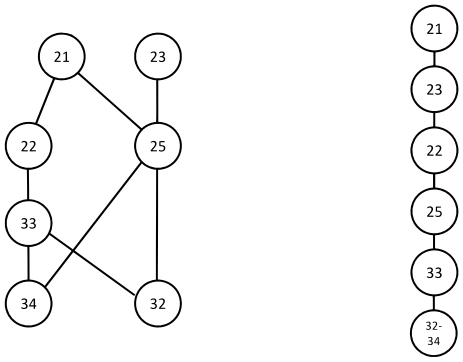


Fig. 1 Left panel: Hasse diagram of the selected Belgium regions; Right panel: final ranking

5 Open issues and research opportunities

The construction of indicators on order structures can be seen as a part of a larger problem, that of developing what could be called an "ordinal multidimensional data

Table 2 Mutual ranking probabilities between Belgium regions (entry ij is computed as the probability of picking, uniformly at random, one linear extension of the input poset such that region j dominates region i; in practice, the entry ij can be considered as the degree of dominance of region j over region i).

	BE21	BE22	BE23	BE25	BE32	BE33	BE34	Dominance score
BE21	1.00	0.00	0.33	0.00	0.00	0.00	0.00	0.51
BE22	1.00	1.00	0.67	0.22	0.00	0.00	0.00	0.45
BE23	0.67	0.33	1.00	0.00	0.00	0.11	0.00	0.48
BE25	1.00	0.78	1.00	1.00	0.00	0.44	0.00	0.36
BE32	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.16
BE33	1.00	1.00	0.89	0.56	0.00	1.00	0.00	0.35
BE34	1.00	1.00	1.00	1.00	0.50	1.00	1.00	0.16

analysis", to fill a methodological and practical gap in statistics. Indeed, while a number of tools and algorithms are available to address a great variety of multidimensional data analysis problems with numerical variables, the same cannot be said for the case of ordinal attributes. Beyond the ordinal case, as already mentioned, it is the even more general problem of doing statistics in "partially/quasi-ordered data spaces" to be definitely open. It is in this perspective that below we list some main issues that deserve to be addressed. We organize the list (that is by no means exhaustive and surely reflects unavoidable Author's biases) in three main parts, pertaining to methodological, applicative and computational developments, respectively.

Methodological developments. In the study of multidimensional indicator systems, the typical research questions are related to the extraction of synthetic information, out of the data. And this usually comes under either of two forms: (i) reducing the dimensionality of the input data systems, e.g. producing a ranking of the units scored against the attribute variables, or (ii) clustering the units, reducing the scored population to a smaller set of equivalence classes. While there exist many statistical procedures to pursue these goals in Euclidean spaces, and in more general metric spaces, just a few tools are available for ordered spaces and we can highlight the following main open issues:

- *Dimensionality reduction*. While some algorithms for scoring and ranking partially/quasi ordered units exist (as shown in the toy example), there is no general algorithm to map ordered structures into low dimensional spaces, in the spirit of principal component analysis, Singular Value Decomposition and the other classical dimensionality reduction techniques. Indeed, there exists an algorithm, called POSAC (Partial Order Scalogram Analysis with Base Coordinates, [18]) for the planar visualization of quasi-ordered data, but its theoretical bases are not that strong and it is designed just for bidimensional reductions (in addition, no detailed presentation of the algorithm can be easily retrieved in the literature, although an `R` implementation has been recently made available). A particularly interesting subproblem, in this context, pertains to dimensionality reduction for binary ordinal data. These are increasingly relevant in many application areas,

but almost no tools exist for reducing them to lower dimensional spaces or to provide planar approximations and visualizations of their order structure.

- *Clustering*. Typical cluster procedures assume the existence of a metric in the data space, used to group more similar units together, partitioning the input population. When just ordinal information is available, this approach cannot be pursued. Some cluster analysis procedures exist that employ some metrics or similarity measures between vectors of ordinal scores, but they are quite inconsistent with the nature of the problem. Indeed, the clustering problem on order structures is not just to group similar units together, but to do this *setting* also an order relation on the resulting set of equivalence classes. In other words, a "posetic" cluster procedure should get a quasi-ordered set as input and provide a partially ordered set (of clusters) as output. To do this, the order structure of the input data must be exploited and, currently, there are no algorithms that perform this task.

Dimensionality reduction and clustering are not the only methodological issues which deserve attention, from a posetic perspective. A further very relevant problem pertains to the *analysis of frequency or probability distributions defined on partially ordered sets*, particularly in view of describing their shape and features. For example, how to measure the inequality or the polarization degree of a population scored against a set of ordinal attributes, pertaining to some kind of socio-economic achievements? We touch upon this problem, in the following paragraph.

Applications. As to practical applications of posetic tools, virtually any field where multi-dimensional data systems are to be treated in view of evaluation, prioritization, ranking or similar issues can benefit from these techniques. Here, just a few hints are proposed.

A major open issue in socio-economics, the field where posetic techniques have been mostly applied insofar, is to build multi-dimensional indicators for evaluating the inequality or the polarization of traits like deprivation, well-being or similar multi-faceted concepts. While various proposals exist for the numerical case, when ordinal attributes are considered no satisfactory way to assess these key features of multi-dimensional socio-economic variables exist yet. Indeed, various attempts have appeared in the literature, but these do not contextualize inequality/polarization measurement within partially ordered sets, addressing it in an aggregative way and building the final measures as a composition of the inequality/polarization of the single input variables. This way, the order structure of the domain of the frequency distributions is not considered, losing a great deal of information on the data. The measurement of inequality/polarization is instead to be addressed as the problem of characterizing the shape of probability/frequency distribution of a partially ordered variable or, as equivalently suggested in [11], as the problem of building suitable functionals over partially/quasi ordered data. Once the order structure of the data comes into play, other interesting issues arise, namely on how to decompose the inequality/polarization measure to identify the contribution of the single variables and of their interaction.

A second application area, where posetic tools could be very effectively employed, is that of *prioritization* and *preference* analysis. A typical example is that of policy-design, where actions must be chosen based on complex sets of criteria that often results into a partial ordering of the alternatives. Although in a completely different context, this is the very same problem of consumer choice, e.g. on e-commerce websites, when one has to choose among tens or hundreds of products, based on large feature sets (think, for example, of electronic products, like smartphones or televisions).

Beyond socio-economics, policy-making or choice theory, multi-dimensional ordinal indicator systems can be found in many other disciplines where evaluation is the key goal. For example in *psychometrics* (indeed, the aforementioned POSAC algorithm was developed in the context of *Facet* theory) or in *education* sciences, where students must be assessed and score against various achievement scales. Both the posetic tools already available and those we have highlighted as "nice-to-have" could be applied within these contexts too, improving the quality and the reliability of assessment and evaluation processes.

Computational and algorithmic issues. All the topics and issues introduced in the paragraphs above have a computational counterpart, involving the investigation of various features of the input order structure, which often leads to complex combinatorial problems. In this respect, the main issue is the computation of mutual ranking probabilities, which underlie many of the applications of posetic tools to data analysis. Various algorithms exist, for exact and approximate [4, 6, 15] *mrp* computations, but in many real cases the computational complexity is nevertheless excessive. Ways out to this problem should focus on (i) the *design and implementation of algorithms for specific classes of posets* (e.g. product orders, linear sum orders, lexicographic orders…), where computations can exploit symmetries or useful features of the inputs, or on (ii) the search for *closed approximated formulas*, for different poset classes [2]. It must be noticed that in many cases mutual ranking probabilities are used as drivers for other posetic computations (e.g. in optimal poset approximation, or for ranking extraction [13, 16]) and that the same kind of computation could be driven equally well by simpler numerical criteria, expressing in a less sophisticated way the degree of dominance between poset elements. Some alternatives should be investigated, implemented and tested, comparing their performances in various contexts. This is of key relevance for enlarging the range of applications of the posetic approach, which is currently limited by the complexity of the mutual ranking probability computations.

A different class of computational problems pertains to the *decomposition* of order structures, which is relevant in data analysis, for the identification of the "elementary components" of a statistical phenomenon. For example, in some cases it may well be that the input poset has the structure of a linear sum, where different posets are stacked one on the other, so that the partial order can be seen as a ranking of partially ordered subgroups of units. Usually, this kind of structural analysis is done

by visual inspection, but when the cardinality of the poset increases, some automatic tools become necessary. An interesting problem somehow related to this one is that of designing algorithms for approximating posets through so-called *bucket orders*, i.e. linear sums of antichains. While some algorithms have been proposed in the literature [14], the topic is still largely open.

A further interesting computational problem is related to the development of cluster analysis on partially ordered sets, namely on so-called *lattices*, i.e. posets where each pair of elements admits both *join* (or *sup*) and *meet* (or *inf*). To induce a partial ordering on the clusters, these must be constructed consistently with the input order relation, i.e. the partition generated by the clustering algorithm should be a *congruence* of the input lattice. Computing congruences, i.e. admissible clustering partitions, is not computationally easy, especially for large lattices and developing suitable algorithms is thus a key step, towards ordinal clustering.

The last issue we highlight is more a "hint", than a true open problem and refers to the software languages used for algorithm implementations. Many statistical procedures are implemented in the `R` language that, being an interpreted one, is not particularly fast. To increase the computational speed, numerical routines are often implemented in lower level languages, like *C* or $C++$. The cost for the efficiency of these languages is that programming becomes more difficult and technical. Recently, some new languages have attracted the attention of the scientific community, being at the same time very fast and still of high level. They implement the so-called *functional* paradigm and have the main feature that code can be written using structures that are similar to those of the mathematical language used in the formalization of the problems under investigation, making it easier to design the code and to maintain it. Among the class of functional programming languages, the reference one is the *Haskell* language (*www.haskell.org*), which is being increasingly used, for its elegance and efficiency (and that can also be integrated with `R`). Using Haskell could be a way to provide highly efficient implementations of posetic tools, employing a language with high expressive power and particularly tuned to mathematics.

6 Conclusion

In this short paper, we have outlined the main open issues and the research opportunities, in the development of indicator construction and data analysis on order structures. These structures are abundant in data analysis, although their relevance has been acknowledged only recently, and open many interesting research challenges, both at methodological and computational level. Not to mention the need to apply posetic toolbox in a variety of disciplines and scientific domains, within and outside of the socio-economic field, which calls for measurement, evaluation and prioritization tools that classical aggregative algorithms cannot deliver. Beyond indicator construction, order theory paves the way to the development of what can

be legitimately called "multidimensional ordinal data analysis", in the same spirit as linear algebra provides the foundations for most of multivariate analysis, on numerical data. So, most of the research effort should be devoted to bridging the mathematical theory to the statistical methodology, turning order theoretical results into algorithms for solving data analysis problems (e.g. clustering, dimensionality reduction, indicator construction, but also inferential modeling). Given the combinatorial nature of posetic computations, this leads also to interesting and non-trivial computational problems, which are key to make the posetic approach really effective, in the data analysis practice. All in all, the statistical analysis of partially and quasi-ordered data provides a wide spectrum of research opportunities, making different competencies and attitudes converge into the development of a new branch of data analysis.

References

1. Arcagni A., Barbiano di Belgiojoso E., Fattore M., Rimoldi S. M. L. (2019) "Multidimensional analysis of deprivation and fragility patterns of migrants in Lombardy, using partially ordered sets and self-organizing maps", *Social Indicators Research*, 141(2), 551-579.
2. Bruggemann, R. and Carlsen, L. (2011) "An improved estimation of averaged ranks of partial orders", *MATCH Communications in Mathematical and in Computer Chemistry* 65, 383-414.
3. Bruggemann R., Patil G. P. (2011) *Ranking and Prioritization for Multi-indicator Systems*, Springer.
4. Bubley R. Dyer M. (1999) "Faster random generation of linear extensions", *Discrete mathematics* 201(1-3), 81-88.
5. Davey B. A., Priestley B. H. (2002) *Introduction to Lattices and Order*, CUP.
6. DeLoof K. (2009) "Efficient computation of rank probabilities in posets", *Phd Thesis*, Ghent University.
7. Fattore M., Maggino F., Colombo E. (2012) "From Composite Indicators to Partial Orders: Evaluating Socio-Economic Phenomena Through Ordinal Data", in *Quality of Life in Italy: Researches and Reflections*, Springer.
8. Fattore M. Maggino F. (2014) "Partial Orders in Socio-economics: A Practical Challenge for Poset Theorists or a Cultural Challenge for Social Scientists?", in *Multi-indicator Systems and Modelling in Partial Order*, Springer.
9. Fattore M. (2016) "Partially ordered sets and the measurement of multidimensional ordinal deprivation", *Social Indicators Research*, 128(2), 835-838.
10. *Partial Order Concepts in Applied Sciences*, Fattore M., Brüggemann R. (Eds), Springer 2017.
11. Fattore M. (2017) "Functionals and Synthetic Indicators Over Finite Posets", in Fattore M., Brüggemann R. (Eds) *Partial Order Concepts in Applied Sciences*, Springer.
12. Fattore M., Arcagni A. (2019) "F-FOD: Fuzzy First Order Dominance analysis and populations ranking over ordinal multi-indicator systems", *Social Indicators Research*, 144(1) 1-29.
13. Fattore M., Arcagni A. (2020) "Ranking extraction in ordinal multi-indicator systems", *Book of Short Papers - SIS 2020* - Pearson.
14. Gionis A., Mannila H., Puolamäki K., Ukkonen A. (2006) "Algorithms for discovering bucket orders from data", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 561-566.
15. Korsh J. F., LaFollette, P. S. (2002) "Loopless generation of linear extensions of a poset",*Order*, 19(2) 115–126.

16. Patil G. P., Taillie, C. (2004) "Multiple indicators, partially ordered sets, and linear extensions: Multi-criterion ranking and prioritization", *Environmental and ecological statistics*, 11(2), 199-228.
17. Schröder, B. S. W. (2003) *Ordered sets. An Introduction* Birkäuser
18. Shye S. (1985) *Multiple Scaling: The Theory and Application of Partial Order Scalogram Analysis*, Amsterdam: North-Holland.

Advances in complex sampling strategies

Random forest model-assisted estimation for finite population totals

Stima di totali assistita da modello e basata su random forests

Mehdi Dagdoug, Camelia Goga and David Haziza

Abstract Nowadays, surveys face more and more complex data sets with a large number of variables. These new data raise many challenges and traditional parametric methods of estimation of interest parameters such as totals, ratios or quantiles may prove inefficient. In this work, we propose a new class of model-assisted estimators based on random forests. Under certain regularity conditions on the study variable, the random forest as well as the sampling design, the proposed model-assisted estimator is shown to be asymptotically design unbiased and consistent for the population total. Simulations illustrate that the proposed estimator is efficient and can outperform state-of-the-art estimators, especially in complex and high-dimension settings.

Abstract *Al giorno d'oggi, le indagini interessano insiemi di dati sempre più complessi e caratterizzati da un elevato numero di variabili. Questi dati pongono nuove sfide dal momento che i tradizionali metodi di stima per parametri incogniti della popolazione quali, ad esempio, totali, rapporti e quantili, possono rivelarsi inefficienti. In questo lavoro, si propone una nuova classe di stimatori assistita da modello e basata sul metodo delle random forests. Sotto determinate condizioni di regolarità, si dimostra che lo stimatore proposto è asintoticamente corretto e consistente per il totale della variabile oggetto di studio. Studi di simulazione mostrano, inoltre, che lo stimatore può essere più efficiente degli stimatori presenti in letteratura, specialmente in presenza di relazioni funzionali complesse tra la variabile oggetto di studio e un numero elevato di variabili ausiliarie.*

Dagdoug, M.
Université de Bourgogne Franche-Comté, LMB, 16 route de Gray, 25000 Besançon, FRANCE, e-mail: mohamed_mehdi.dagdoug@univ-fcomte.fr

Goga, C.
Université de Bourgogne Franche-Comté, LMB, 16 route de Gray, 25000 Besançon, FRANCE, e-mail: camelia.goga@univ-fcomte.fr

Haziza, D.
University of Ottawa, Departement of Mathematics and Statistics, Ottawa, CANADA, e-mail: dhaziza@uottawa.ca

Key words: Survey sampling, statistical learning, random forests, model-assisted estimation, variance estimation.

1 Introduction

Nowadays, with the development of digital devices such as smart meters, smartphones which are capable to record and send information at a very fine scale (every minute or every second), it is very common to have very large data sets at hand. Auxiliary information may be used to improve the Horvitz-Thompson estimator of study parameters such as finite population totals by using the well-known model-assisted estimator as described in Särndal et al. (1992). Most model-assisted estimators are based on linear modeling. Recently, nonparametric model-assisted estimators have been suggested: local polynomial (Breidt and Opsomer, 2000), B-splines (Goga, 2005) and penalized B-splines (Goga and Ruiz-Gazen, 2014), penalized splines (Breidt et al., 2005; McConville and Breidt, 2013), generalized additive models (Opsomer et al., 2007), neural nets (Montanari and Ranalli, 2005), nonparametric additive models (Wang and Wang, 2011) and regression trees (Toth and Eltinge, 2011; McConville and Toth, 2019). However, in a high-dimensional framework, traditional parametric or non-parametric model-assisted estimators may fail to provide good estimates. In a classical statistical framework, machine learning methods, such as random forests (Breiman, 2001), are efficient prediction methods in such a high-dimensional framework. Generally speaking, random-forest is an ensemble method that trains a (large) number of trees and combines them to produce more accurate predictions than a single regression tree would.

We suggest in this paper a new class of model-assisted estimators based on random forest estimation methods. The paper is structured as follows: we describe in section 2 the random forest algorithm and we build the new class of model-assisted estimators based on random-forests. Section 3 gives the asymptotic properties of this estimator.

2 Random forest model-assisted estimator of finite population totals

Consider a finite population $U = \{1,...,k,...,N\}$ of size N. We are interested in estimating the population total of a survey variable Y, $t_y = \sum_{k\in U} y_k$. We select a sample S, of size n, according to a sampling design $p(\cdot)$. The first-order and second-order inclusion probabilities are given by $\pi_k = Pr(k \in S)$ and $\pi_{kl} = Pr(k,l \in S)$, respectively.

A basic estimator of t_y is the well-known Horvitz-Thompson estimator given by

$$\widehat{t}_{\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{1}$$

Provided that $\pi_k > 0$ for all $k \in U$, the estimator (1) is design-unbiased for t_y in the sense $\mathbb{E}_p(\widehat{t}_{\pi}) = t_y$. The Horvitz-Thompson estimator makes no use of auxiliary information beyond what is already contained in the construction of π_k.

We assume that a vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^{\top}$ of auxiliary variables is available for all $k \in U$. We also assume that y_k, $k \in U$, are independent realizations from a working model ξ, often referred to as a superpopulation model:

$$\xi: \quad y_k = m(\mathbf{x}_k) + \varepsilon_k, \tag{2}$$

where $m(\cdot)$ is a smooth unknown function and ε_k's, $k \in U$ are independent or zero mean. Suppose that model (2) is fitted at the population-level and let $\widetilde{m}(\mathbf{x}_k)$ be the population-level fit associated with unit k obtained by fitting a parametric or non-parametric procedure. This leads to the pseudo generalized difference estimator of t_y:

$$\widetilde{t}_{pgd} = \sum_{k \in U} \widetilde{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widetilde{m}(\mathbf{x}_k)}{\pi_k}. \tag{3}$$

Most often, the estimator (3) is unfeasible as the population-level fits $\widetilde{m}(\mathbf{x}_k)$ are unknown. Using the sample observations, we fit the working model and obtain the sample-level fits $\widehat{m}(\mathbf{x}_k)$. Replacing $\widetilde{m}(\mathbf{x}_k)$ with $\widehat{m}(\mathbf{x}_k)$ in (3), we obtain the so-called model-assisted estimator of t_y (Särndal et al., 1992):

$$\widehat{t}_{ma} = \sum_{k \in U} \widehat{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k}. \tag{4}$$

Unlike (3), the estimator (4) is no longer design-unbiased, but can be shown to be design-consistent for t_y for a relatively wide class of estimation methods of $m(\cdot)$. The model-assisted estimator (4) is expressed as the sum of the population total of the predictions $\widehat{m}(\mathbf{x}_k)$ and an adjustment term that can be viewed as a protection against model-misspecification.

2.1 Regression trees and random forests

Trees define a class of algorithms that recursively split the p-dimensional predictor space into distinct and non-overlapping regions. In other words, a tree algorithm generates a partition of regions or hyperrectangles of $\mathbb{R}^p$. For an observation belonging to a given region, the prediction is simply obtained by averaging the y-values associated with the units belonging to the same region.

The original classification and regression tree algorithm (CART) of Breiman et al. (1984) searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split) for which the difference in em-

pirical variance in the node before and after splitting is maximized. As a starting point, we consider the hypothetical situation, where y_k and $\mathbf{x}_k$ are observed for all $k \in U$ and assume that the regression tree is fitted at the population level. We use the generic notation A to denote a node with cardinality $\#(A)$ considered for the next split, and $\mathscr{C}_A$ to denote the set of possible splits in the node A, which corresponds to the set of all possible pairs $(j,z) = (\text{variable}, \text{position})$. This splitting process is performed by searching for the best split (j^*, z^*) for which the following empirical CART population criterion is maximized:

$$L_N(j,z) = \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} \left\{ (y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L} \mathbb{1}_{x_{kj} < z} - \bar{y}_{A_R} \mathbb{1}_{x_{kj} \geq z} \right)^2 \right\}, \quad (5)$$

where $A_L = \left\{ k \in A; x_{kj} < z \right\}$, $A_R = \left\{ k \in A; x_{kj} \geqslant z \right\}$ and $\bar{y}_A$ is the average of the y-values of units belonging to A. The best cut is always performed in the middle of two consecutive data points. In practice, it is common to impose a minimal number of observations N_0 (say) in each terminal node. In this case, the splitting process is performed until an additional split generates a terminal node with fewer observations than N_0. The splitting process leads to the partition $\mathscr{P}_U = \left\{ A_j^{(U)} \right\}_{j=1}^{J_U}$ of hyperrectangles of $\mathbb{R}^p$ and the prediction at $\mathbf{x}_k$ is given by:

$$\widetilde{m}_{tree}(\mathbf{x}_k) = \sum_{\ell \in U} \frac{\mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)} y_\ell}{\widetilde{N}(\mathbf{x}_k)}, \quad (6)$$

where $\widetilde{N}(\mathbf{x}_k) = \sum_{\ell \in U} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)}$ denotes the number of units belonging to the terminal node $A^{(U)}(\mathbf{x}_k)$ containing $\mathbf{x}_k$.

While regression trees are easy to interpret and allow the user to visualize the partition (Hastie et al., 2011), they may suffer from a high model variance, hence their qualification of "weak learners". A number of tree-based procedures have been proposed with the aim of improving the predictive performances of regression trees, including pruning (Breiman et al., 1984), Bayesian regression trees (Chipman et al., 1998), gradient boosting (Friedman, 2001) and random forests (Breiman, 2001).

We consider in this work random forests which is a nonparametric estimation method that trains a (large) number, say B, of different trees and combines them to produce more accurate predictions than a single regression tree would. In order to obtain different trees, some amount of randomization is introduced in the tree building process, leading to B different tree-based predictions of $m(\cdot)$.

The original random forest algorithm has been suggested by Breiman (2001), we use in our work a slightly different algorithm as suggested in Biau and Scornet (2016). The random forest algorithm is implemented in two steps. At step 1, we select B data sets without replacement of size N' from the population data set $D_U = \{(\mathbf{x}_k, y_k)\}_{k \in U}$ (called also subsampling step). Next, at step 2, we fit a regression tree on each data set obtained at the previous step. Before each split is performed, m_{try} predictors are selected randomly and without replacement from the full set of p predictors. The m_{try} selected predictors are the split candidates to be considered

for searching the best split in (5). The algorithm stops when each terminal node contains less than a predetermined number of observations. This procedure leads to a set $\widetilde{\mathscr{P}}_U = \left\{ \mathscr{P}_U^{(b)} \right\}_{b=1}^{B}$ of B different partitions of $\mathbb{R}^p$. The randomization used in the tree building process is denoted by the random variable $\theta^{(U)}$, assumed to belong to some measurable space $(\Theta, \mathscr{F})$ and independent of the data (Biau and Scornet, 2016). Let $\theta_b^{(U)}$ be the random variable associated with the bth tree. The random variables $\theta_b^{(U)}, b = 1, \ldots, B$, are assumed to be independent and their distribution is identical to that of the generic random variable $\theta^{(U)}$. The prediction at $\mathbf{x}_k$ obtained by random forest is given by:

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(U)}), \tag{7}$$

where $\widetilde{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(U)})$ is given by (6). However, with survey data, the above estimator is not computable since $\widetilde{m}_{rf}$ is based on partition determined at the population level and asking y_k for all units $k \in U$. In order to cope with this issue, we can build partition by using a variable y^* related to y and known on the whole population or we can build the partition by using the criterion (5) at the sample level $D_n = \{(\mathbf{x}_k, y_k)\}_{k \in S}$ with a stopping criterion asking for minimum n_0 elements in final nodes. In the latter case, we will obtain a sample-based partition denoted by $\widehat{\mathscr{P}}_S = \{\widehat{\mathscr{P}}_S^{(b)}\}_{b=1}^{B}$ and the random forest estimator of m at the sample level is given by

$$\widehat{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}), \tag{8}$$

where

$$\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = \frac{1}{\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})} \sum_{\ell \in S} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)} y_\ell}{\pi_\ell}$$

is the estimated prediction of m at $\mathbf{x}_k$ based on the bth stochastic regression tree and $\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})$ denotes the estimated number of observations in the terminal node $A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)$ containing $\mathbf{x}_k$ in the bth regression tree from the sample based partition $\widehat{\mathscr{P}}_S^{(b)}$ with $\widehat{N}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{\ell \in S} \pi_\ell^{-1} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}$. The variable $\psi_\ell^{(b,S)}$ indicates whether or not unit ℓ has been selected in the bth sub-sample and is such that $\psi_\ell^{(b,S)}$ follows a Bernoulli law $\mathscr{B}(n'/n)$, where n' denotes the number of units in each sub-sample selected at the first step of the random forest algorithm.

The estimator of m given in (8) may be written as a Horvitz-Thompson estimator as follows:

$$\widehat{m}_{rf}(\mathbf{x}_k) = \sum_{\ell \in S} \frac{\widehat{W}_\ell(\mathbf{x}_k) y_\ell}{\pi_\ell}, \tag{9}$$

where

$$\widehat{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}}{\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})}, \quad \ell \in S. \tag{10}$$

2.1.1 Random forest model-assisted estimator of finite population totals

The finite population total t_y is then estimated by the random forest model-assisted estimator obtained by plugging $\widehat{m}_{rf}(\cdot)$ in (4):

$$\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}. \tag{11}$$

The model-assisted estimator $\widehat{t}_{rf}$ given by (11) can be viewed as a bagged estimator:

$$\widehat{t}_{rf} = \frac{1}{B}\sum_{b=1}^{B} \widehat{t}_{tree}^{(b)}(\theta_b^{(S)}),$$

where

$$\widehat{t}_{tree}^{(b)}(\theta_b^{(S)}) = \sum_{k\in U} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) + \sum_{k\in S} \frac{y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})}{\pi_k}$$

is the model-assisted estimator of t_y based on the bth stochastic regression tree.

We may show (Dagdoug et al., 2022) that the random-forest estimator $\widehat{t}_{rf}$ may be writen as a weighted sum of sampled y-values as follows

$$\widehat{t}_{rf} = \sum_{k\in S} w_{ks} y_k,$$

where the weights w_{ks} are given by

$$w'_{ks} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell\in U} \widehat{W}_k(\mathbf{x}_\ell)\left(1 - \frac{I_\ell}{\pi_\ell}\right)\right\}, \quad k \in S, \tag{12}$$

and $I_\ell = 1$ if the unit ℓ is selected in the sample and zero otherwise. The weights w_{ks} satisfy $\sum_{k\in S} w_{ks} = N$ for every sample S (Dagdoug et al., 2022). The weights w_{ks} depend on both the sample selection indicators $I_\ell, \ell \in U$, and the partition $\widehat{\mathscr{P}_S}$ that varies from one sample to another. This is due to the fact that the nodes $A^{(S)}$ are constructed so as to optimize the sample restriction of criterion (5). For this reason, the weights $w_{ks}, k \in S$, are variable specific in the sense that depend on the survey variable Y. To cope with this issue, we suggest in Dagdoug et al. (2022) a model calibration procedure for handling multiple survey variables while producing a single set of weights.

Dagdoug et al. (2022) have shown that the estimator $\widehat{t}_{rf}$ given by (11) holds a nice property related to *out-of-bag* units, the units from the sample S that have not

participated at the prediction $\widehat{m}_{rf}$. More exactly, $\widehat{t}_{rf}$ can be written as follows:

$$\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S} \frac{\left(1-\psi_k^{(b,S)}\right)\left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})\right)}{\pi_k}. \tag{13}$$

The second term on the right hand-side of (13) is equal to the weighted sum of residuals computed for the non-resampled units from the sample S by the random forest algorithm, also called the *out-of-bag* individuals (James et al., 2015), from each of the B trees. This term can then be viewed as a correction term which brings additional information from the units not used in computing the predictions $\widehat{m}_{tree}^{(b)}(\cdot, \theta_b^{(S)}), b = 1, \ldots, B$. The second term on the right hand-side of (13) vanishes if $\psi_k^{(b,S)} = 1$ for all $k \in S$, namely if the random forest algorithm does not involve a resampling mechanism. In this case, the estimator $\widehat{t}_{rf}$ reduces to the so-called projection form:

$$\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k).$$

The estimator $\widehat{t}_{rf}$ reduces to the projection form also if $y_k = c$ for all k, for some $c \in \mathbb{R}$ or if the trees in the forest are fully grown (i.e., each terminal node contains a single observation), which implies that the observations y_k and the corresponding prediction $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ coincide.

3 Asymptotic properties

To establish the asymptotic properties of the proposed estimators and to derive the associated variance estimators, we consider the asymptotic framework of Isaki and Fuller (1982). We start with an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample of size n_v is selected according to a sampling design $Pr(S_v = s_v \mid \mathbf{Z}_U)$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes and the samples sizes go to infinity. It is also supposed that the n_{0v}, the minimal number of units in the terminal nodes, grows to infinity. In order to get the asymptotic properties, we suppose additional assumptions on the sampling design, on the study variable y as well as on the random forest algorithm (Dagdoug et al., 2022).

Result 3.1. *Consider a sequence of random forest model-assisted estimators $\{\widehat{t}_{rf}\}$. Then, there exist positive constants $\tilde{C}_1, \tilde{C}_2$ such that*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf} - t_y\right)\right| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}} + \frac{\tilde{C}_2}{n_{0v}}, \quad \textit{with } \xi\textit{-probability one},$$

where $\mathbb{E}_p$ is the expectation with respect to the sampling design. If $\dfrac{n_v^u}{n_{0v}} = O(1)$ with $1/2 \leqslant u \leqslant 1$, then there exists a positive constant $\tilde{C}$ such that

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf}-t_y\right)\right| \leqslant \frac{\tilde{C}}{\sqrt{n_v}}, \quad \textit{with } \xi\textit{-probability one.}$$

Result 3.1 implies that the random forest model-assisted estimator $\widehat{t}_{rf}$ is asymptotically design-unbiased, i.e., $\lim_{v\to\infty}\mathbb{E}_p[N_v^{-1}(\widehat{t}_{rf}-t_y)]=0$, with ξ-probability one and design-consistent in the sense that $\lim_{v\to\infty}\mathbb{E}_p\left[\mathbf{1}_{\{N_v^{-1}|\widehat{t}_{rf}-t_y|>\eta\}}\right]=0$, with ξ-probability one for all $\eta>0$. Moreover, if n_{0v} is large enough with respect to the sample size n_v, the random forest estimator $\widehat{t}_{rf}$ is $\sqrt{n_v}$-consistent. For a given partition, note that the number of terminal nodes is of order $O(n_v/n_{0v})$, and if n_{0v} satisfies the condition from the Result 3.1, the number of terminal nodes is of order $O(n^{1-u})$ for $1/2\leqslant u\leqslant 1$.

The next result shows that the random forest model-assisted estimator $\widehat{t}_{rf}$ is asymptotically equivalent to the pseudo-generalized difference estimator:

$$\widetilde{t}_{rf}=\sum_{k\in U}\widetilde{m}_{rf}(\mathbf{x}_k)+\sum_{k\in S}\frac{y_k-\widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}, \tag{14}$$

where $\widetilde{m}_{rf}(\mathbf{x}_k)$ is given by (7).

Result 3.2. *Consider a sequence of random forest estimators $\{\widehat{t}_{rf}\}$. Assume also that $\dfrac{n_v^u}{n_{0v}}=O(1)$ with $1/2<u\leqslant 1$. Then, $\{\widehat{t}_{rf}\}$ is asymptotically equivalent to the pseudo-generalized difference estimator $\widetilde{t}_{rf}$ in the sense that*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}-t_y\right)=\frac{\sqrt{n_v}}{N_v}\left(\widetilde{t}_{rf}-t_y\right)+o_{\mathbb{P}}(1).$$

From result 3.2, it follows that the asymptotic variance of $\widehat{t}_{rf}$ can be approximated by the variance of $\widetilde{t}_{rf}$:

$$\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right)=\frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}(\pi_{kl}-\pi_k\pi_\ell)\frac{y_k-\widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widetilde{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{15}$$

The asymptotic variance given in (15) cannot be computed in practice because the residuals, $y_k-\widetilde{m}_{rf}(\mathbf{x}_k)$, $k\in U$, are unknown. Assuming that $\pi_{k\ell}>0$ for all pairs $(k,\ell)\in U_v\times U_v$, a design-consistent estimator of the asymptotic variance is given by:

$$\widehat{\mathbb{V}}_{rf}\left(\frac{1}{N_v}\hat{t}_{rf}\right)=\frac{1}{N_v^2}\sum_{k\in S_v}\sum_{\ell\in S_v}\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k-\widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widehat{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{16}$$

Dagdoug et al. (2022) conducted large simulation studies on simulated as well as on real data considering several nonlinear relationships between study variables and high-dimension auxiliary variables. Simulation results show that the random forest estimator is efficient and can outperform state-of-the-art estimators, especially in complex and high-dimension settings. The variance estimator performance has been also investigated. As we suspected, the minimum number n_0 of observations

in each terminal node, may have an impact on the variance estimator. More exactly, $\widehat{\mathbb{V}}_{rf}\left(\hat{t}_{rf}\right)$ is severely biased for small values of n_0 and as a consequence, the confidence intervals of t_y perform poorly for small values of n_0 because of the substantial underestimation of the true variance in these scenarios. The significant bias for small values of n_0 is most likely due to overfitting, which is characterized by the presence of artificially small residuals $y_k - \widehat{m}_{rf}(\mathbf{x}_k)$ in each terminal node, which in turn, leads to underestimation. To cope with this issue, we suggest in Dagdoug et al. (2022) a variance estimator based on a K-fold criterion which greatly improved the coverage rates.

References

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28:1023–1053.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Model-assisted estimation through random forests in finite population sampling. *Journal of American Statistical Association*, to appear.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *The Canadian Journal of Statistics*, 33:163–180.

Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society: Series B*, 76:113–140.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:49–61.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Regression*, 25:745–763.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46:389–413.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *Journal of the American Statistical Association*, 100:1429–1442.

Opsomer, J. D., Breidt, F. J., Moisen, G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, (478):400–409.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106:1626–1636.

Wang, L. and Wang, S. (2011). Nonparametric additive modelassisted estimation for survey data. *Journal of Multivariate Analysis*, 102:1126–1140.

Design-based Consistency of the Horvitz-Thompson Estimator in Spatial Sampling

Coerenza Basata sul Disegno dello Stimatore di Horvitz-Thompson nel campionamento spaziale

Lorenzo Fattorini

Abstract Spatial populations are usually located on a continuous support. They can be surfaces representing the values of the survey variable at any location, finite collections of units with the corresponding values of the survey variable, or finite collections of areal units partitioning the support, where the value attached is the total amount of an attribute within. We derive conditions on the design sequence ensuring consistency of the Horvitz–Thompson estimator of spatial population totals, supposing minimal requirements on the survey variable. Consistency and its implications in real surveys are discussed with focus on environmental surveys.

Abstract *Le popolazioni spaziali sono di solito collocate in un supporto continuo. Queste popolazioni possono essere costituite da superfici che forniscono il valore della variabile di interesse in ogni punto del supporto, da insiemi finiti di unità con i corrispondenti valori della variabile di interesse, da insiemi finiti di aree che ripartiscono il supporto con i corrispondenti valori dell'ammontare della variabile di interesse al loro interno. Le condizioni che assicurano la coerenza dello stimatore di Horvitz-Thompson dei totali di queste popolazioni sono state derivate supponendo condizioni minimali riguardanti le caratteristiche della variabile di interesse. Le implicazioni della coerenza nelle indagini reali sono state discusse con particolare riferimento alle indagini ambientali.*

Key words: continuous populations, finite populations, Horvitz–Thompson estimator, population totals.

---

[1] Lorenzo Fattorini, University of Siena; email: lorenzo.fattorini@unisi.it

This is a joint paper with Marzia Marcheselli, Caterina Pisani and Luca Pratelli

Introduction

Consistency is an intuitively appealing property ensuring that the distribution of an estimator tends to be concentrated around the parameter as the sample size *n* increases. This definition cannot be immediately carried over finite population setting. Indeed, when a without replacement sampling scheme is adopted to select samples from a population of *N* units, we cannot let *n* approach infinity without further, artificial assumptions. Usually, consistency can be investigated by considering a sequence of increasing, nested populations $\{U_k\}$ each of them characterized by a target parameter θ_k. Then, a sequence $\{d_k\}$ of sampling designs selecting samples of increasing size n_k is introduced and an estimator $\hat{\theta}_k$ of the parameter θ_k is said to be design-consistent if the sequence of random variables $\{\hat{\theta}_k - \theta_k\}$ converges in probability to 0.

This asymptotic framework was probably firstly delineated rigorously by Isaki & Fuller (1982), who, in the spirit of design-based inference, proved consistency of the Horvitz-Thompson (HT) estimator of population means mainly on the basis of the properties of the design sequence $\{d_k\}$, requiring a minimal assumption regarding populations. Indeed, the authors only require that the survey variable is bounded, a feature always satisfied in real surveys. In the spirit of Isaki & Fuller (1982), we aim to give consistency conditions for the HT estimator of totals in spatial populations based on the design sequence under minimal assumptions regarding populations.

We consider three types of spatial populations: (i) continuous populations, constituted by a continuous set of locations on a study area; (ii) finite populations of units scattered over the study area; (iii) finite populations of areal units partitioning the study area. Fattorini et al. (2020) derive consistency conditions for the three types of populations, each of them requiring a different asymptotic scenario. For brevity, here we treat only the first two type of populations. For the results concerning populations of type (iii) as well as for any technical detail here omitted, see Fattorini et al. (2020).

Consistency for continuous populations

Let y be a Borelian and bounded function on A with values on $[0, L]$, where $y(p)$ is the value of the survey variable Y at the location $p \in A$. We aim to estimate the population total $T = \int_A y(p)\lambda(dp)$. Suppose a sequence of designs $\{d_k\}$, each of them selecting an increasing number n_k of points onto A, say $P_{k,1}, \ldots, P_{k,n_k}$. Following Cordy (1993), the designs should be such that the n_k-tuple $[P_{k,1}, \ldots, P_{k,n_k}]$ is a random vector with probability density $g^{(k)}$ with respect to the product measure $\lambda^{\otimes n_k}$. Let $g_i^{(k)}$ be a version of the marginal probability density of $P_{k,i}$ with respect to λ and $g_{ih}^{(k)}$ be a version of the marginal probability density of $[P_{k,i}, P_{k,h}]$ with respect to $\lambda \otimes \lambda$, with $i \neq h = 1, \ldots, n_k$. Moreover, $\pi_k(p) = \sum_{i=1}^{n} g_i^{(k)}(p)$ is the inclusion function and

$\pi_k(p,q) = \sum_{i\neq h=1}^{n} g_{ih}^{(k)}(p,q)$ is the pairwise inclusion function. If $\pi_k(p) > 0$ for each $p \in A$, then the extension of the HT estimator to the continuous case

$$\hat{T}_k = \sum_{i=1}^{n_k} \frac{y(P_{k,i})}{\pi_k(P_{k,i})} \tag{1}$$

is an unbiased estimator of T and, if $\int_A \frac{1}{\pi_k(p)}\lambda(dp) < \infty$,

$$\mathrm{Var}(\hat{T}_k) = \int_A \frac{y^2(p)}{\pi_k(p)}\lambda(dp) + \int_{A^2}\left\{\frac{\pi_k(p,q)}{\pi_k(p)\pi_k(q)} - 1\right\} y(p)y(q)\lambda(dp)\,\lambda(dq) \tag{2}$$

If the design sequence is such that

$$\lim_{k\to\infty} \sup_p \frac{1}{\pi_k(p)} = 0 \tag{3}$$

$$\lim_{k\to\infty} \sup_{p\neq q} \left\{\frac{\pi_k(p,q)}{\pi_k(p)\pi_k(q)} - 1\right\}^{+} = 0 \tag{4}$$

then $\lim_{k\to\infty} \mathrm{Var}(\hat{T}_k) = 0$ and $\hat{T}_k$ converges in probability to T.

The most straightforward scheme to sample spatial location on a continuum is the uniform random sampling (URS), i.e., the random and independent selection of n_k points on the support. Under URS, conditions (3) and (4) are satisfied. But despite simplicity and consistency, URS may lead to uneven surveying. Spatial balance can be achieved using quite complex schemes explicitly tailored for that purpose, such as the generalized random tessellation stratified sampling (Stevens & Olsen, 2004) and the sampling based on space-filling Hilbert curves (Lister & Scott, 2009). More simply, spatial balance can be obtained using tessellation stratified sampling (TSS): the support A is partitioned into n_k spatial subsets of equal extent and a point is randomly and independently located in each subset. Under TSS, conditions (3) and (4) are satisfied. Alternatively, when the support can be tessellated into n_k regular polygons of equal extent, systematic grid sampling (SGS) is widely used to achieve spatial balance. SGS consists of randomly selecting a point in one polygon and systematically repeating it in the others. However, SGS cannot be considered for consistency in the framework introduced by Cordy (1993), because while $\pi_k(p) = n_k/\lambda(A)$ for each $p \in A$, no probability density exists for the pair $[P_{k,i}, P_{k,i+1}]$.

Under URS, TSS and SGS, (1) reduces to the Monte Carlo estimator

$$\hat{T}_k = \frac{\lambda(A)}{n_k}\sum_{i=1}^{n_k} y(P_{k,i}) \tag{5}$$

Therefore, these schemes can be viewed as Monte Carlo integration methods with URS coinciding with crude Monte Carlo integration. Consistency results on Monte Carlo integration have been already exploited. In particular, the superiority of TSS vs URS has been proven and convergence rates of these schemes have been investigated for estimating totals (Barabesi et al., 2012). As to SGS, consistency cannot be proven in the framework introduced by Cordy (1993) while we have proven consistency in the Monte Carlo estimation framework (see Fattorini et al., 2020).

Consistency for finite populations of units

Let $\{U_k\}$ be a nested sequence of populations of units of increasing size N_k scattered throughout A. Moreover, let Y be a survey variable with values on $[0, L]$ and let y_j be the value of Y for the unit $j \in U_k$. We aim to estimate the population total $T_k = \sum_{j \in U_k} y_j$. The population sequence determines a corresponding sequence of totals $\{T_k\}$. Suppose a sequence of designs $\{d_k\}$, each of them selecting a sample S_k from U_k of increasing size n_k, with first and second order inclusion probabilities $\pi_j^{(k)}$ and $\pi_{jh}^{(k)}$ for $h > j \in U_k$. Then the HT estimator

$$\hat{T}_k = \sum_{j \in S_k} \frac{y_j}{\pi_j^{(k)}} \tag{6}$$

is unbiased with variance

$$\mathrm{Var}(\hat{T}_k) = \sum_{j \in U_k} \left(\frac{1}{\pi_j^{(k)}} - 1 \right) y_j^2 + 2 \sum_{h > j \in U_k} \left(\frac{\pi_{jh}^{(k)}}{\pi_j^{(k)} \pi_h^{(k)}} - 1 \right) y_j y_h \tag{7}$$

If the design sequence ensures

$$\lim_{k \to \infty} \max_{h > j} \left\{ \frac{\pi_{jh}^{(k)}}{\pi_j^{(k)} \pi_h^{(k)}} - 1 \right\}^+ = 0 \tag{8}$$

and there exists $\pi_0 > 0$ such that $min_j \, \pi_j^{(k)} \geq \pi_0$, then

$$\lim_{k \to \infty} \mathrm{Var}(\hat{T}_k / T_k) = 0$$

and $\hat{T}_k / T_k$ converges in probability to 1.

If the population list is available, the most straightforward scheme is simple random sampling without replacement (SRSWOR). If a constant fraction $0 < \pi_0 < 1$ of units is selected from each population U_k, then condition (8) is satisfied. Despite simplicity and consistency, SRSWOR may lead to uneven scattering of sampled units throughout the region. Spatial balance can be ensured by using explicitly tailored schemes, such as the generalized random tessellation stratified sampling (Stevens & Olsen, 2004), the draw-by-draw sampling that excludes the selection of contiguous units (Fattorini, 2006), the local pivotal method (Grafström et al., 2012), the spatially correlated Poisson sampling (Grafström, 2012) and the doubly balanced spatial sampling (Grafström and Tillé, 2013). More simply, spatial balance can be achieved by partitioning the support into a fixed number of strata and then selecting the same fraction of units $0 < \pi_0 < 1$ within each stratum by SRSWOR, ensuring consistency within each stratum when strata and sample sizes increase. Then consistency also holds under stratified sampling with proportional allocation.

The knowledge of the list of population units rarely occurs in environmental surveys where, for example, units are trees or shrubs scattered over the study area and the creation of the list involves prohibitive efforts, especially over large areas.

Probably, the unique case in which the list of units becomes available is under 3P sampling, from the acronym of *probability proportional to prediction.* This scheme is a variation of Poisson sampling and is adopted in forest surveys when supports are of moderate sizes. Under 3P sampling, all the units are visited by a crew of experts, a prediction x_j for the value of the survey variable is given by the experts for each unit and units are independently included in the sample with probability x_j/L (Gregoire and Valentine, 2008). A lower bound $l > 0$ for the survey variable *Y* naturally arises in most forest and environmental surveys in which units with *Y* values (e.g., tree height or basal area) smaller than a given threshold are not considered in the population. In this case $\pi_j^{(k)} \geq l/L$ for any $j \in U_k$ and condition (8) holds. Therefore, under 3P sampling $\hat{T}_k/T_k$ converges in probability to 1.

When the list of units is unknown, the sampling schemes usually adopted for selecting units from a population U are based on points (eventually identifying plots or transects) randomly located onto a reference area B. In most cases B constitutes an enlargement of the support A introduced in order to avoid edge effects (see, e.g., Gregoire and Valentine, 2008, section 7.5). For each unit $j \in U$, the scheme univocally defines the inclusion region B_j, i.e. the subset of B onto which the random point p must fall to give rise to the selection of the unit. Because p is randomly selected, the first-order inclusion probability of unit j is given by $\pi_j = \lambda(B_j)/\lambda(B)$. Denote by $S(p) \subset U$ the sample of units selected by means of a random point p onto B. If $\lambda(B_j)$ can be computed for each $j \in S(p)$, then the HT estimator

$$\hat{T} = \sum_{j \in S(p)} \frac{y_j}{\pi_j} = \lambda(B) \sum_{j \in S(p)} \frac{y_j}{\lambda(B_j)} \tag{9}$$

is an unbiased estimator of the population total $T = \sum_{j \in U} y_j$. Since T can be rewritten as

$$T = \int_B g(p) d(\lambda p) \tag{10}$$

where $g(p) = \sum_{j \in U} \frac{y_j}{\lambda(B_j)} I_j(p)$ and $I_j(p)$ is the sample indicator function, that is equal to 1 if $j \in S(p)$ and equal to 0 otherwise, then the HT estimator (9) can be rewritten as

$$\hat{T} = \lambda(B) g(p) \tag{11}$$

i.e., it can be viewed as the Monte Carlo estimator of the integral (10) at p (Gregoire and Valentine, 2008, Chapter 10; Mandallaz, 2008). Practically speaking, when dealing with a without-list population, the total estimation can be rephrased by (5) as the estimation of a total over a continuum, in such a way that consistency can be achieved, as in section 2, as the number of sample points selected onto B increases. Obviously, considerations analogous to those in section 2 regarding URS, TSS and SGS still hold.

Consistency and real surveys

Design-based inference is of large use in environmental surveys, especially in large-scale forest surveys such as national forest inventories (e.g., Tomppo et al., 2010). Usually, the main target is to estimate land use, especially forest cover, that can be expressed as integrals of dichotomous variables, together with totals of some attributes regarding finite populations of objects within some land cover classes (e.g., total volume of trees within forested lands). Therefore, populations of type (i) and (ii) are both involved, and the goal is to estimate their totals by the same survey. That is done by locating points onto the study region in accordance with a sampling scheme, recording the land cover class at each point and selecting samples of objects within plots of pre-fixed size centered at the selected points (Fattorini, 2015). Consistency of the resulting estimators holds under the most widely adopted sampling schemes, such as TSS and SGS. Therefore, as the number of points increases, i.e., when the subset grain is sufficiently small with respect to the size of the study area, thus providing a sufficiently large number of sample points, estimators can be considered concentrated around true parameters. These considerations support the results of some recent surveys such as the Italian National Forest Inventory where about 300,000 points, one per square kilometer, were selected on the Italian territory (Fattorini et al., 2006) and the IUTI survey, a land use survey promoted and carried out in 2008 in Italy, where 1,200,000 points, one each 250 square meters, were selected (Corona et al., 2012).

In most large-scale surveys, a second phase of sampling is performed because it may be demanding to visit and perform estimation for all selected points (e.g., Pagliarella et al., 2016). In this case, estimator (5) is only virtual, and in a second-phase a sub-sample of these points is selected using a finite population sampling scheme. Fattorini et al. (2017) give sufficient conditions for the second-phase designs to ensure consistency of the two-phase estimators when a TSS is performed in the first phase.

As pointed out by Opsomer et al. (2007) in the last years there is a “tremendous” opportunity to exploit auxiliary data derived from remote sensing sources such as photo-interpreted land-cover class, elevation, slope, and Lidar metrics in order to improve the precision by means of calibration strategies performed introducing assisting super-population models. Because the resulting model-assisted estimators, such as regression and ratio estimators, can be invariably expressed as smooth functions of HT estimator of totals (e.g., Särndal et al., 1992, Chapter 6), if consistency holds for the HT estimators it also holds for these model-assisted counterparts.

Conclusions

Consistency of the design-based estimators of totals in spatial populations is pursued under minimal assumptions regarding the population characteristics, by focusing on

the design sequences. Even if there are no population sequences in real surveys but a unique population, however the presumed sequence is inspired by the sampling scheme actually adopted to select the sample from the unique population. Therefore, consistency can be considered somewhat real, not modelled, or assumed.

For continuous populations, it suffices to hold the study area and the *Y* surface as fixed and simply considering a design sequence selecting an increasing number of sample points in the support. A slightly more complex machinery is necessary for finite populations of units scattered onto a support, where the Isaki and Fuller (1982) asymptotic scenario is exploited, taking the support fixed and considering a sequence of nested populations increasing within. However, when the scattered units have no list, as frequently happens in environmental surveys, and hence it is necessary to sample them by points, eventually identifying plots or transects, the population can be held fixed, and consistency can be achieved, as in the continuous case, from the scheme adopted to locate an increasing number of points on the support.

Consistency is important also for estimating more complex parameters than totals. Indeed, if consistent estimators $\hat{T}_1, \ldots, \hat{T}_K$ are available for the totals $T_1, \ldots, T_K$ of K attributes, then for many functions of totals f, the plug-in estimator $f(\hat{T}_1, \ldots, \hat{T}_K)$ is consistent for $f(T_1, \ldots, T_K)$.

Finally, we have proved consistency for some widely applied and naive sampling schemes. Owing to their complexity, we cannot prove consistency for more complex spatially balanced schemes appeared in literature. However, owing to their effectiveness in providing spatial balance and their good performance (Fattorini et al., 2015), we may presume that consistency holds also for these schemes.

References

1. Barabesi, L., Franceschi, S., Marcheselli, M.: Properties of design-based estimation under stratified spatial sampling. Ann. Appl. Stat. **6**, 210–228 (2012)
2. Cordy, C.B.: An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Stat. Probab. Lett. **18**, 353–362 (1993)
3. Corona, P., Barbati, A., Tomao, A., Bertani, R., Valentini, R., Marchetti, M., Fattorini, L., Perugini, L.: Land use inventory as framework for environmental accounting: an application in Italy. iForest **5**, 204–209 (2012)
4. Fattorini, L.: Applying the Horvitz–Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. Biometrika **93**, 269–278, (2006)
5. Fattorini, L.: Design-based methodological advances to support national forest inventories: a review of recent proposals. iForest **8**, 6–11 (2014)
6. Fattorini, L., Marcheselli, M., Pisani, C.: A three-phase sampling strategy for large-scale multiresource forest inventories. J. Agric. Biol. Environ. Stat. **11**, 1–21 (2006)
7. Fattorini, L., Corona, P., Chirici, G., Pagliarella, M.C.: Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use and land cover estimation. Environmetrics **26**, 216–228 (2015)

8. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based asymptotics for two-phase sampling strategies in environmental surveys. Biometrika **104**, 195–205 (2017)
9. Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based consistency of the Horvitz-Thompson estimator under spatial sampling with applications to environmental surveys. Spat. Stat. **35**, 100404 (2020)
10. Grafström, A.: Spatial correlated Poisson sampling. J. Stat. Plan. Inference **142**, 139–147 (2012)
11. Grafström, A., Tillé, Y.: Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics **24**, 120–131 (2013)
12. Grafström, A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (2012)
13. Gregoire, T.G., Valentine, H.T.: Sampling Strategies for Natural Resources and the Environment. Chapman & Hall/CRC, Boca Raton (2008)
14. Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. J. Amer. Statist. Assoc. **77**, 89–96 (1982)
15. Lister, A.J., Scott, C.T.: Use of space-filling curves to select sample locations in natural resource monitoring studies. Environ. Monit. Assess. **149**, 71–80 (2009)
16. Mandallaz, D.: Sampling Techniques for Forest Inventories. Chapman & Hall, Boca Raton (2008)
17. Opsomer, J.D., Breidt, F.G., Moisen, G.G., Kauermann, G. Model-assisted estimation of forest resources with generalized additive models. J. Amer. Statist. Assoc. **102**, 400–416 (2007)
18. Pagliarella, M.C., Sallustio, L., Capobianco, G., Conte, E., Corona, P., Fattorini, L., Marchetti, M.: From one- to two-phase sampling to reduce costs of remote sensing-based estimation of land-cover and land-use proportions and their changes. Remote Sens. Environ. **184**, 410-417 (2016)
19. Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992)
20. Stevens, D.J., Olsen, A.R.: Spatially balanced sampling of natural resources. J. Amer. Statist. Assoc. **99**, 262–278 (2004)
21. Tomppo, L.M., Gschwantner, T., Laurence, M., McRoberts, R.E. National Forest Inventories: Pathways for Common Reporting. Springer, Heidelberg (2010)

The Responsive-Adaptive Survey Design approach for planning the Permanent Census of Population and Housing

L'approccio Responsive-Adaptive Survey Design per progettare il Censimento Permanente della popolazione e delle abitazioni

Claudia De Vitiis, Stefano Falorsi, Alessio Guandalini, Francesca Inglese, Paolo Righi, Marco D. Terribili

Abstract The present paper aims to test the use of the responsive-adaptive design approach for the post-21 Population Census in Italy. The main goal is to optimize the sample size for the list survey in terms of CAWI and CAPI, under budget constraints. Following the approach proposed by van Berkel et al. (2020), the CAPI sampling fractions to be applied in predetermined target groups are obtained through an optimization problem that balances the response rate according to the coefficient of variation of response propensities. The solution is evaluated through a Monte Carlo simulation aiming at assessing the gain in accuracy obtained using an adaptive design in comparison with other naïve solutions that can be applied in this context.

Abstract Il presente lavoro si propone di testare l'uso dell'approccio responsive-adaptive design per il Censimento Permanente della popolazione in Italia. L'obiettivo principale è ottimizzare la dimensione del campione per l'indagine da lista in termini di CAWI e CAPI rispettando dei vincoli di budget. Seguendo l'approccio proposto da van Berkel et al. (2020), le frazioni di interviste CAPI da effettuare in sotto-gruppi di popolazione predeterminati sono ottenute risolvendo un problema di ottimizzazione che bilancia il tasso di risposta tenendo conto del coefficiente di variazione delle propensioni di risposta. I risultati sono valutati attraverso una simulazione Monte Carlo per misurare il guadagno in accuratezza rispetto ad altre soluzioni.

[1] Claudia De Vitiis, ISTAT; email: devitiis@istat.it
Stefano Falorsi, ISTAT; email: stfalors@istat.it
Alessio Guandalini, ISTAT; email: alessio.guandalini@istat.it
Francesca Inglese, ISTAT; email: fringles@istat.it
Paolo Righi, ISTAT; parighi@istat.it
Marco D. Terribili, ISTAT; terribili@istat.it

Key words: responsive-adaptive designs, response propensity, target groups, optimization.

1 Introduction

Starting from October 2018, the Population Census in Italy is based on a combined approach that integrates administrative data and sample surveys. The goal of the Permanent Census of Population and Housing (PCPH) is to produce annual data besides the estimates of hyper-cubes referred to the year 2021, in accordance with the Eurostat regulations on population Census. The PCPH replaces the previous census process, carried out for the 2011 census round, using a cross-sectional complete enumeration of the Italian population carried out once every ten years.

In particular, the 2018-2021 cycle of the PCPH was carried out through two-component sample surveys (area and list) conducted annually. The two components share almost the same sample of municipalities. Self-Representatives (SR) Municipalities (>17,800 inhabitants) are observed each year; the remaining ones, non-SR (NSR), are observed once in 4 years. The two surveys observed every year 2,850 municipalities and 1,500,000 households. At the end of the first cycle, all the Italian municipalities have been surveyed at least once.

The area survey was conducted on a sample of addresses drawn from the Statistical Base Register of Addresses to count and interview (CAPI technique) every resident household that usually lives in the sampled addresses (every year the expected sample size is around 450,000 households). The list survey was conducted every year on a sample of 950,000 resident households drawn from the Population Base Register by means of a sequential survey design (CAWI/CAPI).

Integrating information from the sample surveys and data from administrative sources in an estimation process based on indirect estimators, the permanent census yearly provided data representing the entire population and all the 7,900 Italian municipalities, while reducing costs and response burden. The gathered information played a key role for policymakers, enterprises and institutions in planning programs and projects, identifying the services needed as well as in assessing policy developments.

For the new post-21 cycle of the PCPH, budget cuts are expected with an impact on the household sample size each year. For this reason, the Italian National Statistical Institute (ISTAT) launched a project aimed at proposing and studying more efficient survey designs for PCPH. The present project is framed in this context and aims to test the responsive-adaptive design (RAD) approach (Brick and Tourangeau, 2017; Groves and Heeringa, 2006; Tourangeau et al., 2017; Tourangeau, 2021) for the post-21 PCPH surveys. The main goal is to optimize the sample size for the list survey in terms of CAWI and CAPI under budget constraints trying to preserve a high level of quality of the estimates.

The basic idea is to study, for the next cycle of PCPH, a survey with a larger sample size with respect to the previous one, but cheaper. The idea is to exploit as

much as possible the CAWI responses (about 45% in the previous surveys). In this way, part of the budget can be invested in selecting a sample of CAWI non-respondents and in interviewing them with the CAPI mode. The basic assumption is that the budget is not enough to interview all the non-respondents. The CAPI mode should increase the non-response rate, solve the bias and reduce the variance of the adjusted non-response estimator. The goals are: 1) determining in advance the CAPI no-respondent sample to improve the quality of the estimates with respect to the strategy of selecting a simple random sample of non-respondent households; 2) preserving as much as possible the same level of quality of the estimates already disseminated for the first cycle (2018-2021), paying off the reduction of CAPI sample size by increasing the CAWI sample size.

A further interesting outcome to be evaluated is the accuracy of direct estimates at the municipality level. In fact, the use of RAD can act in increasing also the quality of the domain indirect estimates.

In the next paragraphs, we describe the chosen approach for experimenting with the adaptive survey design (section 2), the definition of the target groups (section 3.1), the optimisation problem (section 3.2), the simulation framework (section 4), the simulation results (section 5) and, finally, conclusions and further development.

2 The RAD approach

The RAD aims to optimize the balance of response for subpopulations to control non-response bias, while, pointing to equalising response rates between domains of interest to control non-response variance.

The most frequent quality indicators used for following the former aim are the (minimum) coefficient of variation of response propensities (CV_ρ) for subpopulations (van Berkel et al., 2020), the (maximum) indicators of response representativeness (Schouten and Shlomo, 2014; Schouten et al., 2011; Shlomo et al., 2012), the balance indicators – minimum distance between the calibration adjusted estimator and the unbiased estimator under full response (Särndal and Lundquist, 2017). The quality indicator used to control non-response variance is the (minimum) variance of a non-response adjusted estimator conditional on the selected sample (Beaumont et al., 2014).

In the present paper, the RAD approach proposed by van Berkel et al. (2020) is applied. The basic idea is to reduce the coefficient of variation of response propensities among the relevant subpopulations or target groups, CV_ρ, for reducing the non-response bias. A crucial role is played by the auxiliary variables used for identifying the target groups with different response propensities. A lower CV_ρ implies a smaller non-response bias on these variables, of course. But, moreover, it implies a smaller non-response bias on survey variables before any weighting adjustment. The magnitude of the non-response bias depends on the correlation between each survey variable and the auxiliary variables. This looks clear from the following formula,

$$|B(\bar{y})| \leq \frac{S_\rho S_y}{\bar{\rho}} = CV_\rho \, S_y, \quad (1)$$

that defines an upper limit for the absolute bias of the mean of a generic y variable given by the product of $CV_\rho (= S_\rho/\bar{\rho})$ and the standard deviation of the y variable, where $\bar{\rho}$ and S_ρ are the average and the standard deviation of the propensity response of the target groups, respectively.

CV_ρ is the quality indicator suggested for taking into account the solution of the optimisation problem. This indicator is estimated on the target groups in the population *N*. In each target group g ($g = 1, \dots, G$), with population size N_g and $n_g = nN_g \,/\, N$ (i.e. proportional allocation), the total response probability is calculated assuming that all people have the same CAWI response probability $p_{cawi,g}$, the same probability $p_{elig,g}$ of being eligible for CAPI follow-up and the same CAPI response probability $p_{capi,g}$. Then, the total response probability in the group g is

$$p_g = p_{cawi,g} + p_{elig,g} f_{capi,g} p_{capi,g}$$

where $f_{capi,g}$ is the CAPI sampling fraction in group g is the unknown quantity to be determined.

Starting from p_g, it is possible to estimate the mean response propensity, the population variance of the response propensities

$$\bar{\rho} = \frac{1}{N} \sum_{g \in G} N_g p_g$$

$$S_\rho^2 = \frac{1}{N} \sum_{g \in G} N_g \left(p_g - \bar{\rho}\right)^2.$$

The CAPI sampling fraction in each group is determined by performing an optimisation problem that aims to minimise CV_ρ under specific constraints. The constraints can be of different types, such as the budget, the theoretical sample size, the respondent burden, the capacity of the data collection in terms of CAPI interviews and interviewers, the number of total respondents, the number of respondents per target groups, the response rate, the ratio between CAWI and CAPI respondents.

3 Experimental phase

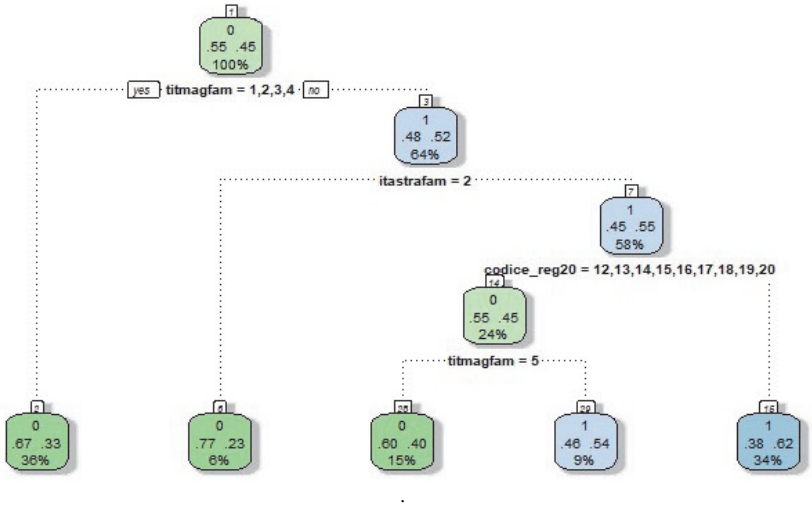
The approach proposed by van Berkel et al. (2020) is, here, implemented on the list component of the 2018 PCPH data enriched by linking additional variables from administrative sources.

3.1 Stratification of the target groups

On the 2018 PCPH surveys data, CAWI non-response using a non-parametric algorithm is studied. In particular, a classification-tree based approach for defining subgroups is used. Considering the CAWI response propensity at the household level, because the response is a household option, target groups are defined.

The auxiliary variables chosen for the CART model are: the highest educational level in the household, citizenship (Italian or not Italian) and region (NUTS 2). To be more specific, citizenship is considered as a binary variable with two modalities (all Italian/at least one foreign member in the household), while the educational level is coded with 8 modalities (Illiterate, Non-illiterate but with no degree, Primary school, Middle school, High school, Bachelor degree, Master degree, Ph.D).

Figure 1: Classification and Regression Tree (CART) for CAWI non-response in the 2018 PCHP



The subgroups that share the same propensity regarding web response behavior (Figure 1) are five:

1. Household for which the highest degree is at most middle school diploma;
2. Household with at least one foreign person and the highest degree is higher than a middle school diploma;
3. Italian household for which the highest degree is a high school diploma, living in Southern Italy (including Islands and Latium region);
4. Italian household for which the highest degree is higher than high school diploma, living in Southern Italy (including Islands and Latium region)
5. Italian household with the highest degree higher than middle school diploma, living in Center-Northern Italy (excluding Latium region).

Because the CAWI response rates are very different at the geographical level (North, Centre and South of Italy), the subgroups already listed are, then, crossed with the geographical areas for defining the target groups.

Table 1, shows the per mode response rates (p_{cawi} and p_{capi}) for the twelve target groups registered in the 2018 PCHP, the weight of each target group in the population ($F_g = N_g/N$) and the eligibility rate (p_{elig}). It is important to point out, that in the present setting the eligible population is calculated considering non-contacts as ineligible for CAPI follow-up.

Table 1:. Weight in the population (F_g), CAWI and CAPI response rates (p_{cawi} and p_{capi}) and elegibility rates (p_{elig}) for the target groups (PCPH list surveys, 2018).

g	Geographical Areal	Subgroup	F_g	p_{cawi}	p_{capi}	p_{elig}
1	North	1	0.155	0.454	0.824	0.974
2		2	0.031	0.373	0.581	0.945
3		5	0.256	0.682	0.800	0.974
4	Center	1	0.069	0.412	0.796	0.972
5		2	0.015	0.377	0.494	0.935
6		3	0.024	0.560	0.707	0.967
7		4	0.014	0.689	0.650	0.955
8		5	0.081	0.631	0.826	0.981
9	South	1	0.140	0.268	0.857	0.977
10		2	0.010	0.275	0.546	0.928
11		3	0.131	0.420	0.855	0.978
12		4	0.073	0.572	0.840	0.971

3.2 The optimisation problem: objective function and constraints

The RAD takes advantage of different CAPI sampling fractions per target group. The CAPI sampling fractions, $f_{capi,g}$ ($g = 1, \dots, 12$), are obtained as the result of the optimisation problem. They are those that minise the CV_p under the budget constraints, that in the present framework are the maximum overall sample size (n) and the overall number of CAPI interviews (n_{capi}) and the minimum number of respondent (r_{tot}). Formally, the optimization problem is defined through the objective function

$$\min_{p_g} \left(\sqrt{\left[\sum_{g=1}^{G} p_g^2 F_g - \left[\sum_{g=1}^{G} p_g F_g \right]^2 \right]} \Big/ \sum_{g=1}^{G} p_g F_g \right) = CV_p$$

while the constraints are

$$\begin{cases} n = n_{cawi} \\ n_{capi} = C \\ r_{tot} \geq R \end{cases}$$

The first constraint implies that, in the first attempt, all the sample is interviewed in CAWI mode. The second and the third constraints can be explicitly written as

$$n_{cawi} \sum_{g=1}^{G} F_g (1 - p_{cawi,g}) f_{capi,g} p_{capi,g} = C$$

$$n_{cawi} \sum_{g=1}^{G} \left[F_g \, p_{cawi,g} + F_g (1 - p_{cawi,g}) f_{capi,g} p_{capi,g} \right] \geq R$$

where $n_{cawi,g} = n \, F_g$ is the sample size in the group g.

Table 2. Comparison between the CAPI sampling fraction (f_{capi}), the CAPI sample size (n_{capi}), the CAPI respondent (r_{capi}) and total response rate (p) for each target group using the optimal CAPI sampling fraction and a constant CAPI sampling fraction, retaining the overall number of respondent.

g	Geographical Areal	Sub group	n	r_{cawi}	f_{capi} constant				f_{capi} optimal			
					f_{capi}	n_{capi}	r_{capi}	p	f_{capi}	n_{capi}	r_{capi}	p
1	North	1	168,141	76,254	0.659	48,638	47,383	0.673	0.689	50,856	49,545	0.756
2		2	74,473	27,772	0.659	16,891	15,959	0.545	1.000	25,624	24,210	0.717
3		5	151,029	102,943	0.659	24,703	24,065	0.806	0.300	11,233	10,943	0.756
4	Center	1	33,946	13,969	0.659	10,183	9,898	0.639	0.757	11,693	11,365	0.756
5		2	16,218	6,108	0.659	3,082	2,882	0.521	1.000	4,675	4,372	0.665
6		3	10,622	5,949	0.659	2,107	2,037	0.711	0.648	2,072	2,004	0.755
7		4	25,542	17,602	0.659	3,249	3,102	0.786	0.415	2,046	1,954	0.769
8		5	142,006	89,628	0.659	27,989	27,456	0.781	0.418	17,736	17,398	0.756
9	South	1	15,584	4,178	0.659	6,300	6,156	0.575	0.796	7,603	7,429	0.756
10		2	79,250	21,764	0.659	19,198	17,824	0.458	1.000	29,123	27,039	0.642
11		3	277,466	116,645	0.659	88,565	86,573	0.663	0.693	93,106	91,012	0.756
12		4	88,062	50,331	0.659	20,292	19,710	0.746	0.528	16,242	15,776	0.756
Total			1,082,340	533,144		271,197	263,046	0.736		272,010	263,046	0.736

Note: n= total sample size, r_{cawi}= CAWI respondent.

The Italian population counts around 60 millions of individuals and 25 millions of households, the sample size for the list component of the PCPH is set to around 1 milion of household ($n = n_{cawi}$ =1,082,340). Furthermore, the overall number of CAPI interviews is set equal to 250,000 ($n_{capi} = C$ =250,000), while the minimum number of respondent should be greater than 650,000 ($r_{tot} \geq R$ =650,000). The optimization problem is solved using the R-package Alabama (Varadhan, 2015).

In Table 2, all the survey process is synthetised. In the first phase, the households are interviewed with the CAWI mode. Since the budget is not sufficient to interview all the CAWI non-respondent with the CAPI mode, just a sample of them can be selected. Following the RAD approach by van Berkel et al. (2020), the CAPI sampling fractions for each target group is determined with the aim of minimazing the variation coefficient of the response rates among the target groups. To have a benchmark, also

a constant CAPI sampling fractions is included in Table 2. The two scenarios provide the same number of respondents, that is the total response rate is equal (p=0.736), but it looks clear that the distribution of the response rate is different among the target groups.

Under the RAD approach, the target groups with a higher CAWI non-response rates have a higher CAPI sampling fraction. Moreover, by definition, the total response rates of the target groups are balanced and close to one another. In fact, the $\boldsymbol{CV_p}$ is equal to 0.00114 under the optimal solution, while is equal to 0.012471 when a constant CAPI sampling fraction is considered. Then, under the RAD approach, a lower non-response bias can be expected.

4 Simulation: selection of replicated samples and evaluation

In this section, the solution obtained applying the RAD approach is compared with other naïve solutions.

Table 3. The upper limit for the absolute bias, $|B(\bar{y})|$, for the unemployment rate. Minimum (min), first quartile (Q1), median (Me), mean (μ), third quartile (Q3) and maximum (MAX) in different domains when a constant or an optimal CAPI sampling fraction (f_{capi}) is considered.

Domain	f_{capi} constant						f_{capi} optimal					
	min	Q1	Me	μ	Q3	MAX	min	Q1	Me	μ	Q3	MAX
Italy	0.020	0.020	0.020	0.020	0.020	0.020	0.008	0.008	0.008	0.008	0.008	0.008
Geographical Areas	0.015	0.015	0.015	0.016	0.017	0.018	0.005	0.007	0.009	0.008	0.010	0.010
Regions	0.013	0.015	0.015	0.016	0.017	0.019	0.004	0.005	0.008	0.007	0.010	0.011
Provinces	0.013	0.015	0.016	0.016	0.017	0.020	0.003	0.005	0.007	0.007	0.010	0.012
Metropolitan cities	0.013	0.015	0.016	0.016	0.017	0.020	0.003	0.005	0.007	0.007	0.010	0.012
Sex	0.014	0.017	0.019	0.019	0.021	0.024	0.005	0.006	0.007	0.007	0.008	0.009
Target groups	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The 2018 PCHP theoretical sample list, which includes respondents and not respondent units for which an archive auxiliary variable on occupational status was available, has been considered as the population universe. Then, the reference population counts 826,979 households and 1,924,906 individuals. The survey variables are derived from the employment status in the administrative sources.

Following expression (1) and knowing the y variables on all the population units, since we are in a simulation context, it is possible to derive the upper limit for the absolute bias, $|B(\bar{y})|$. In Table 3, the upper limits for the absolute bias for the unemployment rate for different domains are shown. In particular, it is possible to see that, on average, the upper limit of the absolute bias is at least 2 times lower under the

RAD approach than under the case in which a constant CAPI sampling fraction is used.

A Monte Carlo simulation (considering R=500 replications) is also performed for comparing the RAD approach with other scenarios. The aim of the simulation is to shed light on the sampling errors of the obtained estimates too, besides their non-response bias.

Without loss of generality, for handling the simulation, not the real sample size of the PCPH (1,082,340 households), but "just" 10,082 households have been selected. This does not impact on the optimal CAPI sampling fractions (f_{capi} optimal) that remain the same as in Table 2.

Four are the scenarios considered:

i. constant CAPI sampling fraction [cost],
ii. optimal CAPI sampling fraction [opt],
iii. constant CAPI sampling fraction within province and SR/NSR municipalities [ng],
iv. only CAWI interviews [cawi].

Under scenarios i., ii. and iii., at the beginning, for each iteration a stratified one-stage sample is drawn from the 2018 PCHP theoretical sample list. This sample coincides with the CAWI component of the survey. The strata are the target groups crossed with the provinces and SR/NSR municipalities (SR=1,0) in the original PCHP sample design. Then, respondents are identified by applying the observed CAWI response rates (Table 1 - p_{cawi}).

Following scenarios i. and ii, CAWI non-respondents and eligibles households are stratified using the same strata, that is considering the target groups. However, just a fraction of them is contacted for the CAPI interviews based on the constant or optimal CAPI sampling fractions respectively (Table 2 - f_{capi} constant and f_{capi} optimal). Finally, the respondents are identified by applying, in both cases, the CAPI response rates (Table 1 - p_{capi}).

Under scenario iii., the CAWI non-respondents and eligibles households are stratified by provinces and SR/NSR municipalities in the original PCHP sample design. That is, in this case, the target group are non considered. Then, the constant CAPI sampling fractions applied in each stratum (Table 2 - f_{capi} constant).

Finally, under scenario iv., the survey is composed of only CAWI interviews. Then, for equalizing the comparisons, that is for obtaining the same number of respondents for the other scenarios, a larger sample size is considered. It is important to point out that the simulation has been set to have the same number of respondents under all the scenarios (796,190 households).

Furthermore, on the sample derived following the four scenarios two estimators have been implemented:

a. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) with an adjustment for making the number of individuals and household consistents at the regional level [ht].
b. Calibration estimator (Deville and Särndal, 1992; Särndal, 2007; Devaud Tillé 2017) in which the sampling weights are made consistent with the

regional distribution of the individuals by sex and 5 age classes (0-14, 15-34, 35-64, 65-74, 75 and more) and the number of households [cal].

Estimates on different parameters[1] related to the individuals belonging to the households are computed at a different domain levels (National – NUTS0, Geographical areas – NUTS1, Regional – NUTS2, Provinces – NUTS3, Metropolitan cities – MC). Considering the values obtained on the overall PCHP theoretical sample list as a benchmark (true value, θ), the estimates obtained in each replication, $\hat{\theta}_i$ ($i = 1, \dots, 500$) are used for evaluated the four scenarios that use the two estimators in terms of:

- relative bias, $Rbias(\hat{\theta}) = \left(\frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i - \theta\right)/\theta$;
- coefficient of variation, $cv(\hat{\theta}) = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}\Big/\bar{\hat{\theta}}$, with $\bar{\hat{\theta}} = \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i$;

For the sake of brevity, just the most interesting results and just those related to the estimates of the unemployment rate are here shown. However, the results and the conclusions that can be inferred looking at the other parameters are similar.

In Figure 2 and Figure 2, the relative bias and the coefficient of variation of the unemployment rate under the different scenarios and using the two estimators are compared. The optimal CAPI sampling fraction provides always less biased estimates, at least for domain above the NUTS1 level (i.e. the level at wich the target groups are created). Under the NUTS2 levels, the relative bias is similar to the case in which the constant CAPI sampling fraction is used. Instead, with respect to the other two scenario is always more convenient. When using the calibration estimator, the relative bias is mitigated and the values are closer among the four scenarios.

In term of coefficient of variation, the optimal CAPI sampling fraction it is a little bit higher due to its more variability in sampling weights, but this is not so serious. Once again, calibration works for mitigating the differences among the scenarios.

[1] Number of employees, number of unemployed, number of people in labour force, proportion of employees, proportion of unemployed, proportion of people in labour force, employment rate and unemployment rate.

Figure 2: Relative bias (***Rbias***) for the estimate of the unemployment rate at NUTS0, NUTS1, NUTS2, NUTS3 and MC level, under the different scenarios and with different estimators.

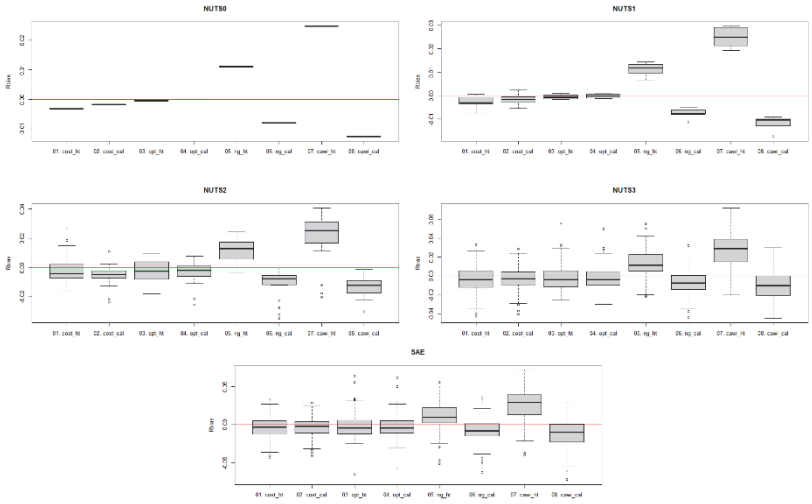
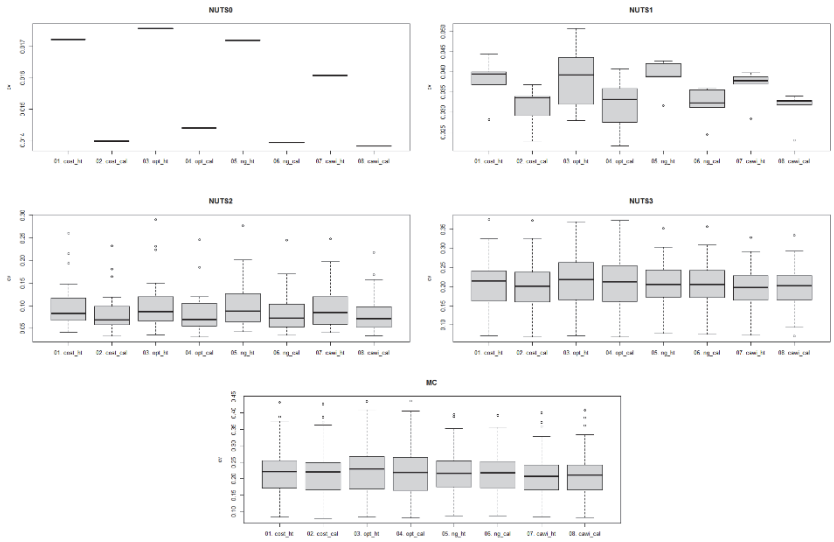


Figure 3: Normalized normalized root-mean-squared error (***NRMSE***) for the estimate of the unemployment rate at NUTS0, NUTS1, NUTS2, NUTS3 and MC level, under the different scenarios and with different estimators.



Note: cost, constant CAPI sampling fraction; opt, optimal CAPI sampling fraction; ng, constant CAPI sampling fraction within the province and Self-Representative and Non-Self-Representative municipalities; cawi, only CAWI interviews. ht, Horvitz-Thompson estimator; cal, calibration estimator.

5 Conclusions and further developments

This work aims to test the use of the responsive-adaptive design approach for the post-21 Population Census in Italy.

The main goal is to optimize the sample size for the list survey in terms of CAWI and CAPI, under budget constraints. The RAD approach proposed by van Berkel et al. (2020) seems to provide promising results in this context.

However, further studies are needed to set this method on a more complex and close to the real case framework in which two-stage of selection (municipalities and households) and a minimum number of CAPI interviews have to be assigned to each sample municipality are considered.

References

1. Beaumont, J.F., Bocci, C., Haziza, D.: An adaptive data collection procedure for call prioritization. J. Off. Stat. **30**, 607--621 (2014)..
2. Brick, J.M., Tourangeau, R.: Responsive Survey Designs for Reducing Nonresponse Bias. J. Off. Stat. **33**, 735--752 (2017).
3. Devaud, D., Tillé, Y.: Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. Test **28**, 1033-1065 (2019).
4. Deville, J. C., Särndal, C. E.: Calibration estimators in survey sampling. J. Amer Stat. Ass. **87**, 376-382 (1992).
5. Groves, R.M., Heeringa, S.G.: Responsive design for household surveys: tools for actively controlling survey errors and costs. J. R. Stat. Soc. Ser. A Stat. Soc. **169**, 439--457 (2006).
6. Horvitz, D. G., Thompson, D. J.: A generalization of sampling without replacement from a finite universe. J. Amer Stat. Ass. **47**, 663-685 (1952).
7. Särndal, C. E.: The calibration approach in survey theory and practice. Surv. Methodol. **33**, 99-119 (2007)..
8. Särndal, C.E., Lundquist, P.: Inconsistent Regression and Nonresponse Bias: Exploring Their Relationship as a Function of Response Imbalance. J. Off. Stat. **33**, 709--734 (2017) http://dx.doi.org/10.1515/JOS-2017-0033.
9. Schouten, B., Shlomo, N.: Selecting Adaptive Survey Design Strata with Partial R-indicators. Technical Report (2014).
10. Schouten, B., Shlomo, N., Skinner C.J.:. Indicators for Monitoring and Improving Representativeness of Response. J. Off. Stat.. **27**, 231--253 (2011).
11. Shlomo, N, Skinner, C.J., Schouten, B.: Estimation of an Indicator of the Representativeness of Survey Response. J. Stat. Plan. Inference, **142**, 201-211 (2012).
12. Tourangeau, R., Brick, J.M., Lohr, S., Li, J.: Adaptive and responsive survey designs: a review and assessment. J. R. Stat. Soc. Ser. A Stat. Soc. **180**, 203-223 (2017).
13. Tourangeau, R.: Science and survey management. Surv. Methodol. **47**, 3--28 (2021).
14. van Berkel, K., van der Doef, S., Schouten, B.: Implementing Adaptive Survey Design with an Application to the Dutch Health Survey. J. Off. Stat. **36**, 609--629 (2020).
15. Varadhan R.: Alabama: Constrained Nonlinear Optimization. R package version 2015. 3-1 (2015) https://CRAN.R-project.org/package=alabama.

Socio-demographic aspects of aging in Italy

Socioeconomic and spatial stratification of frailty in the older population

Stratificazione socioeconomica e spaziale della fragilità nella popolazione anziana

Margherita Silan

Abstract To measure the frailty level of old individuals and identify elderly with peculiar health care needs, the frailty indicator has been proposed. This indicator presents a simple structure that counts only eight variables; in this way, it is easy to replicate and implement. The indicator is based on administrative healthcare data that are available to the entire population. It is useful to predict the seven negative health outcomes related to the frailty condition, following the definition of a frail subject as susceptible to negative outcomes. Moreover, the indicator is useful to stratify the population on the basis of care needs and captures also some socioeconomic dimensions of frailty, although only health variables are used for its construction.

Abstract *Al fine di misurare il livello di fragilità degli individui anziani e identificare gli anziani con particolari bisogni di assistenza sanitaria, è stato proposto l'indicatore di fragilità. Questo indicatore presenta una struttura semplice che conta solo otto variabili, così da essere facile da replicare e implementare. L'indicatore si basa su dati sanitari amministrativi disponibili per tutta la popolazione. È utile per prevedere i sette esiti di salute negativi legati alla fragilità, seguendo la definizione di soggetto fragile come maggiormente esposto a esiti negativi. L'indicatore è inoltre utile per stratificare la popolazione sulla base dei bisogni assistenziali e cattura anche alcune dimensioni socioeconomiche della fragilità, pur utilizzando solo variabili sanitarie nella sua costruzione.*

Key words: Frailty indicator, administrative healthcare data, poset theory, ageing, multiple outcomes, deprivation index.

[1] Margherita Silan, Università di Padova; email: silan@stat.unipd.it

1 Introduction

Given the progressive ageing of the European population, the healthcare system faces new challenges in the management of healthcare resources. In Italy, in 2016, the National Plan for Chronicity was created [10]. It aims to improve the quality of life of patients, especially those suffering from chronic diseases, by making health services more effective and efficient in terms of prevention and care. Among the first objectives of the plan is the stratification of the population through models that take into account the clinical risks and the health and socioeconomic needs of the patients. In this panorama, particular attention should be paid to frail people, who present a more complex health situation, with several concomitant comorbidities and consequently special care needs. However, the National Plan for Chronicity not only considers frailty in health terms, but also social frailty.

Despite the growing interest in the identification of frail individuals, frailty is defined as a syndrome in desperate need of description and analysis [8]. However, there are two fundamental aspects about it that are shared by most of the literature on this topic: frailty as a complex and multidimensional condition, involving multiple functional domains; and frailty as a state of susceptibility to adverse health outcomes, such as death or urgent hospitalization [7, 9].

Therefore, identifying old frail individuals is not a trivial task, but it is certainly indispensable for the implementation of preventive policies that preserve the health conditions in an efficient way, without wasting resources in preventable emergency interventions.

In this panorama there is room for the frailty indicator [12] described below, which proves useful to stratify the population on the basis of care needs [2], useful to predict negative outcomes related to frailty [12] and to capture, although using only health variables for its construction, also some socioeconomic dimensions of frailty. It is an indicator based on information coming from administrative healthcare data flows collected in the Health Unit 6 (whose territory comprises 101 municipalities in the Padua province) during 2016 and 2017.

Consistently with the definition of a frail individual as one who is more likely to experience negative health outcomes, people with a high value of the frailty indicator have a higher risk of experiencing negative health outcomes during 2018. The variables used for the creation of the indicator are aggregated by exploiting partially ordered set theory (poset), which allows variables to be aggregated without the need for assumptions, simply by exploiting the ordinal information present in the source dataset. However, the result is a highly population-specific indicator that should be evaluated and observed also in different populations, for example, in a different time lag. In this work, the behaviour and performance of the frailty indicator are evaluated

in a different period, with data collected in 2017 and 2018, observing its ability to predict outcomes that occurred in 2019 and to catch a glimpse of socioeconomic dimensions of frailty.

2 Data and Population

In the literature, most of the works focused on building a frailty indicator using data collected through self-administered questionnaires. However, having a measure of individual frailty level on a sample of the total population is not particularly useful from a policy implementation perspective. The frailty indicator presented in this work, exploiting administrative healthcare data, quantifies the level of individual frailty for the entire population. Thanks to an agreement with Health Unit 6, we were able to use administrative data regarding their assisted population.

The analysed population is made up of residents assisted by the Health Unit 6, which serves citizens in the province of Padua. In particular, two cohorts were identified through the Regional Health Registry: the first consisting of 215,346 subjects residing on 1 January 2018 at least since 1 January 2016, with at least 65 years of age in 2018, used to build and define the composite indicator; and the second consisting of 218,043 older people with at least 65 years of age in 2019 and residing at least since 1 January 2017, analyzed in this work to evaluate the behavior of the frailty indicator in a different time lag.

Seven administrative healthcare data sources were used (from 2016 to 2019): the regional health registry, which includes the death registry, necessary to identify the reference population; hospital discharge records, containing information on the type and duration of admissions and up to six diagnoses; emergency room (ER) admissions, with information on triage and diagnoses; territorial psychiatry, with information on the type of service required and diagnoses; integrated home care, with information on the number and duration of interventions used by users; ticket exemptions, with information on the pathology or economic situation benefiting from the exemption, on the date of the request and on the duration of the exemption; and territorial pharmaceuticals, with information on prescribed drugs. These sources collect different types of events suffered by the cohorts of patients. Using a deterministic record linkage, information from all sources under analysis was coded, combined, and linked to the studied populations.

3 Methods

Methodological steps to build the frailty indicator start from a definition of frailty that finds wide support among scholars and defines frail individuals as individuals with increased susceptibility to adverse health outcomes [9] (Figure 1).

Starting from elements mentioned in the definition, the following step consists of identifying, according to literature and experts, a list of negative outcomes related to frailty condition: death, urgent unplanned hospitalization, access to the emergency room (ER) with red code, avoidable hospitalization, hip fracture, dementia, and disability. The seven outcomes do not enter directly into the calculation of the indicator, but are used to select the explanatory variables that will compose it.

As a third step, risk factors for the selected outcomes are listed, according to literature. Among those, 72 variables were constructed with administrative healthcare data. The variables concern sociodemographic characteristics, chronic illnesses, use of medication, mental and physical state, and the patient's hospital history (number of visits to the emergency department, number of hospitalizations, etc.). Variables with a very low prevalence (less than 1%) and those that are not associated with outcomes in terms of odds ratio are discarded, reducing the set of variables to 55.

Since the indicator we need to build will be composed of few but relevant variables, in order to be simple, replicable, and parsimonious, the fourth step involves a selection of variables. Since the subset of selected variables must be predictive for the seven outcomes at the same time, the variable selection algorithm repeats the following stages 100 times for every outcome.

- Sample 75% of the total population.
- Balance the sample (50% cases, 50% controls).
- Estimate a logistic regression model with a stepwise variable selection criterion.
- Save the presence and the order of entrance of variables in the model.

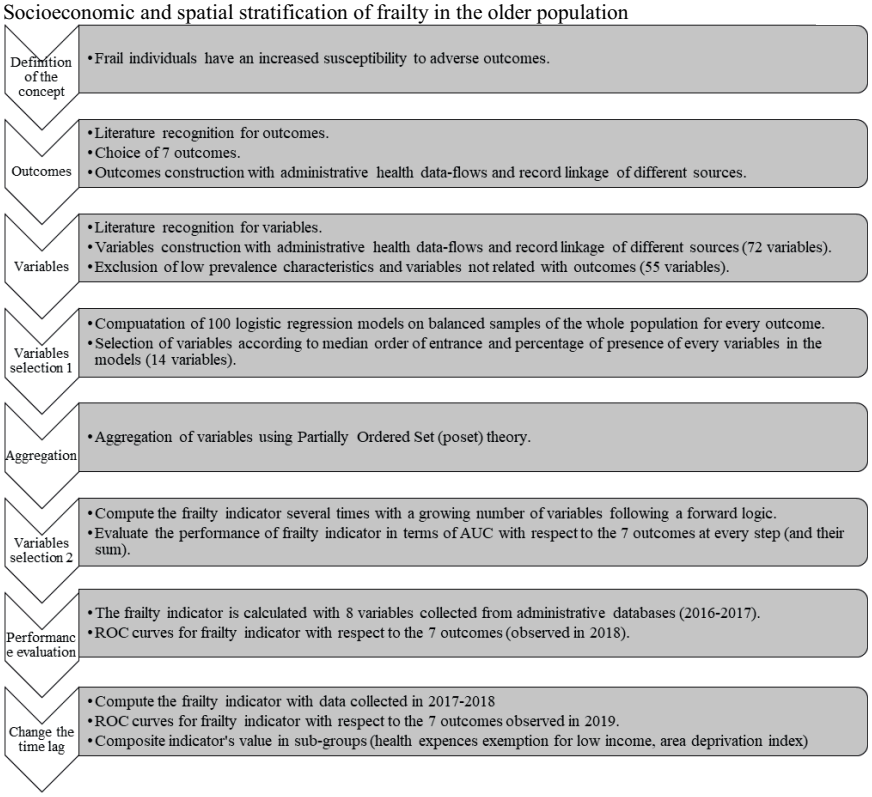


Figure 1: Steps for the construction of the frailty indicator.

After this procedure, it is possible to compute two measures that will guide the variable selection: the percentage of presence and the median order of entrance of every variable for every outcome, among all models. Using these two parameters, we select variables with percentage of presence > 60% for at least 3 outcomes and median order of entrance <20 for at least 3 outcomes.

The result is a set composed of the following 14 variables: age, disability, number of accesses to the emergency room with yellow code, number of accesses to the emergency room with green code, cancer, diabetes, renal failure, Parkinson's disease, blood diseases, mental diseases, polyprescriptions (number of different drugs prescribed in a year), drugs for metabolic and gastrointestinal problems, Charlson Index, diseases of the nervous system.

The indicator is constructed by aggregating the variables exploiting poset theory. This method allows us to summarise information that comes from both dichotomous (disability and other conditions) and ordinal variables (age, number of emergency room admissions, polyprescriptions, Charlson index) without requiring the introduction of subjective components. Thanks to poset theory, it is possible to order the subjects of a population on the basis of a set of ordinal variables. However, if the number of individuals and variables is high, the exact calculation of the average rank (AR) is often impossible and, for this reason, approximations have been introduced

[6]. Specifically, the Mutual Probabilities approach will be used for the calculation of the indicator. It provides better results in terms of accuracy than other methods [6], its implementation in R is available [4], and it is able to handle even large data sets [1, 5, 3].

The set of 14 variables identified by logistic regression is still quite large for the Poset application, whose performance does not always improve with the addition of more variables as the entropy increases with the number of incomparable profiles. Thus, a second step of selection of variables is performed, this time with forward logic. In order to choose the best set of variables that will compose the indicator, we consider the sum of the area under the ROC curves (AUCs) for all the outcomes, i.e., at each step, we add to the indicator the variable that most improves the sum of the seven AUCs.

The starting indicator is constructed with only two variables: The pair (age and polyprescriptions) with the best prediction performance for the seven outcomes is chosen. The third variable is selected from the 12 remaining variables using the same criterion: all possible indicators of three components are constructed, two of which are age and polyprescriptions. The set of three variables that guarantees the highest sum of AUCs is then chosen.

This procedure continues until an improvement is observed in the sum of the AUCs. In this case, it ends after adding the eighth variable, because regardless of which variable is added as the ninth, the sum of the AUCs is always lower than the one obtained with eight variables.

Therefore, the final indicator is made up of the following 8 variables: age, polyprescriptions, number of accesses to the emergency department with yellow code, renal insufficiency, mental illness, Charlson index, disability, and Parkinson's disease.

4 Performance of the Frailty indicator in time

Since the choice of the variables is based on the ability of the frailty indicator to predict the set of seven outcomes selected as negative events related to the frailty condition, the results shown in the first column of Table 1 are expected. Indeed, the frailty indicator predicts well the outcomes observed in 2018, producing fine AUCs, particularly high for death (0.838), dementia (0.832), and access to ER with red code (0.811).

To assess the robustness and validity of the frailty indicator even in a different population, it was calculated using the same set of selected variables with data collected in 2017 and 2018. Thanks to the aggregation method that does not require particular assumptions, it is possible to replicate the indicator only assuming that the eight variables remain the most important in order to predict the outcomes related with the frailty condition. Once the frailty indicator for 2017-2018 is calculated, it is possible to observe its ability to predict negative outcomes observed in 2019. The performance of the indicator in this new period is very good, in some cases even

better than in previous year (such as for disability, urgent hospitalization and access to the ER with red code), also showing a higher sum of AUCs (second column in table 1).

This is an important point of strength for the frailty indicator; indeed, it means that it is possible to just reproduce it in different populations, without repeating the variable selection process and thus without observing the health outcomes in the year that follows those where the data are collected.

Table 1: Area under the ROC curve for the frailty indicator 2016-2017 (predicting outcomes in 2018) and for the frailty indicator 2017-2018 (predicting outcomes in 2019).

Outcome	***Indicator 2016-2017***	***Indicator 2017-2018***
Death	0.838 (0.833-0.842)	0.837 (0.832-0.841)
Disability	0.636 (0.630-0.641)	0.670 (0.665-0.676)
Urgent Hospitalization	0.673 (0.669-0.676)	0.676 (0.673-0.679)
Access to ER with red code	0.811 (0.803-0.819)	0.825 (0.817-0.833)
Dementia	0.832 (0.826-0.839)	0.812 (0.802-0.821)
Fracture	0.767 (0.756-0.778)	0.750 (0.737-0.761)
Avoidable Hospitalization	0.789 (0.785-0.794)	0.785 (0.781-0.789)
Sum	5.346	5.354

5 Socioeconomic and Spatial stratification

The frailty indicator computed in this way is able to stratify the population of old assisted by the Health Unit 6 according to their health status. In fact, people with chronic conditions present on average higher values of the frailty indicator, as represented in Figure 2. In some cases, the interquartile ranges of sick and health subjects are quite distant and well separated, even if the condition is not included in the computation of the frailty indicator. In other words, thanks to careful variables selection and to the poset aggregation approach, the frailty indicator is able to represent health characteristics of individuals, even if they are not directly included in the computation of the composite indicator.

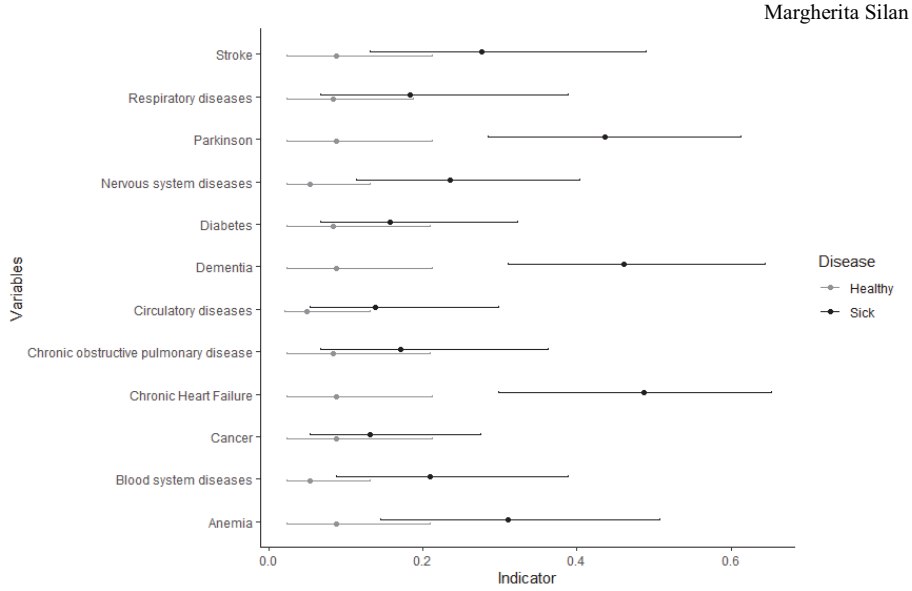


Figure 2: Frailty indicator by chronic conditions. Median and interquartile range.

The frailty indicator is composed only of variables collected in healthcare dataflows; thus, it includes directly variables referred to health aspects of the population, and use of health services. However, as it is able to represent also chronic conditions that are not included in the computation of the indicator, it assumes higher values for those who also present some economic distress.

Socioeconomic variables are not available in healthcare administrative data, so, to find a plausible representation for socioeconomic aspects, an additional effort was needed. First, the presence of health expenses exemption for low income was considered. In Figure 3, the empirical cumulative distribution function shows that people who obtain health expenses exemption for low income also present higher values of the frailty indicator (having the curve lower and closer to the right side).

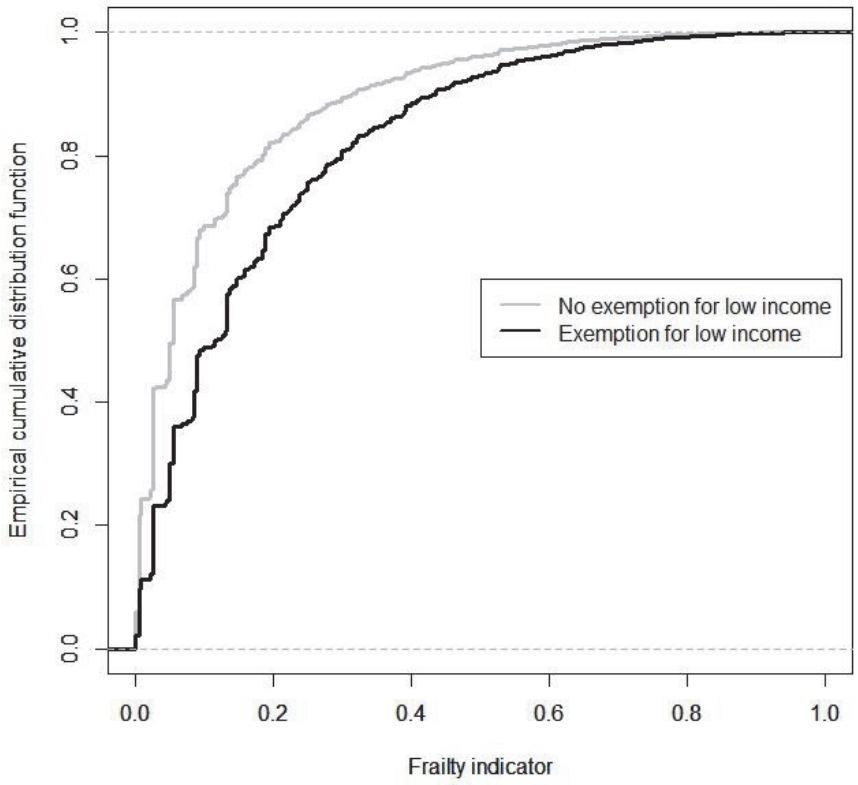


Figure 3: Frailty indicator by health expenses exemption for low income.

The second socioeconomic variable that we considered is the Caranci deprivation index at the census block level [11]. In order to link this index to individuals, it was necessary to geo-reference all the addresses of the assisted old population. The deprivation index uses the General Census of Population and Housing, and it includes five variables that are available at census block level: low level of education, unemployment, non-home ownership, one parent family, and overcrowding. The index is often represented as a categorical variable by quintiles of population. In figure 4 are represented the means (and 95% confidence intervals) of the frailty indicator for the five quintiles of the deprivation index based on Census 2011. Even if the index is at census block level and referred to several years before the time period of the frailty indicator, the relationship between frailty and deprivation is quite clear in Figure 4. Increasing the deprivation of the census block where old individuals live, increases also the mean of the frailty indicator showing differences that are statistically relevant.

From the observation of Figures 3 and 4, it is reasonable to assume that the frailty indicator catches more than just the health dimension of frailty, but also a glimpse of its socio-economic traits.

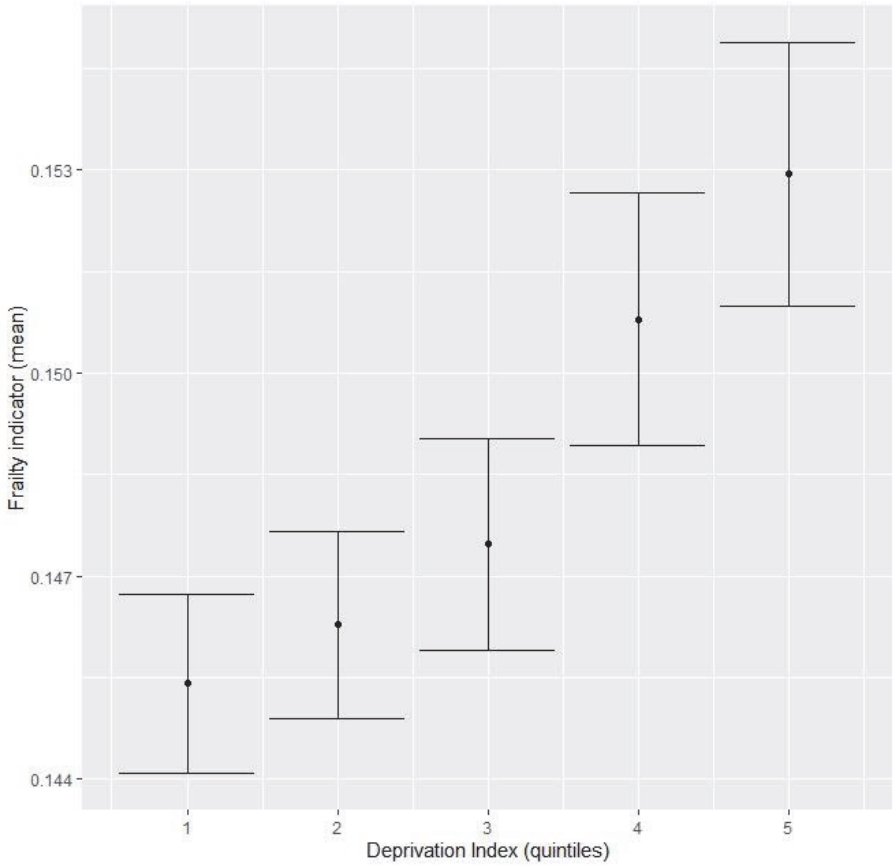


Figure 3: Frailty indicator (means and 95% confidence intervals) by Caranci deprivation index (based on 2011 Census).

6 Conclusion

The frailty indicator is able to identify frail elderly people and classify them from the most frail to the least frail. Individuals with high levels of the frailty indicator have worse health conditions and a higher risk of having negative outcomes related to the frailty condition, reflecting the definition of a frail subject as an individual more susceptible to adverse outcomes.

Moreover, it has a simple structure, it only needs 8 variables, easily retrievable from administrative healthcare dataflows. Thus, it is easily replicable if administrative healthcare data are available and it is possible to compute it for the whole population of assisted individuals.

Thanks to the use of poset approach, it is possible to rebuild the frailty indicator for different populations or periods with the only assumption that the eight selected variables remain suitable to depict frailty condition. In this work, the regeneration of the frailty indicator for a different time lag was extremely successful.

The frailty indicator represents the health condition of old individuals assisted by Health Unit 6, but it is also clearly related to the socioeconomic condition as well, according to the variables that it was possible to connect. Unfortunately, this relationship between socioeconomic deprivation and frailty is not easy to deepen using only administrative healthcare data, because of the lack of socio-economic variables.

To make the frailty indicator accessible also to nonstatistical users, the next step will be to implement a user-friendly application that simplifies and guides the computation of the indicator. However, the possibility of extending the use of this frailty indicator is conditioned by the homogeneity and sharing of methods for collecting and coding clinical information on the population among all regional health systems.

Further steps will deepen behaviour of the frailty indicator in different populations and subgroups of population to validate this promising instrument for health services.

References

1. Boccuzzo, G., Caperna, G.: Evaluation of life satisfaction in Italy: Proposal of a synthetic measure based on poset theory. In: F. Maggino (ed.) Complexity in society: From indicators construction to their synthesis, pp. 291–321. Springer (2017) doi:10.1007/978-3-319-60595 -1_12
2. Boccuzzo, G., Gargiulo, L., Iannucci, L., Silan, M., Costa, G.: La salute degli anziani tra prospettive di resilienza e fragilità. In: Billari, F. C., Tomassini, C. (eds.) Rapporto sulla popolazione. L'Italia e le sfide della demografia, pp. 213-237. Il Mulino, Bologna (2021)
3. Caperna, G.: Partial order theory for synthetic indicators. Doctoral dissertation, University of Padova, Italy. (2016).
4. Caperna, G.: Approximation of AverageRank by means of a formula In: Zenodo (2019) https://zenodo.org/record/2565699#.YmLzaNNBxEY. Last Accessed: 22 Apr 2022
5. Caperna, G., Boccuzzo, G.: Use of poset theory with big datasets: A new proposal applied to the analysis of life satisfaction in Italy. Soc. Indic. Res. **136**(3), 1071–1088 (2018)
6. De Loof, K., De Baets, B., De Meyer, H.: Approximation of average ranks in posets. MATCH Commun. Math. Comput. Chem. **66**, 219–229 (2011)
7. Fried, L.P., Tangen, C.M., Walston, J., Newman, A.B., Hirsch, C., Gottdiener, J., et al.: Frailty in older adults: evidence for a phenotype. J. Gerontol. A. Biol. Sci. Med. Sci. **56**, 46–56 (2001)
8. Gillick, M.: Guest editorial: Pinning down frailty. J. Gerontol. A. Biol. Sci. Med. Sci. **56**(3), M134–M135. (2001)
9. Gobbens, R.J.J., Luijkx, K.G., Wijnen-Sponselee, M.T., Schols, J.M.G.A.: In search of an integral conceptual definition of frailty: Opinions of experts. J. Am. Med. Dir. Assoc. **11**, 338-–343. (2010)
10. Ministero della Salute: Piano Nazionale della Cronicità. (2016) http://www.salute.gov.it/imgs/C_17_pubblicazioni_2584_allegato.pdf. Last Accessed: 22 Apr 2022
11. Rosano, A., Pacelli, B., Zengarini, N., Costa, G., Cislaghi, C., Caranci, N.: Aggiornamento e revisione dell'indice di deprivazione italiano 2011 a livello di sezione di censimento. Epidemiol. Prev., **44**(2-3), 162--170 (2020)

12. Silan, M., Signorin, G., Ferracin, E., Listorti, E., Spadea, T., Costa, G., Boccuzzo, G.: Construction of a Frailty Indicator with Partially Ordered Sets: A Multiple-Outcome Proposal Based on Administrative Healthcare Data. Soc. Indic. Res. **160**, 989--1017 (2020)

Time allocation and wellbeing in later life: the case of Italy

Gestione del tempo e benessere in età anziana: il caso italiano

Annalisa Donno and Maria Letizia Tanturri

Abstract Ageing processes are fundamentally linked to the concept of 'dealing with time'. In old age time use patterns change radically and how these changes are linked with wellbeing is still mostly unexplored. By using the most recent Italian Time Use Survey (2014-15) we get an insight in the association between time allocation in old people's daily routines and wellbeing in later life, in Italy. We use Sequence Analysis techniques to identify some "time use profiles" in old ages. Multinomial regressions are then used to understand which factors influence the risk to be in one of the profiles identified. Moreover, we analyse how those profiles are linked with different levels of subjective wellbeing, thus identifying high-risk groups and providing a new perspective on old people needs.

Abstract *I processi di invecchiamento sono strettamente legati al concetto di "gestione del tempo". In età anziana le routine quotidiane cambiano radicalmente e come questi cambiamenti siano collegati al benessere è ancora per lo più inesplorato. Obbiettivo di questo lavoro è studiare l'associazione tra uso del tempo e benessere in età anziana, in Italia, con i dati dell'Indagine ISTAT sull'Uso del Tempo (2014-15). Tecniche di Analisi delle Sequenze saranno utilizzate per identificare "profili omogenei di utilizzo del tempo" in età anziana, e modelli di regressione multinomiale aiuteranno a comprendere quali fattori influenzino 'il rischio' di essere inclusi in uno dei profili identificati. Infine, analizzeremo come la gestione della routine quotidiana sia associata alla soddisfazione di vita.*

Key words: Time use, old age, well-being, sequence analysis

[1] Annalisa Donno, University of Padova; email: donno@stat.unipd.it

Maria Letizia Tanturri, University of Padova; email: tanturri@stat.unipd.it

1 Introduction

Until a few decades ago, old age was considered as a period of rest in an individual's life course, where the elderly would retire and slowly disengage from society [1]. With the increase in life expectancy, however, time spent in good health and in retirement has increased considerably, and both the idea and the meaning attributed to the concept of 'ageing' have deeply changed.

In the late 1990s the World Health Organisation adopted the concept of *active ageing* [5, 1], which can be broadly defined to include not only the engagement in paid employment and physical activity but also in leisure activities that require mental (and not necessarily physical) effort or that involve social interaction, as well as in long-life education, participation in community life - for example, through volunteering work - and active engagement in household work and in the care of others.

In such a context, ageing processes can be considered as fundamentally linked to the concept of 'dealing with time' [2]. After retirement (for those who are in the labour force), time 'freed up' from paid work can be reallocated to different, passive or active, activities. It could be devoted to self-expression, self-fulfilment, thus fostering the creation of new post-work identities, new roles in societies, and allowing old people to assign new meanings to their own existence. However, even if retirement relieves individuals from obligations and leisure restrictions, the increased availability of time for out-of-paid-work activities could raise some problems too: how to fill it, how to replace structured routines with new ones, and how to find a satisfactory balance between what the elderlies would like to do, and what they can really succeed in doing. Loneliness and poor health could, for example, prevent old people from performing some desired activities, thus generating an overall sense of dissatisfaction, which could result in lower levels of subjective wellbeing. The capacity/possibility to adapt to the challenges involved in the ageing process, the way in which activities are practically substituted and redistributed, the gap between ideal daily time-use and objective constraints are likely to influence wellbeing over the later periods of a person's life.

In this paper we are interested in getting an insight in the association between time allocation and wellbeing in later life in Italy, a country that is one of the most aged in the world. It is well known that in old age time use patterns change radically, but how these changes are linked with satisfaction and wellbeing remains unexplored. As said before, the elderlies' daily allocation of time among different activities is driven by personal aspirations, wants, needs, attitudes, but is also hardly influenced by health, solitude, income levels, and family responsibility constraints.

Moreover, the way old people spend their time is shaped by the societal rhythms culturally constructed: even if everyone has their own daily routine, human activity imply patterns and moments of synchronicity. Synchronization is fundamental for interactive behaviours of humans as it strengthens interpersonal relations, thus fostering the creation of social identities, that could be an important source of wellbeing in later life.

All those elements taken into account, we want to answer the following research questions:

- Is it possible to identify homogeneous patterns of time use in later life?
- Which elements contribute in shaping such time structures?
- Do patterns of time use are correlated with the elderlies' perceived wellbeing?

Several studies show that the elderlies' subjective life evaluation is affected by their state of health, material conditions, social and family relationships, living arrangements, social roles and activities [3]. In such studies, however, only stylized questions have been used to collect information about time spent on various activities during a given time period, for example, during the past week or month. Evaluation studies suggest that those questions do not provide accurate estimates of time use compared to diaries, as they can be affected by difficulties to recall (telescoping effect) as well as by the effort to give social desirable answers (e.g. for physical activities people tend to overstate the time they dedicate to).

No research to has investigated the relation between the time allocation in different activities (analysed through time use survey diaries) and wellbeing in later life. This study proposes an original analysis of the time use in later life in Italy, that goes beyond most of the existing studies that describe individuals' time use in terms of average durations (time budget approach). We recognize the importance of chronology, timing, synchronicity in the study of daily lives by adopting a time-reckoning system based considering time use as combination of durations, ordered sequences of activities, and social meanings.

2 Data and methods

We rely on the most recent ISTAT Italian Time Use Survey (2013-14) and select 12,247 people aged 60 years and more. By using a specific type of questionnaire, the daily activity diary, the Time Use Survey collects information on how individuals allocate their time in different activities during a 24-hour day (by following a 10 minute-intervals time grid).

We go beyond the study of time use in terms of average duration and propose an innovative approach taking into account information on chronology (timing of each activity and how activities are ordered/scheduled during the day). We thus consider the individual daily allocation of time among different activities as an ordered sequence of events (144 time slot, each lasting 10 minutes). Specifically, we focus on several kind of activities that can be conceptualized in the following way:

- Basic/personal needs (sleeping, personal care, eating)
- Productive activities (paid work, housework, caring for others and volunteering)
- Socio-cultural active leisure (socializing or having gatherings at or away from home with family or friends, cinema, theatre, hobbies)

– Physical active leisure (sport, travels)
– Passive leisure (watching TV, resting, listening to music).

By following both the structure and rhythm of individual time allocation during the day, we use Sequence Analysis techniques to measure the degree of dissimilarity between all the possible pairs of sequences (i.e. all possible pairs of individuals in the sample) and to transform sequences into distances between individuals, which can then be clustered in order to uncover homogeneous patterns of time use. Specifically, we use the Dynamic Hamming Method [4] for computing distances among time use sequences. Such distances represent the cost required to make two sequences identical and are derived from the observed transition rates between states (activities), in each time slot. It is thus possible to obtain time-varying costs, inversely proportional to the probability of transition between two states (activities) in each time slot. Such an approach allows to analyse each individual time allocation scheme, in the light of the time patterning of all the other elderly, thus attributing each activity a different 'social meaning', depending on its level of synchronization with the other social actors in the sample.

At each time point, t, the cost of substituting the activity a with the activity b, in order to transform one sequence in another one, in computed as follow:

$$s_t(a,b) = \begin{cases} 4 - [p(X_t = a|X_{t-1} = b) + p(X_t = b|X_{t-1} = a) + \\ \quad p(X_{t+1} = a|X_t = b) + p(X_{t+1} = b|X_t = a)] & \text{if } a \neq b \\ 0 & \text{otherwise} \end{cases}$$

As a consequence, the distance at every moment between two individuals depends on what the entire population has done at the last stage and is about to do in the next one, which is a way to have both a dynamic and a relative definition of which behaviour is common and uncommon.

As many transition matrices as time slots are used to compute the proximity between states at every point in time. In such a way transition matrices can be considered as the statistical translation of collective rhythms. Each activity is assigned to a different meaning, depending on its temporal setting, and on the time patterning of all the other people, as substitution costs vary with the time and with the probability of transition between two states for the particular time considered. Once the dissimilarity (distance) matrix has been computed, Cluster Analysis techniques (Ward's Method) are used to see if the sequences belong to a small number of distinct types. Such an approach will allow us to identify some 'profiles of time use' in old age.

Multinomial logistic regression techniques are then used for understanding which personal attributes (age, sex, education, marital status, family type, degree of solitude, help availability, satisfaction for time use) predict the elderly's membership in one of the identified profiles. Moreover, clusters are studied as determinant of old people's wellbeing, measured by an indicator of subjective satisfaction (self-assessed life satisfaction ranging from 0 – not satisfied at all – to 10 – very satisfied – and representing the answer to the question: 'in this moment, how much satisfied are you with your life as a whole?').

3 Results

The use of the Ward clustering method, applied to the distance matrix obtained through the Sequence Analysis Dynamic Hamming approach, allowed us to identify four main time use patterns in old age, as displayed in Figure 1. Chronograms show, for each cluster, also the percentage of old people performing different activities, in each time slot.

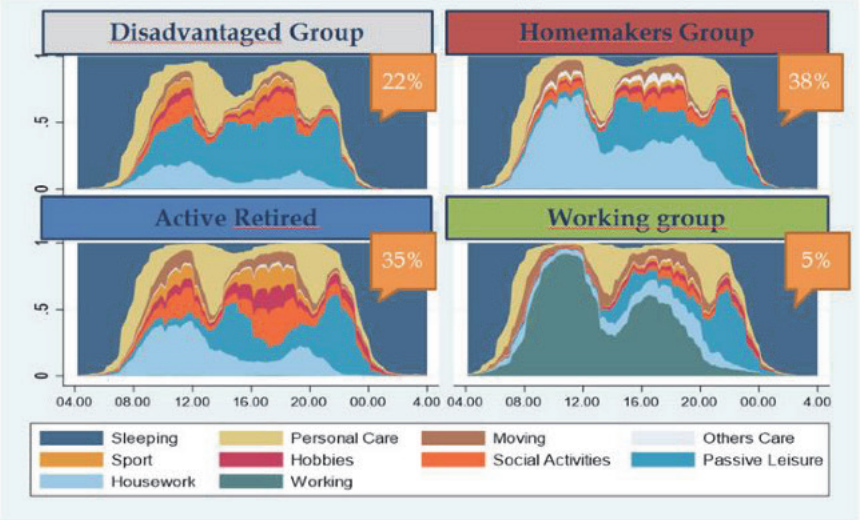


Figure 1: *Elderlies' time use profiles along the 24 hours. Chronograms.*
In the y axis the proportion of old people performing a certain activity in the time slot.
In the orange square the proportion of those belonging to each identified group.
Source: Authors' own elaboration on ISTAT Time use survey (2012-2013).

The first group (Figure 1, picture on the top to the left) consists of 22% of the selected sample of people 60 and over. Such a group seems to include 'disadvantaged' individuals, mostly performing passive activities (sleeping, personal care, passive leisure) both in the morning and in the afternoon.

Multinomial logistic regression results (Table 1, first column) show that low educated men, aged more than 75 years, not in couple, separated or divorced, unable to work, showing a certain degree of isolation (having nobody to refer to in case of need), but spending most of their day being not alone, are more likely to belong to this group. Moreover, they are likely not to be satisfied with their interpersonal relationships (they are likely to have a paid caregiver, that reduces their time spent alone, but increases their level of dissatisfaction with their 'social' relationships) and to declare having too much time to spend in resting. Probably, due to health problems, people in this group are more likely to experience, to a wider extent, the gap between what they would like to do, and what they can really do.

Table 1: *Multinomial regression results. Marginal effects.*

	Disadvantaged		Homemakers		Active retired		Workers	
Female	-0.128	***	0.171	***	-0.032	***	-0.011	***
Age 60-75	-0.096	***	0.067	***	0.026	***	0.003	
Education (Ref. Medium)								
High	-0.020		0.007		0.004		0.009	**
Low	0.048	***	0.012		-0.055	***	-0.006	
Marital status (Ref. In couple)								
Not in couple	0.082	***	-0.093	***	0.008		0.002	
Divorced	0.035	*	-0.100	***	0.069	**	-0.004	
Widowed	0.138	***	-0.135	***	-0.006		0.002	
Professional condition (Ref. Retired)								
In Paid Work	-0.076	***	-0.197	***	-0.062	***	0.335	***
Housewife	0.010		0.019		-0.026	**	-0.004	***
Inactive	0.040	**	-0.029		-0.009		-0.003	
Domestic help	0.012		-0.062	**	0.048	***	0.002	
Elderly help	0.155	***	-0.086	**	-0.044	**	-0.025	
Satisfaction for time in interpersonal relations (Ref. Yes)								
No, too much	-0.041	*	0.011		0.038		-0.008	
No, too little	0.030	***	0.001		-0.027	**	-0.003	
Not applicable	0.050	***	0.001		-0.055	***	0.003	
Satisfaction for time in leisure (Ref. Yes)								
No, too much	0.045		-0.044		-0.011		0.009	
No, too little	0.003		0.004		-0.017		0.009	**
Not applicable	0.076	***	-0.025	*	-0.050	***	-0.001	
Satisfaction for time in resting (Ref. Yes)								
No, too much	0.040	**	-0.034		-0.002		-0.004	
No, too little	-0.047	***	0.054	***	-0.014		0.007	*
Not applicable	0.049	**	-0.111	***	0.050	*	0.012	
Satisfaction for time in self-care (Ref. Yes)								
No, too much	0.024		-0.016		-0.019		0.010	
No, too little	-0.023	**	0.046	***	-0.026	*	0.004	
Proportion of time in active activities (quartile)								
2	-0.053	***	-0.130	***	0.207	***	-0.024	***
3	-0.088	***	-0.202	***	0.330	***	-0.041	***
4	-0.119	***	-0.278	***	0.452	***	-0.055	***
Proportion of time in activities alone (quartile)								
2	-0.052	***	0.053	***	0.008		-0.009	***
3	-0.093	***	0.127	***	-0.021	*	-0.013	***
4	-0.133	***	0.192	***	-0.045	***	-0.014	***

Note 1: *** p<0.001, ** p<0.05, * p<0.1

Note 2: controlling for geographic area of residence, weekday, family type, self-assessed economic resources

Individuals belonging to the second clusters (*homemakers group*) perform mainly housework activities in the morning, while in the afternoon some of the time spent in housework is substituted with passive leisure and social activities (Figure 1, graph on the top to the right). Only a low percentage of people in this group perform 'active' leisure. Results from the multinomial logistic regression (Table 1, column 2) evidence that women aged between 60 and 75 years, living in couple or alone with their children, having no paid domestic aid neither a paid caregiver, being housewives and spending time alone are more likely to perform their daily activities by following the time allocation scheme reassumed in the second chronogram. They are also more likely to feel that the time they spend in resting and in personal care is not enough. It is possible that some time shortage issues (due to their intense participation in housework activities) characterize women's life also at older ages.

The third group (Figure 1, on the bottom to the left), representing 35% of the sample, include the most active individuals: the highest proportion of people performing active leisure (sport, transports, hobbies, social activities), both in the morning and in the afternoon, is observed in this group. Membership to the 'active retired' cluster (Table 1, column 3) is significantly more likely for more educated, retired men, aged 60-75 years, separated or divorced, but in couple, declaring to be satisfied with the time devoted to personal care and social relationships.

The fourth group identified (*working group*) is residual as it includes only 5% of the analysed sample (Figure 1, on the bottom to the right), but nevertheless it is strongly characterised by the most intense participation to labour market activities. People in this group spend most of their time in paid-work activities and they are more likely to be not satisfied with the time they can devote to leisure and resting (Table 2, column 4).

In order to study the relationship between time use patterns and wellbeing at older ages, we run an OLS regression using the level of self-assessed life satisfaction as a dependent variable, and the time-use cluster membership as an independent one, together with other control variables, which are also hypothesized to influence the old people's well-being (solitude, help availability, family type, etc.).

Results (Table 2) confirm that patterns of daily time use correlate significantly with the level of self-rated satisfaction, even when we control for other relevant individual characteristics that also correlates significantly with life satisfaction. The time use clusters (summarizing the way old people allocate their time in different activities) catch detailed information on individual's lives that could not have been observed by using traditional time use measures (durations), and allows us to add new and important evaluation elements in the study of well-being.

Individuals in the 'disadvantaged' group are more likely to have lower levels of life satisfaction, with respect to those in the 'active retired' group (the disadvantaged's time use patterns could also be affected by some health disease, we could not account for, due to lack of such information in our data). Moreover, working a lot in old age seems to effect negatively life satisfaction, probably because of the squeeze of leisure and resting time. Active aging policies should take into account this results to suggest shorter work scheduled for older workers.

Table 2: *Time use patterns and life satisfaction in old age. OLS results*

Variables	Categories	Coef.	
Time use profile	Ref. Active retired		
	Disadvantaged	-0,68	***
	Homemakers	-0,08	*
	Workers	-0,2	***
More than 75% of the day alone	Ref. Not		
	Yes	-0,21	***
Living arrangement	Ref: Couple		
	Alone	-0,26	***
	Alone in children hh	-0,54	***
	Couple with children	-0,03	
	Lone parent	-0,41	***
	Other	-0,12	
In case of necessity	Ref. Nobody		
	Children available	0,205	***
	Sibling available	0,144	***
	Grandchildren available	0,082	*
	Other Relatives available	0,064	
	Friends available	0,166	***
	Neighbours available	0,027	
Paid domestic help	Ref. No		
	Yes	-0,14	*
	Constant	7,875	***

*** p<0.001, ** p<0.05, * p<0.1

(Controlling for sex, age, education, geographic area of residence, self-assessed economic resources)

Acknowledgements

We acknowledge the contribution of the CREW project (Childcare, REtirement and well-being in old ages) funded by the program JPI-More years better Lives)

References

1. Boudiny, K.: 'Active ageing': From empty rhetoric to effective policy tool. Ageing and Society, 33(6): 1077-1098 (2013)
2. Ekerdt, D. J., Koss, C.: The task of time in retirement. Ageing & Society, 36(6), 1295-1311 (2016)
3. Gauthier, A.H., Smeeding, T.M.: Time use at older ages: Cross–nationaldifferences. Research on Aging, 25(3): 247-274 (2003)
4. Lesnard, L.: Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociological Methods & Research, 38(3): 389-419 (2010)
5. Walker, A.: A strategy for active ageing. International Social Security Review, 55(1): 121-139 (2002)

The role played by migration and fertility on Italy's aging trends: a provincial-level analysis

Il ruolo delle migrazioni e della fecondità nel processo di invecchiamento in Italia: un'analisi a livello provinciale

Thaís García-Pereiro and Anna Paterno

Abstract The main purpose of this paper is to quantify and compare the contributions made both by fertility and migration in the rapid aging process taking place in Italy at the provincial level. The relative variations of different indicators (concerning migration, fertility, mortality and age structure of the population) between 2011 and 2019 are analyzed in two differentiated empirical steps. In the first, through principal components factors analysis, both the relationships among variables under examination and the dynamics of their evolution at the provincial level are defined. In the second step, estimating a regression model, the roles of the determinants linked to fertility and migration on the evolution of the aging process are identified and quantified. Our results indicate the levels of fertility of Italian women as the most important decelerator of population aging, within a highly heterogeneous context at the provincial level.

Abstract *L'obiettivo di questo lavoro è quantificare e comparare tra loro i contributi forniti a livello provinciale sia dalla fecondità, sia dalle migrazioni nel processo di rapido invecchiamento in atto in Italia. A tal fine si osservano in due step differenziati le variazioni relative di diversi indicatori (riguardanti migrazioni, fecondità e mortalità e struttura per età della popolazione) verificatesi tra il 2011 e il 2019. Nel primo, attraverso un'analisi delle componenti principali, si definiscono le relazioni esistenti tra le variabili osservate e la loro evoluzione a livello provinciale. Nel secondo step, applicando un modello di regressione, si identificano e quantificano i ruoli delle determinanti connesse alla fecondità e alla migrazione sull'evoluzione del processo di invecchiamento. I risultati indicano il livello di fecondità delle donne*

[1] Thaís García-Pereiro, Università degli Studi di Bari; email: t.garcia.pereiro@uniba.it

Anna Paterno, Università degli Studi di Bari; email: anna.paterno@uniba.it

The authors acknowledge the financial support provided by the MiuR-PRIN Grant N. 2017W5B55Y (GDR project "The Great Demographic Recession", PI: Daniele Vignoli).

italiane come il più importante fattore nel decelerare l'invecchiamento della popolazione, all'interno di un contesto provinciale altamente eterogeneo.

Key words: aging, fertility, migration, Italy, provinces, demographic trends.

1 Introduction

Aging is one of the main long-term demographic challenges that most western countries are called to address within the near future because, given current trends, it might compromise the response capacity (in terms of both quantity and quality) of at least two important components of welfare systems: health and pension. As stated by Spijker and MacInnes (2013), population aging is hardly impacting the sustainability of health systems, and this will make essential for governments to deal with improving the relationship between morbidity and remaining life expectancy at older ages. Also the pension system is going to be well-overstressed due to the sharp combination of an increase of its recipients and a decline of its contributors (Bongaarts, 2004).

In Italy, the economically active population is progressively declining, which is also placing significant pressure on economic growth and public expenditures, given the boosted demand of public health-care related services and pensions, in terms of unbalance between actors involved (ISTAT, 2020). The country has become one of the worlds' oldest countries and, at the same time, hides important territorial differences (Dalla Zuanna and Righi, 1999; García-Pereiro, 2018), representing an interesting case of study.

It is well known that the population age-sex structure of any territory depends on three demographic components: fertility, mortality and international migration. Within the context of increasing population aging and considering that any *adhoc* modifications on mortality are off-limits, only the increase of very-low fertility levels reached and/or of net migration can help slowing down the process.

Therefore, the main purpose of this paper is to assess whether and how fertility and migration trends have affected population aging during a recent interval of time (between 2011 and 2019). We are completely aware that there are other "solutions" to aging, but here we are only interested on those responding to changes on demographic components.

2 A brief review of the state of the art

A vast body of research has focused on the demographic determinants (mortality, fertility and migration) of population aging (Preston and Stokes, 2012; United Nations Department of Economic and Social Affairs, 2015; Murphy, 2017; Lee and Zhou, 2017). Most studies have indicated that the major responsible for population aging has been declining fertility (Boogarts, 2008; Bengtsson and Scott, 2011; Billari and

Dalla Zuanna, 2011; Bloom et al., 2015; Murphy, 2017). As Lee and Zhou (2017) have shown, population aging has been a direct consequence of fertility decline, independently of mortality trends.

Lee and Mason (2014) have stressed the important role played by fertility levels on a population age structure also highlighting that, in countries with very low fertility levels, increasing fertility will result on a moderate increase on the standards of living. A recent study on Poland by Fihel et al. (2018) has noticed that variations on age-specific growth rates were predominantly explained fertility fluctuations, second, mortality trends, and, last, international migration flows, in this order.

The arrival of individuals to a territory implies short-term changes on its population size and structure but also long-term variations because, if individuals remain at destination, they might contribute to local fertility (increasing the number of births) (Mussino and Strozza, 2012; Giannantoni and Strozza, 2015).

Concerning migration, international literature has reported mixed results. Part of these studies has shown a rejuvenating effect of migration inflows on aged populations (Alho, 2008; Chen, 2015; Fihel et al., 2018), while others have stressed that effect of the presence of foreigners on populations' age structures is negligible (Goldstein, 2009; Bengtsson and Scott, 2010; Murphy, 2017). Projection-based studies (UN 2000, Bijak et al., 2008; Bijak, et al., 2013; Kupiszewski, 2013; Craveiro et al., 2019) have found that the number of immigrants necessary to contrast population aging must be excessively large, profiling unrealistic forecasts.

Several studies have focused on the Italian case (Billari and Dalla Zuanna, 2011; De Santis, 2011; Gesano and Strozza, 2011; Paterno, 2011; Gesano and Strozza, 2019) concluding that immigration alone will not be enough to deal with populating aging, especially if fertility levels remain low, but might help by slowing it down for a while. Authors also highlighted the need to consider that the effects of fertility and migration on aging vary greatly at local levels, given their particular combinations of demographic trends.

3 Data and methods

Data were drawn from demographic statistics available at the provincial level (107 provinces) (NUTS3) from the Italian National Institute of Statistics (ISTAT) for the interval of time between 2001 and 2019. Data is referred not only to the total resident population in Italy, but also to the resident foreign population.

Our dependent variable, signalling population aging, is mean age (Mean Age). We chose this indicator among others (Old Age Dependency Ratio, etc.) based on Murphy (2017) results showing no significant variation when interpreting aging determinants using different measures. Other indicators included in our analyses are: Total Fertility Rate of Italian women (TFR_it), to account for the role of fertility trends; Total Fertility Rate of foreign women (TFR_for), to consider the contribution of foreign fertility to national levels; life expectancy at birth of males (Lexp M) and females (Lexp F), to take into account longevity; the share of foreigners among total population (Foreigners), to weight for the stock of individuals coming from foreign

countries, net migration rates (NetMigR), to control for interprovincial migration flows, and, finally, the mean age of foreigners (Mean Age_for) to evaluate their particularly young age-structure.

All these variables are analyzed in terms of relative variations (computing the ratio between the absolute variation during the interval and the value registered at the first year of the interval). These allowed us to make more accurate the comparison of indicators using different units of measurement.

Our empirical strategy follows two well differentiated steps. In the first, after performing descriptive analysis, we conduct principal components factor analysis to better describe relationships among demographic components of population change throughout a reduction of information while highlighting similarities/dissimilarities across provinces. In the second, we estimate a linear regression model on provincial data to identify the determinants of aging of the resident population in Italy in the last decade. Thus, the dependent variable is continuous and measures variations in the mean age of the resident population between 2011 and 2019. As independent variables, model includes the values of demographic components considered (TFR of Italian women, incidence of foreigners, life expectancy at birth for males and females, net interprovincial migration rate, TFR of migrant women, mean age of foreigners) at the beginning of the period (2011) together with their respective absolute variations during the period under observation (between 2011 and 2019) plus the mean age of resident population in 2011.

4 Main findings

The first part of this section is dedicated to a description of the evolution of aging, fertility and migration trends between 2011 and 2019 at the provincial level. Figures shown here only plot the ranking of the top 5 (highest and lowest variations) of the mean age of the resident population, of TFR of Italian women and the share of foreigners among total resident population.

Between 2011 and 2019, the mean age of the resident population increased the most (around 6-7%) in two provinces of Puglia (Barletta–Andria-Trani, and Bari) and three of Sardegna (Sud Sardegna, Cagliari, Oristano). Instead, those provinces where the increase was the lowest (1-2%) were located in the North, in particular, three in Emilia Romagna (Bologna, Parma and Piacenza), one in Liguria (La Spezia) and the last one in Friuli-Venezia-Giulia (Trieste).

Regarding relative variations of TFR of Italian women (Figure 2), most provinces experimenting declining fertility are in Central (Massa Carrara, Roma and Grosseto) and Northern (Valle d'Aosta, Verbano) regions, with relative variations that oscillate between 18.9% and 25%. There are only three provinces that register increases on their fertility levels: Isernia (8.3%), Bolzano (5.3%) and Crotone (4.5%), while at Consenza and Vibo Valentia, fertility remain almost unvaried.

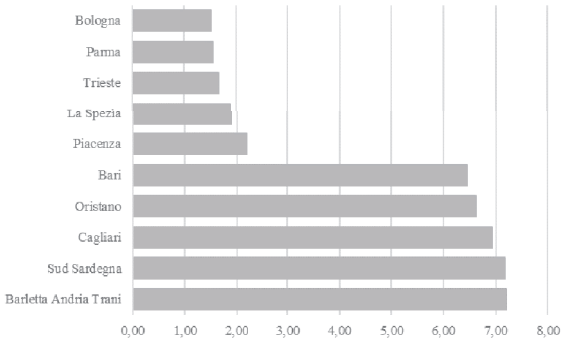


Figure 1: *Relative variations in the mean age of resident population between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

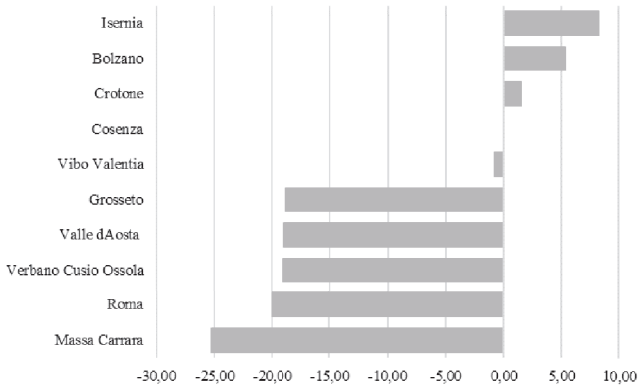


Figure 2: *Relative variations in the Total Fertility Rate of Italian women between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

Figure 3 illustrates changes on the share of foreigners among total population under the period interval under study. Out of 107 provinces, only 13 experience a decrease on the stock of foreigners. Those with the highest negative values are in North (Vicenza, Treviso, Brescia) and Center (Macerata, Pesaro and Urbino) of the

country. In contrast, the highest values (around 6 and 7%) are registered in Southern provinces (Crotone, Trapani, Benevento, Campobasso and Cagliari).

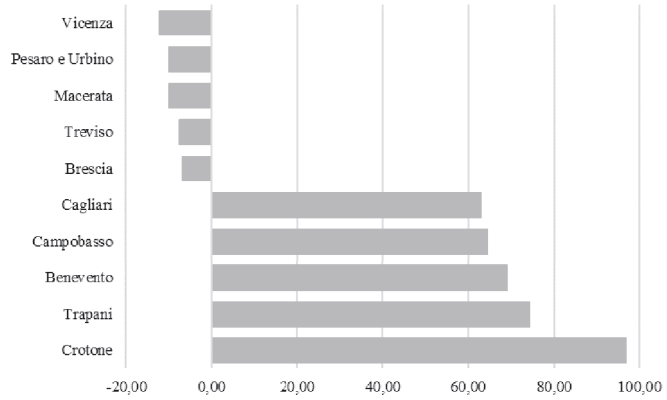


Figure 3: *Relative variations in the share of foreigners among total resident population between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

In following multivariate analysis, we summarize relationships observed at the provincial level among observed variables.

The components matrix resulting from principal components factor analysis (Table 1) allows us to identify four factors explaining 76.82% of the total variability of provincial data. The first factor absorbs 31.16% of the total variance and summarize positive variations of the mean age of resident population (in contrast with foreigners mean age variations) and the presence of foreigners. The second factor, explaining 19.12% of the total variance, is associated to negative variations of fertility of both Italian and foreign women. The third factor (covering 14% of total variance), is tied to the increase in life expectancies at birth of males and females, while the fifth fully represented period variations on inter-provincial net migration.

Table 1: *Matrix of components resulting from principal components factor analysis. Rotation method: Varimax with Kaiser normalization.*

Relative variations	Factor1	Factor2	Factor3	Factor4
Mean Age	0.712	-0.508	-0.008	0.062
TFR_it	0.425	0.525	-0.267	-0.309
TFR_for	0.057	0.805	-0.038	0.312
Lexp M	-0.156	-0.192	0.746	0.195
Lexp F	0.176	0.079	0.807	-0.288
Foreigners	0.937	0.065	0.025	0.049
NetMigR	0.100	0.156	-0.048	0.892
Mean Age_for	-0.852	-0.368	0.009	-0.036
Variance exp.	76.82%	31.16%	19.12%	14%

Figure 4 plots factor scores of the first two factors for each province. Following a counter-clockwise order, we find, in the upper-right quadrant, provinces showing positive values with respect to increasing population aging and contrasted by growing shares of foreigners -which are also growing older- (Factor 1), and less accentuated or even positive variations on fertility (Factor 2). Most of these provinces are situated in the South, for example, Isernia in Molise, Trapani in Sicilia and Benevento and Avellino in Campania, share high values in both factors.

The second quadrant, with positive values on Factor 1 and negative ones in Factor 2, comprises provinces that are predominantly located in the South and Islands, with the Sardinian ones (i.e.: Cagliari, Sud Sardegna) showing the highest combination of values. Here, increases in the mean age and share of foreigners, characterizing positive values of Factor 1, are illustrated in contrast to high negative variations on TFR of foreign women.

The third quadrant comprises provinces with negative figures of both factors and is predominantly represented by northern provinces (i.e.: Reggio nell'Emilia, Rovigo, Viterbo, Brescia e Bergamo). In these provinces both the aging pace (Factor 1) and TFR variations for Italian women (Factor 2) are less pronounced.

The last quadrant, illustrating positive values on the horizontal semiaxis and negative ones on the vertical semiaxis, also includes provinces located in northern areas of the country (such as: Piacenza, Parma, Bologna, Livorno). They share the lowest positive variations on mean age and on the percentage of foreign residents (Factor 1), in opposition to negative variations on fertility, smaller in terms of its magnitude (Factor 2).

Figure 4 is underpinning the deep North-South gap regarding the recent evolution of observed dynamics, which is clearly evident on the net division of the quadrants.

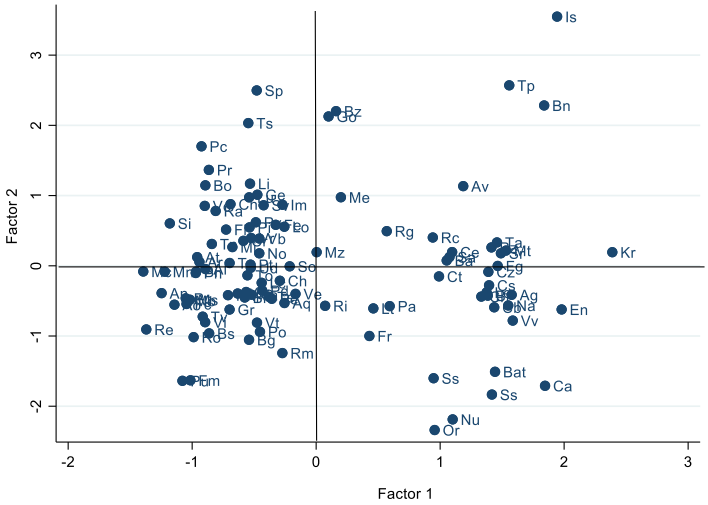


Figure 4: *Positions of provinces on the factor plane (first and second factor) resulting from principal components factor analysis.*

The last step of our analyses regards results coming from a linear regression model on provincial data (Table 2) on the determinants of mean age variations[1]. Our findings show that provinces in which the increase in mean age is larger between 2011 and 2019 are those where the TFR of Italian women was lower both at the beginning of the observation period (ß = -0.476), and when considering its evolution over the observation period (ß = -0.490). Also the presence of foreigners acts shielding against population aging, even if having a weaker impact on it respect to natives fertility. In fact, mean age increases are smaller in provinces with the highest shares of foreigners both in 2011 (ß = -0.222) and successively (ß = -0.289). The third determinant showing a negative relationship is inter-provincial net migration, which also explain mean age variations but to a lesser extent (ß = -0.065) and only in 2011.

Regarding female life expectancy at birth, positive coefficients for its values in 2011 (ß = 0.236) and between years of observations (ß = 0.227), indicate the existence of a direct relation with mean age variation at the province level.

Table 2: *Determinants of the absolute variation of the mean age of the population resident in Italy between 2011 and 2019 from linear regression model with provincial-level data.*

Independent variables	Coeff	SE	Sig.
Mean Age 2011	-0.480	0.037	***
TFR_it 2011	-0.476	0.624	***
TFR_for 2011	0.292	0.283	
Lexp M 2011	-0.123	0.091	
Lexp F 2011	0.236	0.104	**
Foreigners 2011	-0.222	0.025	***
NetMigR 2011	-0.065	0.028	**
Mean Age_for 2011	-.016	0.046	
AV TFR_it (2001-2019)	-0.390	0.717	***
AV TFR_for (2001-2019)	0.151	0.238	
AV Lexp_M (2001-2019)	-0.035	0.114	
AV Lexp_F (2001-2019)	0.227	0.112	**
AV Foreigners (2001-2019)	-0.289	0.058	***
AV NetMigR (2001-2019)	0.038	0.031	
AV Mean Age_for (2001-2019)	0.099	0.073	
Constant	*22.168*	*7.397*	**
R^2	90.45%		
N	107		

Notes: * p<0.1; ** p<0.05; *** p<0.001

[1] Independent variables include in the model are both values at 2011 of mean age (Mean Age 2011), TFR of Italian (TFR_it 2011) and foreign women (TFR_for 2011), life expectancy at birth of males (Lexp M 2011) and females (Lexp F 2011), share of foreigners (Foreigners 2011), interprovincial net migration rate (NMigR 2011) and mean age of foreigners (Mean Age_for 2011); and absolute variations of TFR of Italian (AV TFR_it 2001-2019) and foreign women (AV TFR_for 2001-2019), life expectancy of males (AV Lexp_M 2001-2019) and females (AV Lexp_F 2001-2019), percentage of foreigners (AV Foreigners 2001-2019), net migration rates (AV NetMigR 2001-2019) and mean age of foreigners (AV Mean Age_for 2001-2019). Significant and negative coefficients (in order of magnitude) were found for the level of fertility of Italian women, the share of foreigners and the net inter-provincial migration rate (p-value <0.001), and a positive one for female life expectancy at birth (p-value <0.05).

5 Brief discussion

This paper was aimed at providing a description of whether and how fertility and migration trends (international and internal) have affected population aging in Italy between 2011 and 2019 at the provincial level.

When summarizing relationships observed among variables included in this study, we found that there are two factors explaining more than half of total variability. The first deals with increases on the mean age of resident population (which is contrasted with a younger age structure of foreigners) and the increasing share of foreigners in Italian provinces. The second factor is linked to decreasing fertility trends of both Italian and foreign women.

Results on the determinants of recent population aging (measured through the relative variation of the mean age of the resident population between 2011 and 2019) indicate both fertility of Italian women and the presence of foreigners as important decelerators of population aging, but the first has exerted the greatest impact. This finding is in line with previous research on the subject stressing the predominant role that fertility has had (over longevity and migration) as main responsible for population aging (Boogarts, 2008; Bengtsson and Scott, 2011; Bloom et al., 2015; Murphy, 2017; Lee and Zhou, 2017; Fihel et al., 2018). According to our estimations, provinces more efficaciously contrasting their increasing mean age were those with higher levels of fertility, both at the beginning of the interval analysed (2011) and when considering the evolution measured up to 2019.

As the effects of fertility and migration on aging might considerably vary at local levels when considering their specific demographic profiles (Billari and Dalla Zuanna, 2011; De Santis, 2011; Gesano and Strozza, 2011; Paterno, 2011), further research should consider testing whether and how the contribution of these population components of change on slowing down aging have differed according to the magnitude already achieved at the beginning of the period under analysis.

References

1. Alho, J. Migration, fertility, and aging in stable populations. Demography 45(3): 641–650. doi:10.1353/dem.0.0021 (2008)
2. Bengtsson, T. and Scott, K. The ageing population. In: Bengtsson, T. (ed.). Population ageing: A threat to the welfare state? Berlin: Springer: 7–22. doi:10.1007/978-3-642-12612-3_2 (2010)
3. Bengtsson, T. and Scott, K. Population aging and the future of the welfare state: the example of Sweden. Population and Development Review 37(1): 158–170. doi:10.1111/j.1728-4457.2011.00382.x (2011)
4. Bijak, J., Kupiszewska, D., and Kupiszewski, M. Replacement migration revisited: Simulations of the effects of selected population and labor market strategies for the aging Europe, 2002–2052. Population Research and Policy Review 27(3): 321–342. doi:10.1007/s11113-007-9065-2 (2008)
5. Bijak, J. and Kupiszewski, M. International migration trends in Europe prior to 2002. In: Kupiszewski, M. (ed.). International migration and the future of populations and labour in Europe. Dordrecht: Springer: 57–74. doi:10.1007/978- 90-481-8948-9_4 (2013)
6. Billari, F. C., and Dalla-Zuanna, G. Is replacement migration actually taking place in low fertility countries?. Genus, 67(3), 105-123 (2011)

7. Bloom, D. E., Canning, D., & Lubet, A. Global population aging: Facts, challenges, solutions & perspectives. Daedalus, 144(2), 80-92 (2015)
8. Bongaarts, J. Population aging and the rising cost of public pensions. Population and Development Review, 30(1), 1-23 (2004)
9. Bongaarts, J. What can fertility indicators tell us about pronatalist policy options?. Vienna yearbook of population research, 39-55 (2008)
10. Chen, C.-Y. The effect of migration on the mean age of population: An application of Preston's mean age of population improvement model. Journal of Family History 40(1): 92–110. doi:10.1177/0363199014562711 (2015)
11. Craveiro, D., De Oliveira, I. T., Gomes, M. S., Malheiros, J., Moreira, M. J. G., and Peixoto, J. Back to replacement migration. Demographic research, 40, 1323-1344 (2019)
12. Dalla Zuanna, G., and Righi A. Nascere nelle cento Italie. Analisi territoriale del comportamento riproduttivo nelle province italiane. Argomenti (18), Istat, Roma (1999)
13. De Santis, G. Can immigration solve the aging problem in Italy? Not really…. Genus, 67(3), 37-64 (2011)
14. Fihel, A., Janicka, A., and Kloc-Nowak, W. The direct and indirect impact of international migration on the population ageing process: A formal analysis and its application to Poland. Demographic Research, 38, 1303-1338 (2018)
15. García-Pereiro, T. Aging and pensions in Italy: highlighting regional disparities. Rivista Italiana di Economia Demografia e Statistica, 72(3), 17-28 (2018)
16. Gesano, G., and Strozza, S. Foreign migrations and population aging in Italy. Genus, 67(3), 83-104 (2011)
17. Gesano, G., and Strozza, S. Fecondità delle italiane e immigrazione straniera in Italia: due leve alternative o complementari per il riequilibrio demografico?. la Rivista delle Politiche Sociali, 4, 119-140 (2019)
18. Giannantoni P., Stozza S. Foreigners' contribution to the evolution of fertility in Italy: a re-examination on the decade 2001-2011. Rivista Italiana di Economia Demografia e Statistica, vol. LXIX, n. 2, 129-140 (2015)ISTAT. Invecchiamento attivo e condizione di vita degli anziani in Italia. ISTAT, Roma (2020)
19. Kupiszewski, M. International migration and the future of populations and labour in Europe. Dordrecht: Springer Science and Business Media. doi:10.1007/978-90-481-8948-9 (2013)
20. Lee, R. and Zhou, Y. Does fertility or mortality drive contemporary population aging? The revisionist view revisited. Population and Development Review 43(2): 285-301 (2017)
21. Lee, R., Mason, A. Is low fertility really a problem? Population aging, dependency, and consumption. Science, 346(6206), 229-234 (2014)
22. Murphy, M. Demographic determinants of population aging in Europe since 1850. Population and Development Review, 43(2), 257–283. doi:10.1111/padr.12073 (2017)
23. Mussino, E., & Strozza, S. The fertility of immigrants after arrival: The Italian case. Demographic research, 26, 99-130 (2012)
24. Paterno, A. Is immigration the solution to population aging?. Genus, 67(3), 65-82 (2011)
25. Preston, S. H. and Stokes, A. Sources of population aging in more and less developed countries. Population and Development Review, 38(2), 221-236 (2012)
26. Spijker, J. and MacInnes, J. Population Ageing: The Timebomb that Isn't?. British Medical Journal (BMJ), vol. 347, 6598 https://doi.org/10.1136/bmj.f6598 (2013)
27. United Nations. Replacement Migration: Is it a Solution to Declining and Ageing Populations?, Population Division, Department of Economic and Social Affairs, United Nations Secretariat, New York (2000)
28. United Nations, Department of Economic and Social Affairs, Population Division. World population Prospects: Key findings and advanced tables. The 2015 revision. UN, New York (2015)

New challenges in the labour market

Detecting changes and evolution in specialized professional figures: an application on the Italian IT & Digital sector

Cambiamenti ed evoluzione nelle figure professionali specializzate: una applicazione sul settore IT & Digital in Italia

Andrea Marletta

Abstract In this paper, the relationship between job professions and requested skills for getting a job in Italian Labour market is investigated using a dynamic approach searching for trends during the considered period. From a methodological point of view, a multi-way dataset considering internal and external sources is analysed drawing time trajectories using a Weighted Factor Analysis. In particular, for the IT and Digital industry the profiles of workers recruited by The Adecco Group in Italy in the period 2018-2021 have been analysed detecting evidences and movements.
Abstract *Il contributo analizza la relazione fra professioni e competenze richieste nel mercato del lavoro italiano usando un approccio dinamico alla ricerca di nuove tendenze all'interno del periodo considerato. Da un punto di vista metodologico, l'analisi riguarda un dataset multivariato creato unendo fonti interne e fonti esterne all'azienda che definisce delle traiettorie temporali usando una Weighted Factor Analysis. In particolare, per il settore IT & Digital, questa tecnica è stata applicata ai profili dei lavoratori reclutati da The Adecco Group in Italia nel periodo 2018-2021 alla ricerca di evidenze e movimenti.*

Key words: Italian labour market, Job matching, Multi-way data

1 Introduction

In social and economic systems, the role of labour is fundamental, both for the aspects strictly related to labour as a production factor and for the perspectives regarding workers. The access to the labour market represents a key point for the supply and demand side. About the supply, the role of knowledge, abilities and attitudes leads to the consideration of models and formative offers for their creation

Andrea Marletta
Department of Economics, Management and Statistics, University of Milano-Bicocca,
e-mail: andrea.marletta@unimib.it

and implementation. On the other hand, about the demand, the economic context and the effect of technical progress activate examples of improvement in roles and difficulties in the definition of short-term scenarios.

In this context, institutions and governments asked for a more highlighted view on a labour force more renewed. For example, according the World Economic Forum, more than half of all employees will request a re-qualification before 2022. Among these employees, one third will need further education for six more months, and one fifth will need further education for a longer period [17]. In addition, following the instructions of the International Labour Organization (ILO), enterprises and employers will need to make new investments to expand their involvement in the education, training and re-skilling of workers to support economic growth. Workers will need to pro-actively upgrade their skills or acquire new ones through training, education and learning to remain employable [7].

This indications led to the assumption of a central role for competencies in the competitiveness of firms and workers; in this sense, they could represent a keystone of the retribution. Competencies may become a candidate in the integration or substitution the remunerative parameters, thus serving as a new tool in the relationship between jobs and wages. Information regarding goodwill, albeit with a managerial and administrative slant, provides a source of knowledge structured on the basis of the criteria that companies adopt in their choices of workers who apply for job positions in their companies.

In this perspective, there is need to search for innovative tools capable to measure the importance of this parameter based on real and recent data and the aim of this work is try to fill this gap. Following this viewpoint, the use of multivariate statistical techniques on these data led to trace a link between job offers and competencies' candidates. A possible application of this approach is presented through an analysis based on research proposed by The Adecco Group on new hires starting from 2018 to 2021 in Italy. Using this method, it is possible to define a time trajectory for some professional roles detecting trends and dynamics useful in the recruiting process. This representation appears to be also beneficial to check the existence of clusters of skills and to analyse the relationship between a job title and the skills.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results. Finally, some conclusions will follow.

2 Time trajectories in multi-way data

This contribution aims to give some indications about the relationship between candidates approaching to job offers and the requirements owned by whom that obtained that position. The requested requirements by companies during the hiring process could be divided into 3 categories: knowledges, abilities and attitudes. Knowledges are a set of structured principles and theories useful for the correct im-

plementation of the profession. Abilities are procedures and processes defining the capabilities to accomplish the professional tasks and they are commonly called hard skills. Attitudes are cognitive features affecting the professional development and the execution of job activities. This study is focused on the attitudes intended as soft skills.

In this contribution, a set of soft skills for a subgroup of job titles is observed over the 2018-2021 period. These structure led to a multivariate time array $\mathbf{X}$ [3, 10], so composed:

$$\mathbf{X} \equiv \left\{ x_{ijt} : i = 1,\ldots,I; j = 1,\ldots,J; t = 1,\ldots,T \right\} \tag{1}$$

where i is a generic unit (i.e. a professional role), j is one of the observed variables and t is a year within the 2018-2021 period. Such three-way data can be re-arranged to obtain the so-called multivariate time trajectories [1, 2, 3], displaying the path of each professional figures over the years on a J-dimensional space.

The re-arrangement of the multivariate time array $\mathbf{X}$ takes place in two steps. The observed values for all figures in a given year t are selected from the multivariate time array $\mathbf{X}$, obtaining an $I \times J$ matrix which is called "slice" [3, 10].

Once a slice has been created for each year t (with $t = 1,\ldots,T$), the slices are stacked one on the top of the other until the matrix $\widetilde{\mathbf{X}}$ with $I \cdot T$ rows and J columns is achieved. The generic row of $\widetilde{\mathbf{X}}$, denoted by $\mathbf{x}_{it}$, contains the observed values for job title i in year t:

$$\mathbf{x}_{it} \equiv x_{i1t},\ldots,x_{iJt}. \tag{2}$$

When a single job title i is considered, the matrix displaying the time trajectory i is obtained by selecting the J-dimensional vectors $\mathbf{x}_{it}$, with $t = 1,\ldots,T$, from $\widetilde{\mathbf{X}}$ [4]:

$$\widetilde{\mathbf{X}}_i \equiv \left\{ \mathbf{x}_{it} : t = 1,\ldots,T \right\}. \tag{3}$$

A multivariate time trajectory $\widetilde{\mathbf{X}}_i$ can be achieved for each job title i, with $i = 1,\ldots,I$, and then such trajectories can be compared to detect the dissimilarities among professional figures. D'Urso [3] compared the multivariate time trajectories of different statistical units by using a geometric setting where each unit i was located on T parallel J-dimensional spaces. Liberati and Mariani [14] applied a principal component analysis (hereafter, PCA) [8] to the matrix $\widetilde{\mathbf{X}}$ in order to reduce the number of variables J.

When applying PCA to a data set with J variables, Q new latent factors are created, where Q is less than J. Such new factors are obtained in a way that ensures the loss in statistical information is minimized for each Q (with $Q = 1,\ldots,J-1$), as measured by the proportion of total variance that is not explained by the Q new variables. These principal components (henceforth, PCs), are uncorrelated with each other by construction. A principal component, indicated by $\mathbf{y}$, is given by the linear combination $\mathbf{y} = \sum_{j=1}^{J} a_j \mathbf{x}_j = \widetilde{\mathbf{X}}\mathbf{a}$, where $\mathbf{x}_j$ is the j-th column of $\widetilde{\mathbf{X}}$ and $\mathbf{a} = \{a_1,\ldots,a_J\}$ is a vector of coefficients [9]. The elements of $\mathbf{a}$ are chosen to maximize the variance of $\mathbf{y}$, which is:

$$Var(\mathbf{y}) = Var\left(\widetilde{\mathbf{X}}\mathbf{a}\right) = \mathbf{a}^T \Sigma \mathbf{a} \tag{4}$$

where Σ stands for the variance-covariance matrix of $\widetilde{\mathbf{X}}$. To find the vector $\mathbf{a}$ maximizing $\mathbf{a}^T \Sigma \mathbf{a}$, the constraint that $\mathbf{a}$ is a unit-norm vector (i.e. $\mathbf{a}^T \mathbf{a} = 1$) is commonly imposed. Once this is done, the problem can be solved by using the method of Lagrange multipliers, that means finding the maximum of the function $L(\mathbf{a}) = \mathbf{a}^T \Sigma \mathbf{a} - \lambda \left(\mathbf{a}^T \mathbf{a} - 1\right)$ [9]. After differentiating and setting the first derivative equal to $\mathbf{0}$, it is obtained:

$$\Sigma \mathbf{a} = \lambda \mathbf{a}. \tag{5}$$

Equation 5 shows that $\mathbf{a}$ is an eigenvector of Σ and λ is the respective eigenvalue. Given equations 4 and 5, the variance of $\mathbf{y}$ is equal to λ. Choosing the greatest eigenvalue of Σ, denoted by λ_1, the corresponding eigenvector $\mathbf{a}_1$ gives the linear combination with the largest variance $\mathbf{y}_1 = \widetilde{\mathbf{X}}\mathbf{a}_1$, i.e. the first PC. The second PC is obtained by using the same method with the additional restriction that the two eigenvectors must be orthogonal, i.e. $\mathbf{a}_1^T \mathbf{a}_2 = 0$. Such an approach can be used to create up to J PCs, which are uncorrelated [9]. As the target of PCA is reducing the number of variables to be used, only Q PCs (with $Q < J$) are held. When PCA is applied to $\widetilde{\mathbf{X}}$ and only the first two PCs are held, we obtain a two-dimensional plane [6, 13, 14] in which the time trajectory of each unit is depicted in the space spanned by the first two PCs. The advantage of such an approach is that the time trajectory can be displayed by connecting its PC scores, calculated for each year in the period considered, in a Cartesian plane.

3 Application

In this paper, the dataset is obtained as a merge of business sources in combination with external sources. Internal sources are represented by the Adecco Group database on job offers and necessary requirements for the hires. External sources are the ESCO (European Skills, Competences, Qualifications and Occupations) classification for abilities and skills for professional figures.

Regarding the internal sources, two macro-categories of data were detected: Candidate and job offer. About candidate, data are present for registry information and previous work experience. On the other hand, about the job offer, the set of requested recruitments are represented for each position in terms of work experience, linguistic knowledge, etc. About the external sources, the database has integrated the following information through the ESCO database and Italian National Collective Labour Agreement contracts. The ESCO Taxonomy is used as a dictionary, to describe, identify and classify professional figures, abilities and qualifications relevant to the European labour market.

Since data are available for a period of four years, from 2018 to 2021 (provisional data until September 2021), the analysis could be repeated for each year in order to find differences in the selected period. Beyond the differences, it is possible to sketch

a defined path over the entire period. This path could be represented from a graphical point of view through the use of a time trajectory. The statistical unit is represented by a person receiving a job, and there were more than 600.000 job positions divided into the following 9 industries: Production and Logistic, Food services, Commercial and Marketing, Human Resources, Legal and Finance, Medical and Pharmaceutics, Engineering, Tourism and Fashion, IT and Digital. In table 1, the distribution of the job positions over the entire period and the industries is displayed.

Table 1 Distribution of the job positions for industry, Italy, 2018-2021

Industry	**2018**	**2019**	**2020**	**2021**
Production and Logistic	136.831	103.973	99.879	92.141
Food services	27.337	23.096	12.085	12.843
Commercial and Marketing	12.708	10.117	7.971	7.889
Human Resources	7.792	6.336	4.153	3.610
Legal and Finance	4.016	4.183	3.240	2.734
Medical and Pharmaceutics	2.275	1.958	2.074	1.310
Engineering	1.734	1.481	1.034	905
Tourism and Fashion	6.309	3.807	1.801	589
IT and Digital	751	685	497	458
Total	**199.753**	**155.636**	**132.734**	**122.479**

Source: elaboration on The AdeccoGroup data

As it is possible to note from the Table 1, some preliminary differences at industry level are present. If the sector with more job offers is Production and Logistic for the entire period, Tourism and Fashion had a clear decrease in last years passing from $3,1\%$ in 2018 to $0,5\%$ in 2021. This could be a clear effect of the health emergency caused by Covid-19. A first research issue could regard whether starting from these differences, it will be possible to detect changes also in terms of skills required for the recruitment process.

Since it represents one of the most in phase of development sectors in Italian labour market with many job title in evolution searching for new skills, this work is focused on IT and Digital industry. This choice is also convenient in order to experiment this technique on a limited numbers of job titles. Among all positions in this sector, the 4 most requested roles have been selected:

- System Analyst
- Software Developer
- ICT Technician
- ICT Help Desk Agent

For each job title, a full description has been reported from the ESCO classification. Systems analysts conduct research, analyse and evaluate client information technology requirements, procedures or problems, and develop and implement proposals, recommendations, and plans to improve current or future information systems.

Software developers research, analyse and evaluate requirements for existing or new software applications and operating systems, and design, develop, test and maintain software solutions to meet these requirements.

Information and communications technology operations technicians support the day-to-day processing, operation and monitoring of information and communications technology systems, peripherals, hardware, software and related computer equipment to ensure optimal performance and identify any problems.

Information and communications technology user support technicians provide technical assistance to users, either directly or by telephone, email or other electronic means, including diagnosing and resolving issues and problems with software, hardware, computer peripheral equipment, networks, databases and the Internet, and providing guidance and support in the deployment, installation and maintenance of systems.

It is possible to note that there is a substantial difference between the first two and the last two positions. Systems analysts and software developers belong to the group of scientific professions specialized in ICT. On the other hand, Information and communications technology operations technicians and information and communications technology user support technicians are part of intermediate technician professions in ICT. This division could led to a strong difference in terms of skills required. In table 2, the frequency distribution of the job positions in the industry is displayed. For each year, most of the 80% of the job offers of the entire industry has been considered in the analysis.

Table 2 Distribution of the job positions over the period and the industries, Italy, 2018-2021

Industry	**2018**	**2019**	**2020**	**2021**
System Analyst	28,2%	35,2%	30,2%	22,3%
ICT Help Desk Agent	27,2%	31,5%	22,3%	11,1%
Software Developer	19,3%	12,4%	9,3%	8,5%
ICT Technician	6,5%	9,3%	23,7%	41,3%
Total	**81,2%**	**88,4%**	**85,5%**	**83,2%**

Source: elaboration on The AdeccoGroup data

For the IT and Digital industry, the analysed requirements have been selected among 26 skills included in the AdeccoGroup competence dictionary. The selection has been achieved only considering the intersection of the top-10 skills for each year for the entire industry. Since the top-10 of the skills in the IT sector is not equal over the period, this skills intersection led to a reduction to 8 competencies. As happened for the selection of the job positions, this reduction do not led to a loss of information covering about the 90% of the total soft skills required. Communication is the most requested soft skill in 2018 and 2020, while Problem solving and analysis is the most present for 2019 and 2021. Result orientation, innovation and customer orientation are over 10% for each year.

Using the proposed dataset, let $I = 4$ the number of professional roles, let $X = 8$ the number of soft skills and $K = 4$ years from 2018 to 2021, it is possible to

Table 3 Distribution of the soft skills in the IT and Digital industry, Italy, 2018-2021

Soft skills	**2018**	**2019**	**2020**	**2021**
Communication	20,8%	14,4%	17,5%	15,5%
Problem Solving and analysis	18,3%	16,0%	17,1%	15,7%
Result orientation	18,2%	14,8%	16,9%	14,9%
Innovation	15,2%	14,1%	16,5%	14,1%
Customer orientation	14,9%	13,5%	15,9%	13,5%
Team working	3,3%	9,3%	5,4%	7,4%
Quality orientation	1,9%	2,9%	1,4%	2,5%
Adaptability	1,3%	3,6%	0,7%	4,1%
Total	**93,9%**	**88,6%**	**91,4%**	**87,7%**

Source: elaboration on The AdeccoGroup data

obtain a reduction of the dimensions using a well-known technique of multivariate analysis. A Weighted Factor Analysis (WFA) has been applied, where the weights are represented by the number of the job offers for a job title in a year.

The Weighted Factor Analysis has been conducted on the relative frequency distribution of 8 soft skills for each year and professional figure in order to detect 2 latent components grouping the soft skills. This method allows to achieve 4 possible evidences:

1. Similarities between job figures of the same industry
2. Individuation of cluster of soft skills
3. Association of professional roles with some specific soft skill
4. Evolution of the job figures over the considered period using the time trajectory

In table 4, using the loadings of the WFA, it is represented the contribution of the single competence to the 2 latent factors. Factor 1 is positively correlated with Result orientation, innovation and customer orientation. Factor 2 is positively correlated to communication and problem solving. The fist component of the WFA explains the 51% of the variance. Total variance explained by the two facotrs is 77%.

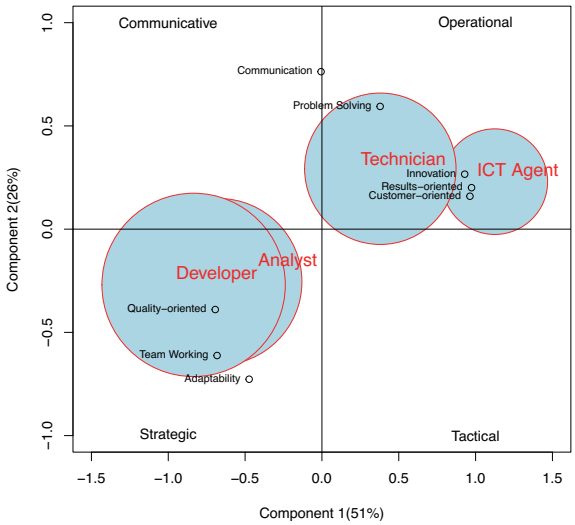
Table 4 Contribution of soft skills to 2 latent factors

Soft skills	**Factor 1**	**Factor 2**
Communication	-	Positive
Problem Solving and analysis	-	Positive
Result orientation	Positive	-
Innovation	Positive	-
Customer orientation	Positive	-
Team working	Negative	Negative
Quality orientation	Negative	-
Adaptability	-	Negative
Explained variance	**50,7%**	**26,6%**

Source: elaboration on The AdeccoGroup data

The bubble graph in figure 1 represents a static point of view only based on the barycentre of the professional figures not taking into account the trend of the entire time-series. It is possible to note from the graph that clusters of soft skills have been detected. These two groups have been classified on the basis of their contribution to the total variance in WFA. The two clusters have been named as operational and strategic requirements. As expected after the descriptions of professional figures, analysts and developers have a similar request with strategic soft skills. Moreover, analysing the relationship between figures and competencies technicians and ICT Help Desk Agents are more involved in operational soft skills, while communication seems to be a cross skill useful for all the figures in the industry. This kind of visualization does not give a dynamic vision of the phenomenon. To do this, it is necessary to represent on the same Cartesian plane, the time trajectory for each job figure.

Fig. 1 Bubble graph of the WFA for job titles and soft skills

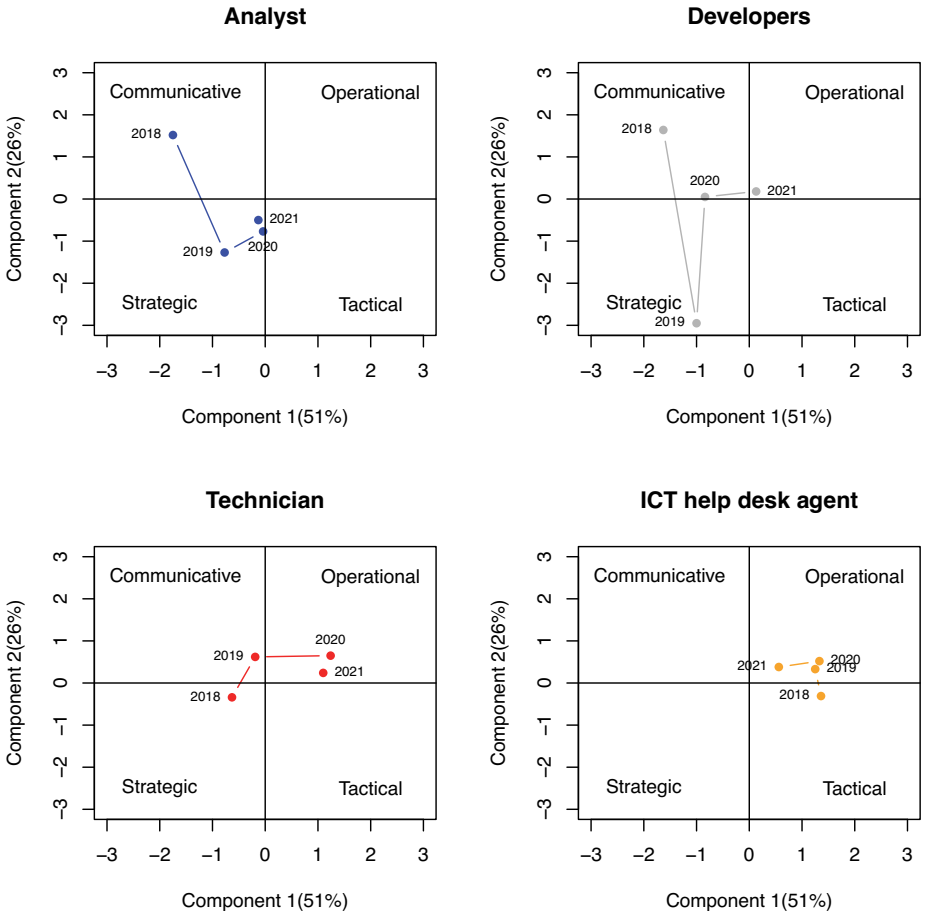


On the basis of the positioning of the soft skills, the Cartesian plane have been divided into four quadrants, the first quadrant has been named operational, the second one as communicative, the third one as strategic and the fourth one as tactical.

On the same plane, the time trajectories for each job figure are displayed in figure 2. Principal evidences from this representation show an overlapping between analysts and developers. They are going from communicative towards operational after a strategic 2019. Technicians have a central trajectory and they are moving in

an operational way, while ICT Help Desk Agents have very small variations in soft skill request and they are travelling in opposite direction.

Fig. 2 Time trajectories of the job positions in IT and Digital sector, Italy, 2018-2021



It is possible to note that even for job offers with a similar barycentre, the use of time trajectories can lead to very different perspectives in terms of movements and directions. This depends on the different composition of soft skill portfolio of the job title over the period.

4 Conclusions and Future research

The time trajectories have been used as an exploratory approach to verify the presence of a soft skill pattern for some professional figures in IT and Digital sector in the Italian labour market. This approach allows to analyse the situation in a twofold way. Firstly, from a static point view, using the barycentre of the trajectories. Secondly, from a dynamic point of view, drawing the trajectories. This method allows to detect similarities between roles and soft skills. In particular analysts and developers seems to have a similar behaviour. On the other hand, technicians and ICT Help Desk Agents show a different pattern. About soft skills, they have been clustered in operational and strategic requirements.

In terms of innovative contribution, this work tried to propose the study of short time series as a statistical tool useful in the decision making issues for candidates for a specialized job position. For example, unemployed individuals with a certified set of soft skills could be addressed by job agencies towards positions where these competencies could be more appreciated.

About future works, an extension of time interval appears to be necessary to lengthen the trajectories. The model could be enhanced also using the personal features of the new hired. The validation of this approach could regard the use of other economic sectors and professional figures or mixing job titles belonging to different industries to verify the existence of transversal skills.

References

1. Coppi, R., D'Urso, P. (2002). Fuzzy K-means clustering models for triangular fuzzy time trajectories. Statistical Methods & Applications, 11, p. 21-40.
2. Coppi, R., D'Urso, P. (2006). Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization. Computational Statistics & Data Analysis, 50, p. 1452-1477.
3. D'Urso, P. (2000). Dissimilarity measures for time trajectories. Journal of the Italian Statistical Society, p. 53-83.
4. D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., (2019). Fuzzy clustering with spatial–temporal information. Spatial Statistics, 30, p. 71-102.
5. Dagsvik, J.K. Random utility models for discrete choice behavior. An Introduction. Statistics Norway Research Department, Norway (1998).
6. Escofier., B., Pagès J. (1994). Multiple factor analysis (afmult package). Computational Statistics and Data Analysis, 18, p. 121-140.
7. International Labour Organization, Skills, knowledge and employability (2018).
8. Jolliffe, I. (2002). Principal component analysis. New York: Springer-Verlag.
9. Jolliffe I., Cadima, J. (2016). Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A 374:20150202.
10. Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. Journal of Chemometrics, 14, p. 105-122.
11. Krantz, D.H. Conjoint measurement: The Luce-Tukey axiomatization and some extensions. Journal of Mathematical Psychology **2**, 248-277 (1964).
12. Kroonenberg, P. M. (1983). Three-mode Principal Component Analysis: Theory and Applications. Leiden: DSWO Press.

13. Lacangellera, M., Liberati, C., Mariani, P. (2011). Banking services evaluation: A dynamic analysis. Journal of Applied Quantitative Methods, 6, p. 3-13.
14. Liberati, C., Mariani, P. (2012). Banking customer satisfaction evaluation: a three-way factor perspective. Advances in Data Analysis and Classification, 6, p. 323-336.
15. Luce, R.D., Krantz, D.H. Conditional Expected Utility. Econometrica **2**, 253-271 (1971).
16. Street, D.J., Burgess, L. The Construction of Optimal Stated Choice Experiments: Theory and Methods. Wiley, New York (2007).
17. World Economic Forum. The future of jobs report. World Economic Forum, Geneva, Switzerland, (2018).

How did the COVID-19 pandemic affect the gender pay gap in EU countries?

Che effetto ha avuto la pandemia da COVID-19 sul differenziale salariale di genere nei Paesi Europei?

Antonella Rocca, Paolo Mazzocchi, Giovanni De Luca, Rosalia Castellano, Claudio Quintano

Abstract The economic crisis caused by the COVID-19 pandemic has yielded dramatic consequences in job losses and firm closures almost everywhere. The first evidence showed a more substantial negative impact on female workers, particularly those with children. However, this impact varied a lot across countries. In this paper, we want to verify the effects of the pandemic on the gender wage gap. At this aim, for a selection of European countries, we compare the levels of the gender wage gap in 2019 and in 2020. For a robust analysis, we propose and compare at this scope the classical Oaxaca-Blinder decomposition and propensity score matching technique.

Abstract *La crisi economica dovuta alla pandemia da COVID-19 ha prodotto gravi conseguenze in termini di perdite di posti di lavoro e chiusure di aziende quasi in tutto il mondo. Le conseguenze economiche della crisi hanno maggiormente interessato alcuni segmenti della popolazione, tra cui, in base alle prime evidenze empiriche, vi sono le donne lavoratrici, specialmente quelle con figli piccoli. In questo articolo si analizzano gli effetti della crisi sul differenziale salariale di genere. A tal fine, si confrontano i livelli di tale differenziale registrati nel 2019 e nel 2020 in una selezione di otto paesi europei. Tale analisi è sviluppata confrontando i risultati di due diverse metodologie: la classica scomposizione del differenziale salariale di Oaxaca-Blinder e la tecnica del propensity score matching.*

Key words: gender pay gap, economic recession, propensity score matching.

[1] Department of Management and Quantitative Studies. University of Naples Parthenope, Italy. antonella.rocca@uniparthenope.it; paolo.mazzocchi@uniparthenope.it; giovanni.deluca@uniparthenope.it; lia.castellano@uniparthenope.it; claudio.quintano@emerito.uniparthenope.it.

1. Introductio

The economic crisis due to the COVID-19 pandemic was very different from the previous financial crises for many reasons. Indeed, it produced an unprecedented health, social and economic downturn, which quickly led to a devastating economic recession [3, 9, 12]. The immediate consequences resulted in layoffs and loss of income, in worsened economic prospects, and, more generally, reduced household consumption and firms' investments. Contrarily to the past economic crises, it provoked a shock both on the supply and demand sides. On the supply side, prolonged lockdowns, business closures, and social distancing caused the slowdown of many productive activities, global supply chain disruptions, and closures of factories [8, 27]. On the demand side, the slowdown hit especially some economic sectors, such as tourism and accommodation and arts and entertainment. In contrast, other sectors, such as food stores, even increased their revenues.

To front this emergency, Governments arranged many more or less effective forms of income support. They introduced various conditions of social restrictions, including the stay-at-home imposition, and many economic activities were suspended or shifted in remote. In many countries, at least for some periods, even the educational activities were converted in remote, provoking a total reorganization of the individual lives. These facts had substantial repercussions also on the work-life balance because if on the one hand, working from home facilitated the reconciliation with housework and child care, on the other hand, the closure of schools and other entertainment activities for children caused relevant problems to workers with children [19].

This crisis was immediate on the GDP, but in the subsequent months, it also generated an increase in the unemployment levels. However, not all the countries suffered the same impact and not all people within each country were hit in the same way.

The previous economic crises, such as the global and financial crisis of 2007-2010, implied, among the other effects, a reduction in gender inequalities. Indeed, at least in a first time, the repercussions are usually stronger in the industrial sector, where more men than women work [2, 21]. Conversely, in the economic recession caused by COVID-19, for many reasons, many economists drew attention to women, renaming the crisis as she-cession [see, among the others, 2].

The motivations for this are manifold. First of all, especially in countries where gender segregation is higher, worker women are usually more concentrated than men in some economic sectors that are more hit by the pandemic (such as the tourism and accommodation sector). Further, in most countries, the incidence of women among workers with temporary contracts and other less protected jobs is higher in comparison to men. Again, women typically take on more childcare and household chores responsibilities. Especially in countries that in 2020 introduced more restrictions to contrast the pandemic, included school closure for prolonged periods, their burden was remarkably stronger [19]. However, on the other side, worker women are usually more concentrated than men in jobs officially classified

as essential, such as health care, education, personal care, and office occupations, not suspended even during the pandemic [13].

In any case, it is reasonable to expect that the crisis by COVID-19 had a strong impact on wages for many workers. In many cases, their wages reduced to the base pay levels, as the payment of allowances was suspended. We expect that the impact on wages was more substantial for workers more involved in housework and child and elderly care [29].

For all these reasons, it is extremely interesting to verify the impact of the COVID-19 pandemic on female and male workers. We expect that the different implications depend mainly on their distribution across the economic sectors and temporary jobs and reflect on wages more than on job losses. We also wonder about the role of the welfare regimes on the effects of the crisis by COVID-19, as during the pandemic, at least for some periods, all forms of care provisions were suspended.

Looking at the EU countries, we observe on the one side the Scandinavian model, usually considered to approximate most closely the 'dual breadwinner' model. On the other side, Continental and Mediterranean countries are still close to the classical "male-breadwinner" model, with lower female participation rates and limited public childcare provision. In an intermediate position, we find a selection of Central European countries, such as Belgium and France. In these latter countries, an extensive system of family-related transfers and childcare provision/subsidies leads to a *defamilisation* model[1] for some aspects near to the Scandinavian one. However, for other aspects mainly related to the mechanisms of taxation of the second earner in the household, these countries show similarities with the other Continental countries, discouraging female work. Finally, the Eastern European countries show different types of welfare regimes due to their different reaction to the fall of the Communist regime. Indeed, while some of them maintained the high levels of female participation rates, such as Lithuania and Latvia, other countries like Romania and Bulgaria adapted more to the Mediterranean model, with low female participation rates and a higher burden for women in child and homecare [28].

Therefore, in this paper, we aim to verify the impact of the COVID-19 pandemic on female participation and on their condition in the labour market in a selection of eight European countries representative of different welfare regimes. In particular, we look at the worsening in the women's condition in the levels of employment and wages, analysing their condition at the end of 2019, just before the pandemic, and at the end of 2020. We focus on the gender wage gap. One of the novelties of this study consists in the methodology used to estimate the gender wage gap: besides the classical Oaxaca-Blinder decomposition, we applied the propensity score matching technique. This latter overcomes most of the critics connected to the first one, and represents an innovative way of studying this phenomenon.

The rest of the paper is organized as follows: Section 2 shows the framework of analysis and is finalized to comprehend what happened in the labour market in 2020

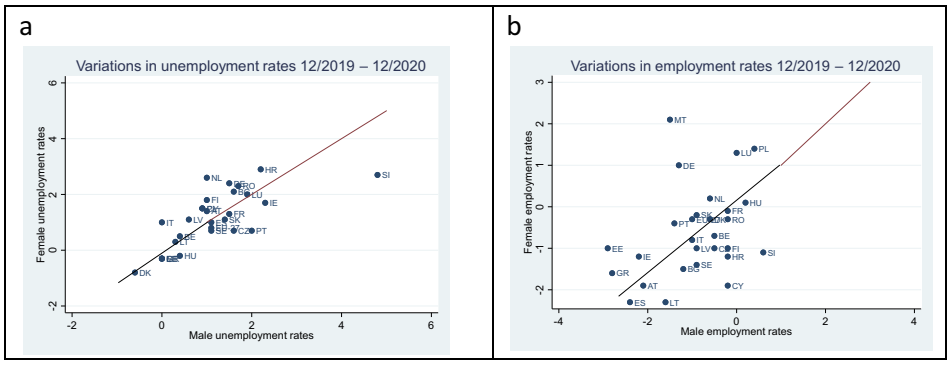
[1] The term defamilisiation has been recently introduced in the economic literature on the gender gap to indicate how the welfare state facilitates female autonomy and economic independence from the family [5].

from a gender perspective. Section 3 discusses the countries' choices, data and methodology. Section 4 shows the analysis results and, finally, Section 5 concludes.

2. The COVID-19 and the gender gap in the labour market

The COVID-19 pandemic produced a sharp decrease in the GDP in almost all EU countries. In 2020, the losses in the GDP growth rate were consistent especially in the Mediterranean countries (from -10.8% of Spain to -8.4% of Portugal). Only in Ireland the GDP growth was positive, and of 5.9%.

The repercussions of the crisis were substantial, even reducing employment rates and the number of hours worked. The majority of Governments tried to contrast the reduction in the worked hours and prevent the layoffs introducing income supports and fiscal measures to sustain enterprises [26]. In some cases, Governments intervened with measures blocking layoffs. However, many jobs were lost, and unemployment rates increased almost everywhere. Figure 1 shows some key indicators of the labour market by gender. The first two indicators (Figure 1a and 1b) are the variations in the unemployment and employment rates, respectively, from the end of 2019 to the end of 2020. At the EU-27 level, the female unemployment rate increased by 0.8%, while the male unemployment rate by 1.1%. Countries where women were more penalized are the Netherlands, Italy, Germany, and Croatia. However, to effectively detect the impact of COVID-19 on the labour market, it is more helpful to refer to the employment rates. Indeed, after the loss of a job, many persons may be transited to the condition of inactivity rather than unemployed during the pandemic. Even in this case, at the EU-27 level, women did not result more penalized than men, as the employment rates decreased by 0.3% and 1%, respectively for women and men. Female employment rates fell mainly in Slovenia, Cyprus, Croatia, Finland, and Sweden. Conversely, in Malta, Germany, Estonia, and Luxembourg, women were less penalized than men. However, if we look at the variation from the end of 2019 to the end of 2021 (Figure 1c), the unemployment rate for men decreased by 1.2% while for women only by 0.50, highlighting a less attitude to the economic recovery for women.



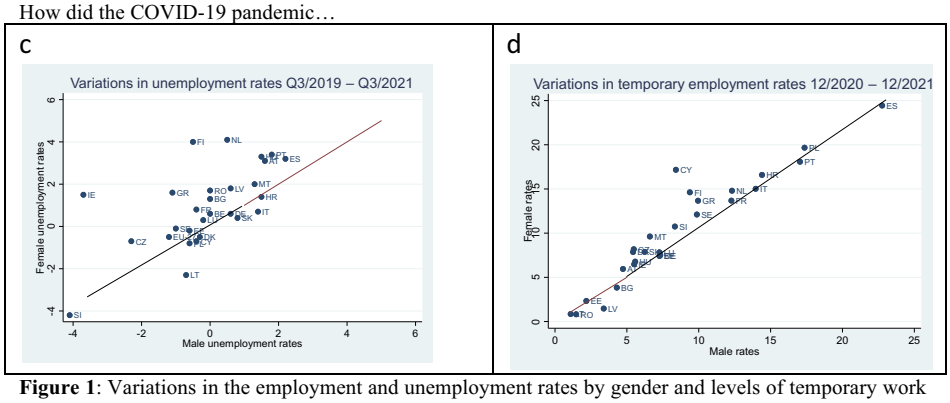


Figure 1: Variations in the employment and unemployment rates by gender and levels of temporary work in EU countries.
Source: Authors ad hoc elaborations on Eurostat data (Eurostat on line database) and Labour Force Survey.

Finally, Figure 1d compares the variations in the share of temporary workers. This is an essential indicator of precariousness and unstable position in the labour market, and shows that, everywhere, except for some Eastern countries (Romania, Bulgaria, Latvia, and Lithuania), the share of women working with a temporary contract is higher than the share of men. The differences are higher, especially in Greece, Malta, and the Czech Republic[1].

3. Data and methodology

3.1. Data

The empirical analysis uses EU-SILC data, the survey on income and living conditions collecting relevant socio-economic information both at the household and individual level. It is based on nationally representative probability samples, including information on the professional status of each family component over the age of 16 and retrospective questions able to reconstruct the educational and professional history of individuals, as well information on the household and the family's financial conditions. Our analysis only focuses on employees aged 15-64 years. We excluded self-employed because their income is not strictly dependent on the number of hours worked. According to the prevalent literature, for measuring the gender wage gap, we referred only to individuals working in enterprises with more than ten employees [15].

[1] Other analyses involved the transition matrixes from the worker status in 2019 to any other status in 2020 and the variation in the number of hours worked on average per week. Even if they are not reported in this version of the paper for brevity, these analyses highlighted that while the share of women passed from employed to unemployed is slightly similar to that of men, systematically more women transited from the status of employed to that of inactive everywhere.

The eight countries selected for the analysis are representative, on the one side, of the different welfare regimes and, on the other side, of the different impact of the pandemic on their economies.

Ireland is representative of the liberal Anglo-Saxon regime and is characterized by a high level of symmetry in the gender roles, even if with low levels of person-specificity, in the sense that it does not consider the different needs of men and women in terms, for example, of child care, because the time spent in paid work is generally high for both men and women.

Sweden for Nordic countries shows high levels of symmetry in the gender roles, representing the classic example of 'dual breadwinner' model, offering workers the opportunity to choose the amount of work spent at work, given the high diffusion of part-time jobs.

France and Austria represent the Continental regime, showing developed Institutions regulating the labour market and services for care needs. In contrast, as all the Mediterranean countries, Greece and Spain have opposite characteristics. However, both Continental and Mediterranean countries highlight a high level of asymmetry in the couple relationship and a system of taxation that does not incentive the second earner, so that the male-breadwinner model still prevails. Finally, as representatives for the Eastern countries, we chose Poland and Lithuania. They show different characteristics in terms of welfare regimes. Indeed, in the transition to a market economy, Poland, such as the Czech Republic and Slovenia, tended to establish a welfare state system similar to that of Mediterranean countries, with low levels of female participation in the labour market and underdeveloped policies for the reconciliation of work and family life [4, 14]. Lithuania presents high levels of female participation in the labour market. However, a welfare system is still not well defined. Hence, the levels of gender equality are under the EU-average [1], and Lithuania appears similar to the other Baltic countries (Estonia, Latvia), Bulgaria, and Romania. These countries show low levels of social protection and even unemployment [20].

3.2. Methodology

The analysis of the gender wage gap is full of contributions in literature. The first pioneering studies go back to Becker [6] and Mincer [23].

The primary indicator for measuring the difference in the income of men and women is the raw or unadjusted gender wage gap, calculated as the difference in the mean income of men and women, divided by the male income. It is usually based on the hourly gross wage in order to control the different number of hours worked by individuals. The choice of the gross income is justified by the need for cross-country comparisons in consideration of the various forms and burdens of wage taxation across countries. The gross hourly wage is usually analysed after the transformation in the logarithmic scale to correct for income asymmetry. However, this measure is defined raw as it does not consider the female and male personal characteristics, which might justify the existence of a wage gap. The economic literature developed

many techniques to control for the observed characteristics of men and women. The Oaxaca and Blinder gender wage gap decomposition is among the most used for its simplicity and because it allows disentangling the part of the gap due to the observed characteristics from the part which remains unexplained and is captured by the different remuneration or rewards that the same characteristics for men and women receive [see 24, 7 for the methodology and, among the others, 10, 11 for the empirical application]:

$$\overline{Y}^M - \overline{Y}^F = \beta^M\left(\overline{X}^M - \overline{X}^F\right) + \left(\beta^M - \beta^F\right)\overline{X}^F \quad (1)$$

In (1), the superscripts M and F stand for male and female; on the right-hand side of the equation, the first term represents the difference in the mean characteristics for males and females, valued at the return rate of male characteristics ("endowment effect", including, however, pre-market discrimination). The second term represents the part of the gap that is due to the different remuneration received by the same characteristics in the two models ("coefficients effect"), valued considering the females' mean characteristics [24, 17].

Even if widely used, in the last years the Oaxaca-Blinder decomposition received many criticisms because it tends to identify as "discrimination" the part of the gap which is not explained by the observed characteristics, despite not all the unobserved aspects may be due to discrimination. Consequently, different alternative non-parametric and semi-parametric methods have been recently used in literature to overcome the issue of possible unobserved heterogeneity [22, 25].

In this paper, in addition the Oaxaca-Blinder decomposition, we employ the propensity score matching technique to examine inequality in pay between men and women. This statistical technique attempts to estimate the effect of a particular condition – being a woman – on a specific outcome – the wage. Its application requires the sample to be split into two groups: the subsample of those who received the treatment (women) and the subsample of those who did not receive the treatment (men). In the first step, we identify the key characteristics connected to being a woman through a logit model. In the second step, the matching algorithm pairs people in the treatment group with people not in the treatment group but whose other variables indicate a high likelihood of being in the treatment group. After that, this technique seeks to establish if a significant statistical difference exists in the outcome variable Y (the logarithm of the gross hourly wage) between the groups of men and women sharing the same observed characteristics. At this aim, the procedure estimates a linear model for the outcome Y on a set of covariates X and the residuals from the binary model (previously estimated) describing the treatment. This method overcomes some of the weaknesses of the parametric models, such as the Oaxaca-Blinder decomposition, because it does not impose the linear function specification and allows to simulate the adjusted mean wage only for the common support population [see 16 and references in it].

Let t denote the random treatment process so that t_i is the treatment received by the i^{th} individual; t=1 is the treatment level (for women) and t=0 the control level (those who have not received the treatment, that is being a man):

$$t=\begin{cases}1 \ if \ w_i'\gamma + \eta_i > 0 \\ 0 \ otherwise\end{cases}$$ is the treatment assignment process

where w is the vector of covariates affecting the probability of receiving the treatment (being a woman), γ is a coefficient vector and η is an unobservable error term not related to X and w. We then proceed to estimate the outcome Y as conditional to a number of covariates supposing to influence it. This allows us to calculate the following two measures:

$$ATE = E(Y_{1i} - Y_{0i}) \equiv E(\beta_i) \quad (2)$$

It is the pay gap between the treated and the untreated groups, that is the average effect of the treatment in the population (gender wage gap)

$$ATET = E(Y_{1i} - Y_{0i}|t=1) \equiv E(\beta_i|t=1) \quad (3)$$

It is defined Average Treatment Effect for the Treated (ATET) and represents the pay gap between the treated group (females) and the control one (counterfactual). It measures the difference between the average outcome for the treated group and the average theoretical outcome of the control group in the hypothesis that this latter receives the treatment. In other words, this latter is the outcome for men with the same characteristics as women in the hypothesis that they were women. In the absence of systematic differences between males and females, ATET should be near zero. Therefore, the ratio ATET/ATE measures the part of the gender gap not due to the observable characteristics but only to the effect of being in the treated or untreated group.

4. Results

The variables considered in the analysis of the gender wage gap include individual characteristics (marital status, level of education), information related to their job (number of years of experience, the economic sector and the professional qualification, the type of contract), and the place of residence (NUTS1 region and degree of urbanization). Work experience is considered for age classes in the propensity score method and in years in the Oaxaca-Blinder decomposition. In this latter, as it is based on extensions of Mincerian equations, besides the years of work experience, we also considered the squared years of work experience. Table 1 shows the raw gender pay gap and the decomposition through the Oaxaca-Blinder technique in its endowments and remuneration parts.

Table 1: Gender gap and its decomposition through the Oaxaca-Blinder technique for the years 2019 and 2020. The sample includes only individuals working in enterprises with 10 or more employees.

Countries	***Raw Gender wage gap***		***Adjusted gender gap***	
	$(\bar{Y}^M - \bar{Y}^F)/\bar{Y}^M$		$(\bar{Y}^M - \bar{Y}^F)$	
	2019	2020	2019	2020
Austria	0.0590	0.0553	2.90-2.73=0.17	2.98-2.81=0.17
% due to endowments			49.3	23.7
% due to discrimination			50.7	76.3
France	0.0510	0.0684	2.73-2.59=0.14	3.03-2.83=0.21
% due to endowments			26.7	38.2
% due to discrimination			73.3	61.8

Greece	0.0399	0.0396	2.02-1.94=0.08	2.09-2.01=0.08
% due to endowments			-85.9	-28.3
% due to discrimination			185.9	128.3
Ireland	0.0342	0.0301	3.06-2.96=0.10	2.88-2.80=0.09
% due to endowments			8.0	-26.3
% due to discrimination			92.0	126.3
Lithuania	0.1280	0.0956	1.55-1.35=0.20	1.89-1.71=0.18
% due to endowments			-57.2	-17.9
% due to discrimination			157.2	117.9
Poland	0.1057	0.0873	1.58-1.41=0.17	1.65-1.51=0.14
% due to endowments			-13.5	-15.9
% due to discrimination			113.5	115.9
Spain	0.0620	0.0661	2.22-2.08=0.14	2.36-2.21=0.16
% due to endowments			-73.1	-29.2
% due to discrimination			173.1	129.2
Sweden	0.0249	0.0633	2.79-2.72=0.07	2.88-2.69=0.18
% due to endowments			57.9	-3.8
% due to discrimination			42.1	103.8

Source: Authors ad hoc elaborations on EU-SILC data.

The wage gap increased in France, but above all in Sweden. Conversely, it remained almost stationary in Austria, Greece, Spain, and Ireland, while in Lithuania and Poland it decreased. Sweden registered even the highest increase of the unexplained part, passing from 42.1% to 103.8%.

Overall, results from propensity score (Table 2) confirm the Table 1 outcomes. ATET, that is the part of the gender wage gap due only to the "treatment" (being a woman), increased in Austria, but above all in Sweden. The proportion between ATET and ATE increased in Austria, Greece, Lithuania, and Sweden, and above all in Ireland (from 105% to 142%) showing a worsening of the female condition not due to their characteristics. Therefore, even controlling for the unobserved heterogeneity through the propensity score, we observe that the pandemic worsened the women's condition, especially in countries where the gender wage gap was lower. This outcome applies even to Ireland, which did not interrupt its economic growth in 2020.

Table 2: Propensity score. Employees working in enterprises with more than 10 employees.

Countries					95% Confidence interval		ATET/ATE %
Austria		Coeff	SD	z	Lower	Upper	
2019	ATE	-0.1401	0.0575	-2.44***	-0.2527	-0.0274	98.72
	ATET	-0.1383	0.0340	-4.07***	-0.2048	-0.0717	
2020	ATE	-0.1705	0.0270	-6.31***	-0.2234	-0.1175	99.24
	ATET	-0.1692	0.0274	-6.18***	-0.2229	-0.11155	
France							
2019	ATE	-0.1113	0.0135	-8.23***	-0.1378	-0.0848	117.34
	ATET	-0.1306	0.0142	-9.22***	-0.1584	-0.1028	
2020	ATE	-0.1262	0.0145	-8.73***	-0.1545	-0.0979	98.81
	ATET	-0.1247	0.0174	-7.16***	-0.1588	-0.0905	
Greece							

2019	ATE	-0.1357	0.0204	-6.64***	-0.1757	-0.0956	89.31
	ATET	-0.1212	0.0167	-7.26***	-0.1539	-0.0885	
2020	ATE	-0.1313	0.0261	-5.03***	-0.1824	-0.0801	98.48
	ATET	-0.1293	0.0181	-7.13***	-0.1648	-0.0937	
Ireland							
2019	ATE	-0.1108	0.0276	-4.01***	-0.1649	-0.0566	105.42
	ATET	-0.1168	0.0309	-3.78***	-0.1774	-0.0562	
2020	ATE	-0.1100	0.0358	-3.08***	-0.1801	-0.0399	142.55
	ATET	-0.1568	0.0460	-3.41***	-0.2468	-0.0667	
Lithuania							
2019	ATE	-0.2156	0.0328	-6.58***	-0.2798	-0.1514	131.68
	ATET	-0.2839	0.0399	-7.12***	-0.3621	-0.2058	
2020	ATE	-0.1627	0.0346	-4.70***	-0.2306	-0.0948	136.32
	ATET	-0.2218	0.0365	-6.07***	-0.2934	-0.1502	
Poland							
2019	ATE	-0.1974	0.0181	-10.91***	-0.2329	-0.1619	118.59
	ATET	-0.2341	0.0203	-11.51***	-0.2740	-0.1943	
2020	ATE	-0.1953	0.0583	-3.35***	-0.3096	-0.0811	111.11
	ATET	-0.2170	0.0167	-12.97***	-0.2499	-0.1842	
Spain							
2019	ATE	-0.2497	0.0273	-9.13***	-0.3033	-0.1961	113.98
	ATET	-0.2846	0.0377	-7.56***	-0.3584	-0.2108	
2020	ATE	-0.1995	0.0181	-11.03***	-0.2350	-0.1641	107.12
	ATET	-0.2137	0.0237	-9.03***	-0.2601	-0.1673	
Sweden							
2019	ATE	-0.0697	0.0340	-2.05**	-0.1364	0.0031	81.20
	ATET	-0.0566	0.0437	-1.29	-0.1423	0.0291	
2020	ATE	-0.2019	0.0508	-3.98***	-0.3015	-0.1024	93.11
	ATET	-0.1880	0.0272	-6.92***	-0.2413	-0.1348	

Source: Authors ad hoc elaborations on EU-SILC data.

5. Conclusions

The COVID-19 pandemic produced an unprecedented socio-economic crisis everywhere. Focusing on eight European countries, in this paper, we tried to verify if and the extent to which women were more severely hit than men in the labour market, both in terms of job losses and wages. Results confirm only partially the so-called she-cession. Women prevail among temporary workers and during the pandemic the transitions from employed to another status were more frequent among them. Therefore, what emerges is that even if, overall, the decrease in the employment rates was similar for men and women, the women's working conditions appear more unstable almost everywhere. After all, even in countries with high levels of gender inequality, a major penalty for women in terms of job losses and wages can be expected for at least two reasons. On the demand side, because worker women are more likely than men to work in many of the most hit sectors by the pandemic (such as hospitality, travel, personal care, cleaning, etc.) and for their major precariousness conditions. On the supply side, the school closures, the stops in the services for children entertainments (sports activities, etc.), in the paid

cleaning services, produced a higher engagement of women in domestic tasks and childcare, given their higher attitude to reduce worked hours for these reasons. This was due to the still consolidated gender norms, but also to the higher opportunity-cost of men giving up paid work, as men usually earn more than women.

Our analysis shows the consistent increase in the gender pay gap in Sweden, while it remained almost stationary in the Mediterranean countries. This could signal that the main driver for its increase was the higher opportunity costs, rather than the gender norms, more settled in the Mediterranean countries.

About the unexplained part of the gender wage gap, results from both the Oaxaca-Blinder and the propensity score methods showed that it increased in Sweden, but even in Ireland although the Irish economy was the only one that in 2020 continued its growth. Conversely, for Austria and France, an increase in the unexplained part of the gender wage gap emerges only from the Oaxaca-Blinder decomposition. In these last two countries, in any case, the gender wage gap, at least in part, is motivated by a lower human capital endowment for women, on average, as highlighted by the explained part of the gender wage gap that in 2020 remained positive only for these two countries.

In conclusion, our results align with those of the currently increasing literature on the gender wage gap in the years of economic crisis by COVID-19 [18, 13]. They found that labour market outcomes of men and women were roughly equally affected in terms of job losses and furloughing. However, it is evident that even looking at the results of studies on unpaid work [21], the women's condition in the labour market worsened more than that of men, both for their general condition as workers and for the work-life balance. Future development of research will concern the identification of the specific mechanisms that led to these results, taking for example in consideration the amount of the gender wage gap due to the gender segregation.

References

1. Aidukaite, J.: Transformation of the welfare state in Lithuania, Communist and Post-Communist Studies, 47(1), 59-69 (2014)
2. Alon, T., Doepke, M., Olmstead-Rumse, J., Tertilt, M.: The impact of COVID-19 on gender equality, NBER Working Paper Series, 26947 (2020)
3. Antipova, A.: Analysis of the COVID-19 impacts on employment and unemployment across the multi-dimensional social disadvantaged areas, Social Sciences & Humanities Open, 4(1), 100224 (2021)
4. Aspalter, C., Jinsoo, K., Sojeung, P.: Analysing the Welfare State in Poland, the Czech Republic, Hungary and Slovenia: An Ideal-Typical Perspective, Soc. Policy Admin., (43)2, 170-185 (2019)
5. Bambra, C.: Defamilisation and welfare state regimes: a cluster analysis, Int. J. Social Welfare, 18th September (2007)
6. Becker, G.: Human Capital – a theoretical and empirical analysis with special reference to education, 3rd ed., Chicago University Press (1964)
7. Blinder, A.S.: Wage discrimination: reduced forms and structural estimates. J. Hum. Resour. 8, 436–455 (1973)
8. Bodnàr, K., Le Roux, J., Lopez-Garcia, P., Szorfi, B.: The impact of COVID-19 on potential output in the euro area, ECB Economic Bulletin, 7 (2020)
9. Borio, C.: The Covid-19 economic crisis: dangerously unique, Business Economics, 55, 181-190 (2020)

10. Castellano, R., Rocca, A.: Gender gap and labour market participation: a composite indicator for the ranking of European countries. Int. J. Manpower. 35(3), 345–367 (2014)
11. Castellano R., Rocca, A.: The dynamic of Gender Gap in European Labour Market in the years of economic crisis, Quality & Quantity, 51(3), 1337-1357, 28 March, DOI 10.1007/s11135-016-0334-1, (2016)
12. Chi-Wei, S., Ke, D., Sana, U., Zubaria, A.: COVID-19 pandemic and unemployment dynamics in European economies, Economic Research-Ekonomska Istraživanja, (2021)
13. Del Boca D., Oggero, N., Profeta, P., Rossi, M.: Women's and men's work, housework and childcare, before and during COVID-19, Review of Econ. of Hous., 18, 1001-1017. (2020)
14. European Parliament: The policy on Gender Equality in Poland, Update, PE 571.372, Brussels (2016)
15. Eurostat: Gender Pay Gaps in the European Union: a statistical analysis, Statistical Working Papers, Luxembourg (2021)
16. Frölich, M.: Propensity score matching without conditional independence assumption – with an application to the gender wage gap in the United Kingdom, Econometrics Journal, Royal Economic Society, 10(2), 359-407 (2007)
17. Heckman, J.J., Lance, J.L., Petra, E.T.: Fifty years of Mincer Earnings Regressions, N. w9732, National Bureau of Economic Research, Cambridge (2003)
18. Hupkau, C., Petrongolo, B.: Work, Care and Gender during the COVID-19 Crisis. IZA DP 13762 (2020)
19. ILO: Teleworking during the COVID-19 pandemic and beyond A Practical Guide, Geneva, (2020)
20. Lauzadyte-Tutliene A., Balezentis, T., Goculenko, E.: Welfare State in Central and Eastern Europe. Economics and Sociology, 11(1), 100-123 (2018)
21. Mascherini, M., Nivakoski, S.: Gender Differences in the Impact of the COVID-19 Pandemic on Employment, Unpaid Work and Well-Being in the EU, Intereconomics, (56)5, September/October, (2021)
22. Meara, K., Pastore, F., Webster, A.: The gender pay gap in the USA: a matching study, J. Pop. Economics, 33, 271-305 (2020)
23. Mincer, J.A.: Schooling, experience and earnings, National Bureau of Economic Research, New York (1974)
24. Oaxaca, R.: Male-female wages differentials in urban labor markets, Int. Econ. Review, XIV(3), 693-709 (1973)
25. Obermann G., Hoang Oanh, N., Hong Ngoc, N.: Gender pay gap in Vietnam: a propensity score matching analysis, J. Econom. Develop., 23(3), 238-253 (2021)
26. OECD: The Territorial Impact of COVID-19: Managing the Crisis and Recovery across Levels of Government, May, OECD, Paris, (2021)
27. Pak, A., Adegboye, O.A., Adekunle, A.I., Rahman, K.M., McBryde, E.S., Damon, P.: Front Public Health, May 29, (2020)
28. Pascall, G., Lewis, J.: Emerging gender regimes and policies for gender equality in a wider Europe. J. of Social Policy, 33(3), 373-394 (2004)
29. Tverdostup, M.: Gender gaps in employment, wages, and work hours: Assessment of COVID-19 implications, WP n. 2020, The Vienna Institute for International Economic Studies, Vienna (2021)

Skill Similarities and Dissimilarities in Online Job Vacancy Data across Italian Regions

Similarità e dissimilarità fra le regioni italiane nelle skill richieste nei Job Vacancy Data

Adham Kahlawi, Lucia Buzzigoli, Laura Grassini, Cristina Martelli[1]

Abstract In European countries there is a growing interest in integrating traditional statistical sources on the labour market with online job vacancy data as they offer detailed and timely information on the use of the Internet for job vacancies and on the specific skills required at different levels (in particular, at a territorial and sectoral level). In this context, the work proposes an analysis of the similarity between the Italian regions in terms of required skills. The study looks at a specific group of innovation-related occupations that are believed to be well represented by online data. The results highlight a regional gap in the use of online offers and in the description of professional profiles in terms of required skills.

Abstract *Nei Paesi Europei vi è un interesse crescente nell'integrazione delle fonti statistiche tradizionali sul mercato del lavoro con i dati sulle offerte di lavoro online in quanto offrono informazioni dettagliate e tempestive sull'uso di Internet per le offerte di lavoro e sulle specifiche competenze richieste a diversi livelli (in particolare, a livello territoriale e settoriale). In questo quadro, il lavoro propone un'analisi della similarità fra le regioni italiane in termini di competenze richieste. Lo studio prende in esame uno specifico gruppo di occupazioni legate all'innovazione che si ritiene siano ben rappresentate dai dati online. I risultati evidenziano un divario regionale nell'uso delle offerte online e nella descrizione dei profili professionali in termini di competenze richieste.*

Key words: Labour market, Job ads data, Occupations, Skills, ESCO.

[1] Adham Kalhawi, Department of Statistics, Computer Science, Applications, Università di Firenze. Email: adham.kahlawi@unifi.it
Lucia Buzzigoli: Department of Statistics, Computer Science, Applications, Università di Firenze. Email: lucia.buzzigoli@unifi.it
Laura Grassini: Department of Statistics, Computer Science, Applications, Università di Firenze. Email: laura.grassini@unifi.it
Cristina Martelli: Department of Statistics, Computer Science, Applications, Università di Firenze. Email: cristina.martellli@unifi.it

1 Introduction

Online Job Vacancy data (OJVs) has recently received growing attention in the study of the labour market (Beręsewicz and Pater, 2021). It is well known that this data may be affected by some potential risks and biases widely explored and discussed in the statistical literature (Giambona et al., 2021). Nevertheless, OJVs offer valuable and timely insights on the use of the Internet for job offers and on job-specific skills requirements at different levels (for instance, territorial or sectoral). For the EU Member States, valuable contributions on the use of OJV data are by CEDEFOP reports (2018, 2019a, 2019b), while in Italy, OJVs have been monitored since 2013 by Wollybi (Boselli et al., 2018).

One of the most important datasets that can be used to analyse OJVs is the one produced by Burning Glass Technologies[1] (BGT); the file contains millions of online job postings collected by scanning daily thousands of Internet sources (dedicated job portals and company websites). The data are collected with various methods (API, scraping, crawling) based on the web portal characteristics and are subjected to a data cleaning process to remove noise, outliers, and duplicate entries (Mezzanzanica & Mercorio, 2019). The content of the ads is coded using text classification algorithms referring to the official classifications used in the various countries for describing the job positions.

Our contribution deals with 2019 BGT data for Italy, collected from 239 online job portals. The total number of ads is more than 1.7 million. They contain about 70 variables, most of them referred to official classifications (shown in brackets in the following): opening and closure date of publication, identification and description of occupation and related skills (ESCO classification[2]), geographic job location (LAU and NUTS), the economic activity of the company (ATECO2007), educational level (ISCED).

The paper's main objective is to explore the similarity between the occupational profiles needed for the various regions: this is possible because the BGT dataset contains both the territorial information and a detailed description of the skills requested by each job ad.

The relative analysis will be carried out for the ESCO 2-digits (in the following ESCO-2) group of occupations 25 (Information and communications technology professionals). The interest in this group of occupations is motivated by three reasons. Firstly, this group of occupations are considered to be well captured by OJV data (Turrell et al., 2018): therefore, we expect that the OJVs could well represent the labour demand for those job positions also at the territorial level. Secondly, as we will see in section 2, this group is the second greatest one for the number of job ads and has the highest number of skills requested in the job ads. Thirdly, this group of occupations includes professional profiles related to innovation, and therefore it can be considered a proxy for the trend in innovation in the local labour market.

[1] Source: Burning Glass Technologies. burning-glass.com. 2021.

[2] https://esco.ec.europa.eu/en/home

The use of OJV data for territorial comparisons is not frequent because of their well-known inherent characteristics. In particular, the different use of the Internet in job applications due to the different levels of digitisation of regions or territories determines problems of representativeness.

For these reasons, regional analyses of OJV data are not numerous. Among recent contributions, we can mention Turrell et al. (2018), who studied the UK labour market and Cedefop (2019), highlighting the differences across economic activities. Moreover, Giambona et al. (2021) used BGT data to study the skill change between 2019 and 2020 at the level of the Italian regions.

The paper is organised as follows. Section 2 presents a descriptive analysis of BGT data to emphasise the dimension of the original data in terms of the number of job ads, occupations and requested skills at the national and regional levels. In section 3, the paper proceeds with a multivariate approach to explore any similarities among skill profiles for the occupation of the 25 ESCO-2 already mentioned. Such similarities are based on an index of skill importance which is specifically computed for the ESCO 4-digit code (in the following ESCO-4) occupations included in group 25. The problem of sparse matrices is addressed. Finally, the last section presents some concluding remarks.

2 Descriptive analysis

Table 1 presents some basic information about quantitative results for the 2019 BGT data: the total number of OJVs (that is, OJVs with complete data in occupation code, requested skills for occupation, and region), the number of different ESCO-2 and ESCO-4 occupations, the number of different skills covered by the data.

Table 1: Main figures of 2019 BGT data

# Job-ads	# ESCO-2 occupations	# ESCO-4 occupations	# Skills
1,078,327	37	326	1,200

Table 2 describes the distribution of OJVs by economic activity and – as expected – shows the prevalence of job ads in Manufacturing (22.9%) and service sectors (Administrative and support services, 19.3%; Professional, scientific and technical activities, 15.5%; Wholesale and retail trade, 12.5%).

Table 3 shows that the distribution across regions (NUTS2) of job ads and skills requested is somewhat heterogeneous. In contrast, the average number of skills by job ad ranges from 9 to 14 with no extreme values. Note that the job ads from just three regions cover 55% of overall job ads (Lombardia 29%, Veneto 14% and Emilia Romagna 13%).

Table 2: Number of job ads by economic activity

Economic activity	# Job Ads	%
Agriculture	1,228	0.12
Mining	308	0.03
Manufacturing	237,256	22.93
Electricity and gas	10,013	0.97
Water and recycling	451	0.04
Construction	8,310	0.80
Wholesale and retail trade	129,701	12.54
Transportation and storage	48,064	4.65
Accommodation and food	44,167	4.27
Information and communication	85,261	8.24
Finance and insurance	19,664	1.90
Real estate	6,249	0.60
Prof., sci., tech services	160,205	15.48
Admin., support services	199,221	19.25
Public administration	8,908	0.86
Education	18,238	1.76
Health	31,163	3.01
Arts, entertainment and recreation	7,680	0.74
Other service activities	18,066	1.75
Other	555	0.05
Total	**1,034,708**	**100.00**
Missing economic activity	*43,619*	
Total	**1,078,327**	

The primacy of Lombardia and Veneto is maintained even when we divide the number of job ads by the size of the labour force of the relative region to remove the dimensional effect. In this case, Friuli Venezia Giulia surpasses Emilia Romagna. The ratio values show the typical Italian divide between the North macro-region[1] and the others because only the Northern regions (the only exception is Liguria) exhibit values over the national value.

Table 4 presents the three ESCO-2 occupations with the greatest number of job ads and requested skills: the occupation group 25 (*Information and communications technology professionals*), which will be the subject of subsequent analyses, is included in both rankings. Note that group 25 contains 9 ESCO-4 occupations (out of 326: 2.8%) and 620 different skills (out of 1,200: 51.7%).

All this data confirms the peculiarities of such an occupational group that includes high skilled jobs, whose production processes are defined at a high granularity level.

[1] The NUTS1 level areas 1-North-East (Piemonte, Val d'Aosta, Liguria, Lombardia)), 2-North-East (Trentino A.A., Vento, Friuli Venezia Giulia, Emilia Romagna), 3-Center (Toscana, Umbria, Marche, Lazio), 4-South (Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria), 5-Islands (Sicilia, Sardegna) have been grouped in three macro areas: North (levels 1 and 2), Center (level 3), South (levels 4 and 5).

Table 3: Regional statistics of OJV data

Regions (NUTS2)	# job ads	% job ads	# requested skills	Avg. # skills by job ad	Labour force (thousands)	# job ads / labour force (‰)
Piemonte	83,275	7.72	874,453	10.5	1,981	42.0
Valle d'Aosta	2,677	0.25	25,937	10.5	59	45.2
Lombardia	317,251	29.42	4,082,627	12.9	4,750	66.8
Liguria	22,641	2.10	254,446	11.2	677	33.5
Veneto	146,401	13.58	1,925,786	11.2	2,297	63.7
Trentino A.A.	26,542	2.46	284,670	13.2	520	51.1
Friuli V.G.	34,624	3.21	315,826	9.1	545	63.6
Emilia R.	136,066	12.62	1,469,480	10.8	2,152	63.2
Toscana	68,799	6.38	783,907	10.8	1,718	40.0
Umbria	12,268	1.14	134,240	11.4	396	30.9
Marche	27,356	2.54	282,023	10.3	696	39.3
Lazio	67,748	6.28	946,068	14.0	2,649	25.6
Abruzzo	20,926	1.94	223,569	10.7	561	37.3
Molise	2,770	0.26	29,772	10.7	124	22.3
Campania	33,730	3.13	420,268	12.5	2,060	16.4
Puglia	25,288	2.35	303,828	10.7	1,450	17.4
Basilicata	6,595	0.61	74,020	11.2	213	31.0
Calabria	12,343	1.14	144,012	11.7	697	17.7
Sicilia	19,041	1.77	201,803	11.7	1,705	11.2
Sardegna	11,986	1.11	128,555	10.6	692	17.3
Total	**1,078,327**	**100.00**	**12,905,290**	**10.7**	**25,941**	**41.6**

Table 4: ESCO-2 occupations with the highest number of job ads and skills.

ESCO-2 code	ESCO-2 label	# job ads	% job ads
33	*Business and administration associate professionals*	145,399	13.48
25	*ICT professionals*	95,867	8.89
52	*Sales workers*	75,610	7.01

ESCO-2 code	ESCO-2 label	# requested skills	% requested skills
25	*ICT professionals*	2,610,205	20.23
33	*Business and administration associate professionals*	1,774,884	13.75
24	*Business and administration professionals*	1,178,197	9.13

3 Skill importance and regional similarities

This second part of the analysis is focused on the ESCO-4 occupations included in the ESCO-2 group 25 *Information and communication technology professionals*. The interest is in assessing whether there are similar skill profiles across regions in job ads referred to this occupation group.

Specifically, we do not refer to the single job ad, but we aggregate the number of job ads with the following indicator of skill importance (*SI*):

$$SI_{R,O,S} = \frac{N_{R,O,S}}{N_{R,O}}$$

where $N_{R,O,S}$ is the number of job ads requiring skill *S* for Occupation *O* (ESCO-4) in region *R,* and $N_{R,O}$ is the number of job ads for occupation *O* in region *R*. $SI_{R,O,S}$ is the proportion of job ads in the region *R,* for occupation *O*¸ requiring skill *S.*

Therefore, we associate the measure of skill importance $SI_{R,O,S}$ to each combination of region, occupation and skill. Each region has its own profile defined by the list of the SIs calculated for each combination of skill×occupation requested in the job ads for that region. Consequently, we build a sparse matrix of SI_s using the combinations occupation-skill as rows and regions as columns. The matrix is 2085×20, and its sparsity is 33.9%.

Then, we train the Collaborative filtering (Bhumichitr et al., 2017; Jiang et al., 2019; Paleti et al., 2021) to obtain the regions factorisation matrix. Indeed, collaborative filtering implements matrix factorisation to determine the relationship between items' and users' entities (in our case, between the combination occupation-skill and region). For matrix factorisation, we use the Alternative Least Squares algorithm (ALS), which is implemented in the Python implicit package and built for large-scale collaborative filtering problems. ALS is doing a pretty good job at solving the scalability and sparseness of the compilation data; it is simple and scales well to enormous datasets.

Finally, we calculate the similarity between regions by applying the cosine formula to the regions factorisation matrix, where the cosine similarity between two vectors A and B is:

$$sim(A,B) = \frac{A'B}{||A||\ ||B||}$$

Consequently, we recode the similarity values in four groups, as shown in Table 5. We see that only 11 regions have at least one similarity greater or equal to 0.3 and that those regions are located in the Northern and Center of Italy. Figure 1 shows the six regions most similar to each other (i.e., having at least one similarity equal to or greater than 0.5). The line width represents the strength of similarity as in Table 5.

Table 5: Number of regions by level of similarity with the row region

Regions with at least one similarity ≥0.3	**Similarity**			
	<0.3	0.3|-0.5	0.5|-0.7	0.7|-|1
Piemonte	11	4	3	1
Lombardia	10	4	1	4
Liguria	18	1	-	-
Veneto	10	4	3	2
Trentino A.A.	13	6	-	-
Friuli-Venezia Giulia	15	4	-	-
Emilia-Romagna	11	3	4	1
Toscana	9	7	3	-
Lazio	10	5	2	2
Campania	12	7	-	-
Puglia	12	7	-	-

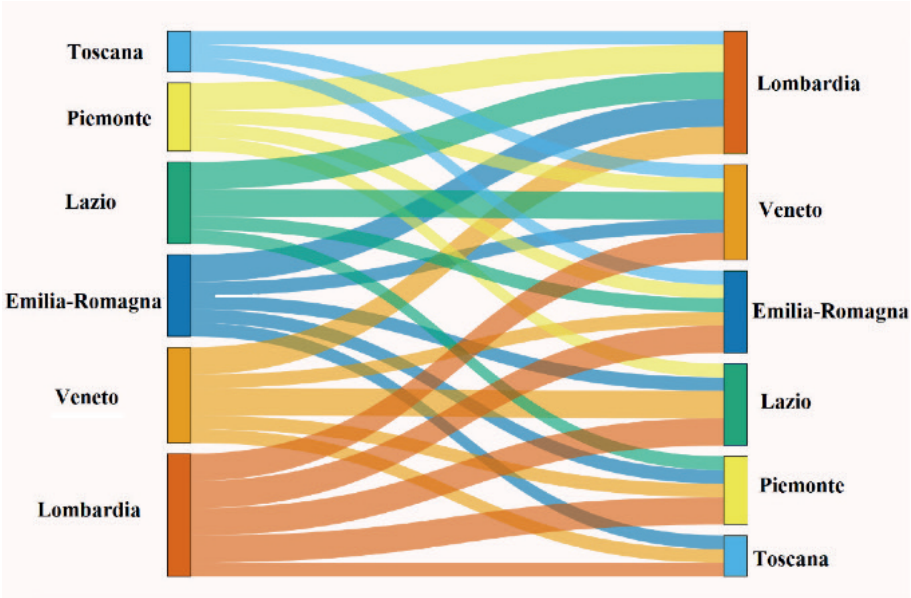


Figure 1: Region similarity ≥ 0.5

4 Conclusions

OJVs are the language companies communicate their employment needs via the Internet. The analysis conducted in this work exploited the specification, in terms of skills, of the required occupations; leveraging this level of detail, we have analysed the similarities between regions.

According to ESCO, the term “skill” refers typically to the use of methods or instruments in a particular setting and in relation to defined tasks; since the skills are described in verbal and functional forms, this specification of the required occupations highlights the production processes in which the candidates will be inserted. In this perspective, therefore, this work does not only describe the similarities between the needs of the labour market but also the similarity between the production processes that require people able to supervise them.

What to say about the regions that are not similar? The reasons may refer to two different orders of explanation: firstly, there could be a diverse policy for recruiting, mainly about the use of the internet channel; this hypothesis, however, was addressed precisely by choosing occupations that are typically sought with this type of media. More convincing is that the production processes present in the two regional contexts are different.

In this perspective, the use of these results can be adopted when a region faces production contexts in which other regions are at a different degree of experience: the employment needs, expressed in similar, more mature contexts are already outlined, and the vocational system can be usefully addressed to target in time any possible job mismatch problem.

References

1. Beręsewicz M., Pater R. (2021), Inferring job vacancies from online job advertisements. Luxembourg: Publications Office of the European Union.
2. Bhumichitr K., S. Channarukul, N. Saejiem, R. Jiamthapthaksin, and K. Nongpong, "Recommender Systems for university elective course recommendation," in 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2017, pp. 1–5, doi: 10.1109/JCSSE.2017.8025933.
3. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2018), Classifying online Job Advertisements through Machine Learning. Future Generation Computer Systems, 86, 319-328.
4. Cedefop (2018), Mapping the landscape of online job vacancies. Background report: Italy, https://www.cedefop.europa.eu/en/publications-and-resources/
5. Cedefop (2019a), Online Job Vacancies and Skills Analysis. A Cedefop pan-European Approach, The European Centre for the Development of Vocational Training, Thessaloniki.
6. Cedefop (2019b), The online job vacancy market in the EU: driving forces and emerging trends. Luxembourg: Publications Office. Cedefop research paper; No 72.
7. Giambona F., Kahlawi A., Buzzigoli L., Grassini L. and Martelli C. (2021), Big Data Analysis and Labour Market: are Web Data Useful to Understand Italian Tendenciens and Regional Gaps, XLII Conferenza Italiana di Scienze Regionali.
8. Jiang L., Y. Cheng, L. Yang, J. Li, H. Yan, and X. Wang, "A trust-based collaborative filtering algorithm for E-commerce recommendation system," J. Ambient Intell. Humaniz. Comput., vol. 10, no. 8, pp. 3023–3034, 2019, doi: 10.1007/s12652-018-0928-7.
9. Mezzanzanica, M.; Mercorio, F. 2019a. Big data for labour market intelligence: An introductory guide (Turin: European Training Foundation). Available at: https://www.etf.europa.eu/sites/default/files/2019-06/Big%20data%20for%20LMI.pdf.
10. Paleti L., P. Radha Krishna, and J. V. R. Murthy, "Approaching the cold-start problem using community detection based alternating least square factorisation in recommendation systems," Evol. Intell., vol. 14, no. 2, pp. 835–849, 2021, doi: 10.1007/s12065-020-00464-y.
11. Turrell A., Thurgood J., Copple D., Djumalieva J. and Speigner B. (2018), Using online job vacancies to understand the UK labour market from the bottom-up, Bank of England, Staff Working Paper No. 742.

Small area estimation methods with socioeconomic applications

Exploring Small Area Estimation techniques to address uncertainty in Spatial Price Indexes

Un'esplorazione delle tecniche di piccola area per la stima dell'incertezza negli indici dei prezzi spaziali

Ilaria Benedetti and Federico Crescenzi

Abstract The availability of scanner data for the compilation of price statistics has increased over the past twenty years and several European Member States have introduced Scanner Data into Consumer Price Index (CPI) production. Besides reducing administrative burden, Scanner Data have proved to be of benefit to CPIs thanks to the higher granularity, the wide coverage, the opportunity to implement superlative index and greater precision. However, in spite of their potential, to the authors' knowledge, only few National Statistical Institutes have started official research project for computing sub-national spatial price indexes (SPIs) using Scanner Data. Given the crucial role of SPIs for comparing standard of living among regions it is also relevant to be able to assess their accuracy. In this study, we explore the use of small area estimation techniques to reduce the uncertainty associated to point estimates of sub-national SPIs which we have been computed via Jackknife Repeated Replications. The data that we use is part of the ISTAT 2018 Scanner Data on the ten provinces of Tuscany (Italy) for selected groups of products.

Abstract *La disponibilità di dati scanner per la compilazione di statistiche sui prezzi è aumentata negli ultimi venti anni e diversi Stati membri europei hanno introdotto i dati scanner nella produzione degli indici dei prezzi al consumo (CPI). Oltre a ridurre l'onere amministrativo, gli Scanner Data hanno dimostrato di essere vantaggiosi per i CPI grazie alla maggiore granularità, all'ampia copertura, all'opportunità di implementare indici superlativi e alla maggiore precisione. Tuttavia, nonostante il loro potenziale, per quanto noto agli autori, solo pochi istituti nazionali di statistica hanno avviato progetti di ricerca ufficiali per il calcolo di indici dei prezzi spaziali subnazionali (SPI) utilizzando i dati Scanner. Dato il ruolo cruciale degli SPI per valutare la disparità territoriale del costo della vita, è importante valutare la loro*

[1] Ilaria Benedetti, University of Tuscia; email: i.benedetti@unitus.it

Federico Crescenzi, University of Tuscia; email: federico.crescenzi@unitus.it

accuratezza della stima della varianza. In questo studio, esploriamo l'uso di tecniche di stima per piccole aree per ridurre l'incertezza associata alle stime della varianza degli SPIs subnazionali calcolati tramite Jackknife Repeated Replications. I dati utilizzati fanno parte dei dati dello Scanner ISTAT 2018 sulle dieci province della Toscana (Italia) per gruppi selezionati di prodotti: Pasta, Caffè e Acqua minerale.

Key words: Scanner data, spatial price indexes, uncertainty.

1 Introduction

The popularity and availability of transaction or Scanner Data - that is electronic point-of-sale price and quantity data collected by retailers - for the compilation of the Consumer Price Index (CPI) has increased over the past twenty years

Switching from traditional surveys to Scanner Data reduces administrative burden for both National Statistical Institutes (NSIs) and retailers and offers new opportunities and challenges for price index calculation especially in the use of expenditure data for constructing product weights within elementary aggregates. Although a number of statistical agencies already integrated the use of Scanner Data into their CPIs, it is worth noting that several European NSIs have been using Scanner Data for replacing on-field collected prices needed for international Purchasing Power Parity (PPPs[1]) computations in the framework of the OECD-Eurostat Program.

Spatial price indexes (SPIs) provide measures of price level differences across countries or across regions within a country and are widely used by researchers and policy-makers for comparing real income, standards of living and consumer expenditure patterns. Several players in the economic and social debate have acknowledged the need of sub-national household consumption PPPs due to the high socio-economic heterogeneity among regions.

The increasing availably of Scanner Data may enable countries to measure price level differences across regions which is essential for assessing regional disparities in the distribution of real incomes and supporting regional policy making (Rokicki and Hewings, 2019). Laureti and Polidoro (2017; 2022) explored the possibility of using scanner data of price data for compiling sub-national SPIs in Italy.

In addition, Scanner Data stimulated various research studies for adopting more developed statistical techniques by using probability sampling design and assessing CPI accuracy (De Gregorio, 2012; Jaluzot and Sillard, 2016). Given the advantages of scanner data in CPI compilation, it is interesting to explore the use of these data for providing accurate point estimates of price differences across space with information about the presence and magnitude of uncertainty (Deaton, 2012; Deaton and Aten,

[1] PPPs are essentially spatial price index numbers (SPIs). The concept of purchasing power parity is used to measure the price level in one location compared to that in another location. More specifically, at international level, purchasing power parities of currencies are defined as the number of currency units of a country that can purchase the same basket of goods and services that can be purchased with one unit of currency of a reference currency (Rao and Hajargasht, 2016). PPPs are calculated for product groups and for each of the various levels of aggregation up to and including Gross Domestic Product (GDP).

2017; Rao and Hajargasht, 2016). Uncertainty in the SPIs comes not only from the choice of the aggregation procedure, but also from the dispersion of relative prices. In countries where a huge heterogeneity in relative prices among regions is observed, PPPs suffer of large uncertainty (Deaton, 2012).

This source of variation induces substantial uncertainty into the price indexes. It is important to note that sampling of representative items for SPIs is often judgmental. The universe of products is structured by selecting representative items within the different categories of the expenditure classification. In practice, NSIs adopt several levels of sampling: location, outlets within locations and item varieties. The use of Scanner Data, which cover all transaction of the modern distribution for grocery products, ensure a probabilistic sampling frame where weights must be used at various level of the index hierarchy so that each part is appropriately represented. The sample of location, outlets, goods and services for which price movements are observed ensure that the prices collected are representative to meet the requirements for the accuracy of the index.

Therefore, Scanner Data may play a crucial role for improving current sampling methods, checking the representativity of the achieved sample and controlling initial sample selection. This paper contributes to the advancement of the literature in SPIs by exploring the issue of evaluating the uncertainty associated to point estimates of sub-national SPIs using as source of data the Italian Scanner Data for the year 2018. To the authors' knowledge, the evaluation of uncertainty among SPIs computed for geographical areas within a country, has not been explored yet.

In order to obtain reliable SPIs estimates across Italian provinces, we have taken into account Small Areas Techniques by using Fay-Herriot model-based estimator. The remainder of this short paper is structured as follows. Section 2 discusses the data. In Section 3 we introduce the formulas to compute SPIs, their variance and the basic Fay-Herriot model for small area estimation. This section also contains the main results of the work. Finally, Section 4 discusses the strong points of the results obtained, their limitations and offers some directions for further research.

2 Italian Scanner Data

In this paper, we use a portion of the Italian Scanner Dataset, provided by ISTAT for the year 2018 coming from modern distribution chains (hypermarkets and supermarkets) for grocery products (packaged food, household and personal care goods). The Italian Scanner Data refers to 16 large-scale retail groups in Italy and 107 administrative provinces of the national territory. The sample of large-scale retail trade outlets is representative of the entire universe of large-scale retail trade hypermarkets and supermarkets and includes 1,781 outlets, of which 510 hypermarkets and 1,271 supermarkets distributed throughout the country.

The sampled outlets are extracted within each of the 888 strata of the universe, which were found to be populated, with probability proportional to the sales turnover

of the previous year. Within each outlet, for each reference identified using global trade item number (GTIN), price is calculated based on turnover and quantities sold

(price = turnover/quantity). In our analysis we used data for each GTIN where the turnover and the provincial quantity were calculated as a weighted sum by using the sample weight. The sample of references is drawn within homogeneous groupings of products corresponding to the markets, which in turn are selected considering their relative weight, calculated in terms of turnover in the previous year. The classification of homogeneous products within markets represents an objective and detectable identification of commodity products shared by industrial and distribution companies.

In our paper, we refer to a portion of this big dataset since we used 2018 for all outlets of the Tuscany region. In this analysis we considered three basic headings[1] (BHs), namely: Mineral water, Coffee and Pasta. The dataset consists in 9,516 annual price quotes from the ten Tuscany provinces concerning 13 outlets. The Italian region of Tuscany is divided in 10 Provinces: Arezzo, Florence, Grosseto, Livorno, Lucca, Massa-Carrara, Pisa, Prato, Pistoia and Siena.

3 Methods and Results

In order to estimate SPIs, we adopt the Eurostat-OECD (2012) method where real weights for items at BH level are not considered. For each pair of provinces, two binary SPIs are calculated: the Laspeyres (P_{jk}^{L}) and Paasche (P_{jk}^{P}) indexes by using expenditure share for each product sold in both compared provinces. By following this procedure each basic heading is provided with a matrix of Fisher SPIs: the Fisher (P_{jk}^{F}) price index have good axiomatic and economic properties (Balk, 1995).

$$P_{jk}^{L} = \frac{\sum_{i \in N_{jk}} p_{ik}\, q_{ij}}{\sum_{i \in N_{jk}} p_{ij}\, q_{ij}} \qquad P_{jk}^{P} = \frac{\sum_{i \in N_{jk}} p_{ik}\, q_{ik}}{\sum_{i \in N_{jk}} p_{ij}\, q_{ik}}$$

$$P_{jk}^{F} = \sqrt{P_{jk}^{L} \times P_{jk}^{P}} \quad (1)$$

With the aim of estimating the variance of the price index in (1), we make use of the standard delete one-PSU at a time Jackknife Repeated Replications (Leaver and Cage, 1997). Each replication is done by eliminating one sample PSU from a particular stratum at a time and increasing the weight of the remaining sample PSUs in that stratum appropriately to obtain an alternative but equally valid estimate to that obtained from the full sample. In the framework of Small Area Estimation methods, we consider the basic area level so called Fay-Herriot model:

$$\hat{\theta}_i = z_i'\beta + b_i v_i + e_i, \quad i = 1, \dots, m \qquad (2)$$

where z_i is a vector of area level covariates, $v_i \sim_{iid} N(0, \sigma_v^2)$ are area level effect independent of sampling errors, $e_i \sim_{iid} N(0, \psi_i)$, b_i is a known positive constant, and

[1] BH is defined as a group of similar well-defined goods or services.

$\hat{\theta}_i$ is a direct estimator of the i-th area parameter θ_i. The Best Linear Unbiased Predictor (BLUP) estimator of θ_i is

$$\tilde{\theta}_i^F = \gamma_i \hat{\theta}_i + (1-\gamma_i) z_i' \tilde{\beta}$$

where $\gamma_i = \sigma_v^2 b_i^2/(\psi_i + \sigma_v^2 b_i^2)$ and $\tilde{\beta}$ is the BLUE of β. As the BLUP estimator depends on the unknown σ_v^2, empirical BLUP (EBLUP) is obtained by replacing σ_v^2 with a proper estimator $\hat{\sigma}_v^2$, therefore the EBLUP estimator of θ_i turns out to be

$$\hat{\theta}_i^F = \hat{\gamma}_i \hat{\theta}_i + (1-\hat{\gamma}_i) z_i' \hat{\beta} \qquad (3)$$

Where $\hat{\beta}$ is the BLUP estimator of β having plugged in the estimator of σ_v^2. EBLUP estimation assumes the area level covariates z_i-s to be measured without error. For this reason, we used administrative data from the official archive of Italian Ministry of Economy and Finance on labour earnings among employees[1]. Tables 1,2,3 show the results for the provinces of Tuscany having taken Florence as reference (Florence=1). All results were obtained using the R package sae (Molina and Yolanda, 2015). It is evident that we obtain the best results in the categories of water (average gain of 25.63%) and coffee (average gain of 31.21%) while we obtain smaller gains in the category of Pasta (average gain of 7.28%). This may be due to that the price of pasta is much less volatile than that of other categories, so yielding to more accurate direct estimates, or that the covariate is less predictive in this category.

Table 1: Fay Herriot estimates of Water BH (Florence =1 as reference).

Province	Direct	MSE	EBLUP	MSE
Arezzo	1.195	0.016	1.143	0.015
Grosseto	0.902	0.044	1.026	0.024
Livorno	1.163	0.059	1.107	0.023
Lucca	1.767	0.105	1.216	0.033
Massa-Carrara	1.456	0.042	1.183	0.024
Pisa	1.055	0.020	1.074	0.017
Prato	0.906	0.012	0.982	0.013
Pistoia	1.009	0.022	1.081	0.026
Siena	0.950	0.027	1.024	0.021

[1] Due to the hierarchical administrative division characterizing Italy (i.e. regions, provinces and municipalities), each municipality is included in a specific province. The availability of the total number contributors as well as the total amount incomes for each municipality enabled us to calculate the average value of income per capita in each province. Further information on the data collected can be found at https://www1.finanze.gov.it/finanze/analisi_stat/public/index.php?search_class[0]=cCOMUNE&opendata=yes

Table 2: Fay Herriot estimates of Coffee BH (Florence =1 as reference).

Province	Direct	MSE	EBLUP	MSE
Arezzo	1.044	0.010	1.013	0.005
Grosseto	0.868	0.005	0.964	0.006
Livorno	1.112	0.017	1.021	0.004
Lucca	1.084	0.016	1.046	0.006
Massa-Carrara	1.071	0.014	1.004	0.005
Pisa	1.071	0.005	1.028	0.005
Prato	0.967	0.012	1.005	0.005
Pistoia	1.021	0.005	1.046	0.007
Siena	1.047	0.009	1.004	0.006

Table 3: Fay Herriot estimates of Pasta BH (Florence =1 as reference).

Province	Direct	MSE	EBLUP	MSE
Arezzo	1.120	0.002	1.085	0.002
Grosseto	0.818	0.002	0.842	0.002
Livorno	1.038	0.001	1.029	0.001
Lucca	1.083	0.003	1.067	0.002
Massa-Carrara	0.961	0.003	0.959	0.003
Pisa	1.024	0.003	1.012	0.002
Prato	0.877	0.001	0.892	0.001
Pistoia	0.968	0.001	0.978	0.001
Siena	0.973	0.002	0.969	0.002

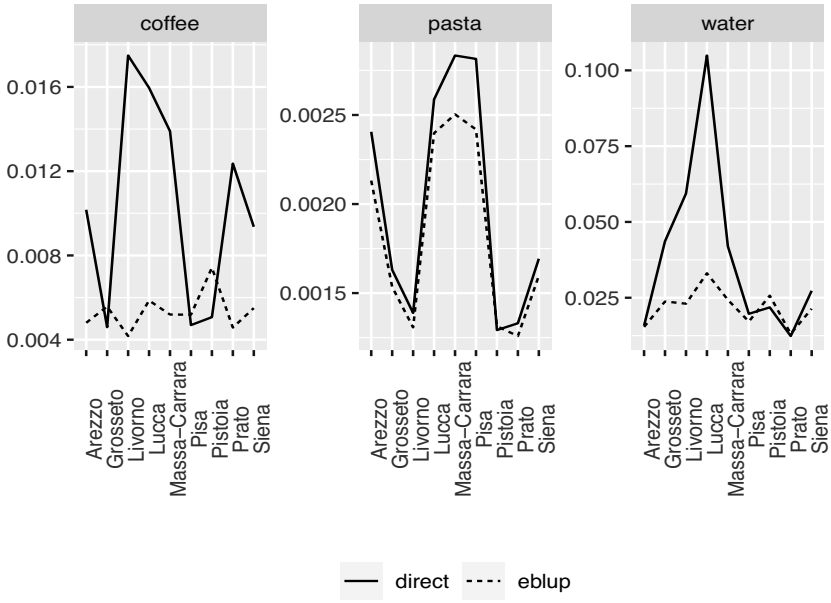


Figure 1: Gain in efficiency (Florence =1 as reference)

4 Conclusions

In this work we have explored the use of small area estimation by estimating an area-level Fay-Herriot model to provide more accurate SPIs estimates than those obtained following a direct approach. To the best of our knowledge, this is the first (and preliminary) attempt to use SAE into a SPI context. Nevertheless, this approach has shown some limitations which we shall address as topics in future research. First, this experiment only used data on the provinces of Tuscany for which had computed the estimates of uncertainty associated to the SPI. The Fay-Herriot model in Formula 2 assumes normality of errors v_i, e_i. Given the few data points and the preliminary nature of this work we have assumed this hypothesis to hold. We are currently working on building a dataset for the whole set of Italian provinces to address the actual validity of these hypotheses in a deeper way as well as other approaches other than the basic Fay-Herriot model. Also, albeit the gains that we obtained for water and coffee were not negligible, it is also true that the variance we have estimated for the direct estimators (particularly that of pasta) were not inadequate. Therefore, we may raise the issue of when it is necessary to improve direct estimates by means of small area estimation methods. With this regard, the literature of variance estimation for SPIs using scanner data has not yet defined any standard and most times SPIs are published without providing any variance estimate.

References

1. Balk, B.M. Axiomatic Price Index Theory: A Survey. International Statistical Review, 63, pp 69–93 (1995)
2. Deaton, A. Calibrating measurement uncertainty in purchasing power parity exchange rates. International Comparison Program (ICP) Technical Advisory Group, Washington, DC, September, pp. 17-18, (2012
3. Deaton, A., Aten, B. Trying to Understand the PPPs in ICP 2011: Why are the Results so Different? American Economic Journal: Macroeconomics, 9(1), pp. 243-64 (2017)
4. de Gregorio, C. Sample size for the estimate of consumer price subindices with alternative statistical designs. Rivista di Statistica Ufficiale (2012)
5. Jaluzot, L., Sillard, P. Sampling of CPI agglomerations for the 2015 base. National Institute of Statistics and economic studies (INSEE). working paper series of the demographic and social statistics directorate. N°F1601 (2016)
6. Laureti, T., Polidoro, F. Testing the use of scanner data for computing sub-national Purchasing Power Parities in Italy. In Proceeding of 61st ISI World Statistics Congress, 16–21 July 2017, Marrakech, Morocco. Available at: https://www.isi-web.org/publications/proceedings (2017)
7. Laureti, T., Polidoro, F. Using Scanner Data for Computing Consumer Spatial Price Indexes at Regional Level: An Empirical Application for Grocery Products in Italy. Journal of Official Statistics (JOS), 38(1), (2022)
8. Molina, I., and Yolanda, M. R package sae: Methodology. The R Journal 7(1), 81-98. (2015)
9. Rao, D. P., Hajargasht, G. Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP). Journal of econometrics, 191(2), pp. 414-425 (2016)
10. Rokicki, B., Hewings, G. J. (2019). “Regional price deflators in Poland: Evidence from NUTS-2 and NUTS-3 regions”. Spatial Economic Analysis, 141, pp. 88-105 (2019). https://doi.org/10.1080/17421772.2018.1503705

Small Area Estimation of Relative Inequality Indices using Mixture of Beta

Stima per Piccole Aree di Indicatori di Diseguaglianza Relativa con Misture di Beta

Silvia De Nicolò and Silvia Pacei

Abstract The paper aims at proposing a small area estimation strategy for the Theil Index, an entropy-based inequality measure. Specifically, we have developed an area-level model of its relative index, i.e. Theil index over its maximum, which has more manageable support between 0 and 1. Classical proposals in area-level context for measures defined on the unit interval are mostly based on proportions modelling and show limitations when dealing with asymmetric heavy-tailed data, such as in our case. We propose a Hierarchical Bayes model with alternative likelihood assumptions based on a particular Beta mixture, providing a more flexible framework.

Abstract *Obiettivo di questo paper è proporre un modello di stima per piccole aree per l'indice di Theil relativo, una misura di diseguaglianza basata sul concetto di entropia e definita sull'intervallo unitario. Ci collochiamo nel contesto dei modelli di tipo "area level" in cui le proposte presenti in letteratura relative a stime in piccole aree di indicatori definiti in (0,1) mostrano limitazioni in caso di stimatori con distribuzione asimmetrica a code alte, come nel nostro caso. Proponiamo, dunque, un modello con assunzioni distributive alternative basate su una particolare mistura di Beta. L'impostazione alla stima che adottiamo è di tipo Bayesiano.*

Key words: Beta Mixtures, Inequality Mapping, Small Area Estimation, Theil Index.

Silvia De Nicolò
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Cesare Battisti 241, 35121 Padova; e-mail: silvia.denicolo@phd.unipd.it

Silvia Pacei
Dipartimento di Scienze Statistiche "P.Fortunati", Alma Mater Studiorum Università di Bologna, Via Belle Arti 41, 40126 Bologna; e-mail: silvia.pacei@unibo.it

1 Introduction

In recent years, we are observing an increasing gap in inequality and social exclusion across EU regions. As a consequence, the demand for reliable estimates of economic inequality measures for small areas is growing due to its importance in better planning public and convergence policies. Their estimation in small areas by using income data from household surveys implies that the number of units sampled at area level is generally not large enough to obtain reliable estimates. Thus, we have to resort to small area estimation techniques, allowing estimators to borrow strength across areas through the use of auxiliary information. See [8] for a comprehensive review. The body of literature concerning the estimation of inequality measures in small areas is very scarce, comprising [3] for Gini Index at area level, [9] for Gini Index and Quintile Share Ratio and [6] for Gini and Theil indexes at unit level.

As opposed to the well known Gini index, the Theil index has the advantage to be strongly transfer sensitive, meaning that it reacts to transfers depending on the donor's (of income transfer) and the recipient's income levels and it is decomposable among groups. Based on the concept of entropy which applied to income distributions has the meaning of deviations from perfect equality, it pertains to the Generalized Entropy family with parameter $\alpha = 1$:

$$GE(\alpha = 1) = \frac{1}{N} \sum_k \frac{z_k}{\mu} \ln \frac{z_k}{\mu}$$

with z_k be a characteristic of interest, in our case income, for the k-th unit of the finite population, where $x_k \in \mathbb{R}^+$, $k = 1, \ldots, N$, and μ its expected value. Since the Theil index is defined between 0 and $\log(N)$, we consider its relative version, namely $RE(1) = GE(1)/\log(N)$ with $GE(1)$ estimated from survey data with a proper weighted estimator and N the true population size. In our estimation strategy, we consider area level models and we adopt a Hierarchical Bayesian approach [8] implemented by using MCMC computational methods.

2 The Flexible Beta Model

In the context of small area estimation of measures in $(0, 1)$, a huge body of literature is dedicated to proportions, implementing Fay-Herriot [8] and Beta regression models, see [5] for a review, with a non-linear linking model. The first solution appears restrictive since it may fit values outside the variable support. On the other hand, Beta regression does not provide enough flexibility when facing heavy-tailed, skewed responses and bimodality. Thus our model proposal involves incorporating an alternative distributional assumption on the likelihood by adopting a Beta mixture-based approach.

Specifically, we implement the Flexible Beta (FB) distribution proposed by [7], a special mixture of two Beta distributions that guarantees great flexibility and at the same time, great tractability. The common dispersion parameter between the components and their ordered arbitrary means leads it to be identifiable in a strong sense. Let us considered the mean-precision parametrization of the Beta distribution [4] such that a generic random variable Beta distributed $Y \sim Beta(\mu\phi, (1-\mu)\phi)$, with $\mathbb{E}(Y) = \mu$ and $\mathbb{V}(Y) = \mu(1-\mu)/(\phi+1)$ with $0 < \mu < 1$ and $\phi > 0$ has probability density function $f_B(y; \mu, \phi)$. The FB distribution has pdf

$$f_{FB}(\lambda_1, \lambda_2, \phi, p) = p \cdot f_B(y; \lambda_1, \phi) + (1-p) \cdot f_B(y; \lambda_2, \phi) \tag{1}$$

with $0 < \lambda_2 < \lambda_1 < 1$ distinct ordered means of the components, $0 < p < 1$ the mixing parameter and $p\lambda_1 + (1-p)\lambda_2$ the expected value. Our small area model proposal for y_d, the direct estimator of Relative Theil index for area d and x_d a set of p generic covariates for m small areas is as follows:

$$\begin{cases} y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p \overset{ind}{\sim} FB(\lambda_{1d}, \lambda_{2d}, \phi_d, p) & \forall d = 1, \dots, m \\ \text{logit}(\lambda_{2d}) = x_d^T\beta + v_d \quad v_d \sim N(0, \sigma_v^2) \end{cases} \tag{2}$$

with $\theta_d = \mathbb{E}(y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p) = p\lambda_{1d} + (1-p)\lambda_{2d}$ the true parameter value and

$$\phi_d = \frac{\theta_d(1-\theta_d) - \mathbb{V}(y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p)}{\mathbb{V}(y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p) - p(1-p)(\lambda_{1d} - \lambda_{2d})}, \tag{3}$$

where the sampling variance $\mathbb{V}(y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p)$ is assumed to be known, as common in literature, in order to allow identifiability.

As opposed to the FB regression proposed by [7], the linear predictor does not model directly the mean parameter but rather a mixture component mean, which in this case can be seen as a pure location parameter. This location-modelling approach unleashes θ_d estimation and speeds up convergence. In order to carry on estimation, the parametrization considered is the following: $y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p \sim FB(\tilde{w}_d + \lambda_{2d}, \lambda_{2d}, \phi_d, p)$ with $\tilde{w}_d = \lambda_{1d} - \lambda_{2d} > 0$. Since estimation requires a variation independent parameter space, we decided to leave λ_{2d}, ϕ_d, and p free to assume any value of their support and to constrain $\tilde{w}_d$, whose range is

$$\left(0, \min\left\{\frac{1-\lambda_{2d}}{p}, \sqrt{\frac{\mathbb{V}(y_d|\lambda_{1d}, \lambda_{2d}, \phi_d, p)}{p(1-p)}}\right\}\right).$$

We model it as $\tilde{w}_d = w \cdot \max\{\tilde{w}_d\}$, with w defined on the unit interval and common to all areas.

The separate estimation of the sampling variances follows a two steps procedure as in [2]. Initially, it is estimated by a proper bootstrap procedure developed taking into account the complex sampling design, using $B = 1000$

repeated samples. Secondly, those estimates are smoothed via a Generalized Variance Function approach to reduce bootstrap sampling error. We estimated it via Hamiltonian MCMC (`Stan`) [1].

3 Conclusions

We proposed a Beta mixture approach for small area estimation of the Relative Theil index, which provides a more flexible framework with respect to Beta regression. We test its performance through a preliminary design-based simulation whose results are encouraging as the estimates we obtain outperform the most common Beta small area model, generally used for parameters defined on the unit interval. The design-based simulations has been carried out by considering NUTS-2 regions as synthetic domains and related demographic and fiscal data as auxiliary variables. Further directions of research involve expanding it to other measures and developing a multivariate context.

References

[1] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. Grantee Submission **76**(1), 1–32 (2017)
[2] Fabrizi, E., Ferrante, M.R., Pacei, S., Trivisano, C.: Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. Computational Statistics and Data Analysis **55**(4), 1736–1747 (2011)
[3] Fabrizi, E., Trivisano, C.: Small area estimation of the Gini concentration coefficient. Computational Statistics and Data Analysis **99**, 223–234 (2016)
[4] Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. Journal of applied statistics **31**(7), 799–815 (2004)
[5] Janicki, R.: Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. Communications in Statistics - Theory and Methods **49**(9), 2264–2284 (2020)
[6] Marchetti, S., Tzavidis, N.: Robust estimation of the theil index and the gini coefficient for small areas. Journal of Official Statistics (2021)
[7] Migliorati, S., Di Brisco, A.M., Ongaro, A., et al.: A new regression model for bounded responses. Bayesian Analysis **13**(3), 845–872 (2018)
[8] Rao, J.N., Molina, I.: Small-area estimation. Wiley Series in Survey Methodology (2015)
[9] Tzavidis, N., Marchetti, S.: Robust domain estimation of income-based inequality indicators. Analysis of Poverty Data by Small Area Estimation pp. 171–186 (2016)

Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics)

Inferenza per big data assistita da metodi di stima per piccole aree: un applicazione sulle OBEC

Monica Pratesi, Francesco Schirripa Spagnolo, Gaia Bertarelli, Stefano Marchetti, Monica Scannapieco, Nicola Salvati, Donato Summa

Abstract Nowadays, the availability of a huge amount of data produced by a wide range of new technologies, so-called big data, is increasing. However, data obtainable from big data sources are often the result of a non-probability sampling process and adjusting for the selection bias is an important problem. In this paper, we propose a novel method of reducing the selection bias associated with the big data source in the context of Small Area Estimation (SAE). Our approach is based on data integration and combination of a big data sample and a probability sample. An application on OBEC (on-line based enterprise characteristics) combining Istat sampling survey and web scraping data has been proposed.

Abstract *Attualmente, la disponibilità di grandi quantità di dati che vengono prodotti da nuove tecnologie, c.d. big data, è sempre più in crescita. Tuttavia, tali big data sono spesso il risultato di un processo di campionamento non probabilistico ed è necessario considerare il problema del bias di selezione. In questo lavoro, proponiamo un nuovo metodo per ridurre il bias di selezione associato ai big data nel contesto della stima per piccole aree. Il nostro approccio si basa sulla metodologia integrazione di dati ed, in particolare, sulla integrazione di un campione di big data e un campione probabilistico. Viene proposta un'applicazione sulle OBEC (caratteristiche dell'impresa on-line based) che combina i dati di indagine campionaria Istat e web scraping.*

Monica Pratesi
Istat and Dipartimento di Economia e Management Università di Pisa e-mail: monica.pratesi@istat.it

Francesco Schirripa Spagnolo; Stefano Marchetti; Nicola Salvati
Dipartimento di Economia e Management Università di Pisa e-mail: francesco.schirripa@unipi.it; stefano.marchetti@unipi.it; nicola.salvati@unipi.it

Gaia Bertarelli
Istituto di Management Scuola Superiore Sant'Anna e-mail: gaia.bertarelli@santanna.it

Monica Scannapieco; Donato Summa
Istat, e-mail: scannapi@istat.it; donato.summa@istat.it

Key words: Data integration; Small Area Estimation; Big data; Official Statistics

1 Introduction

In recent years, there has been a growing demand for more and more detailed official data in order to implement more targeted policies. This has increased the need for appropriate statistical methods to produce reliable statistics for subdomains of a population (such as geographical areas or socio-economic groups). For many decades, probability surveys have been the standard for producing Official Statistics. Due to technological innovations, over the past decade, there has been an unprecedented increase in the volume of "new" data, such as transaction data, social media data, internet of things and scrape data from websites, sensor data and satellite images and so on. Generally, they are called *big data*. Furthermore, the decline in response rates in probability surveys associated with the the increasing cost of data collection have become senior issues for producing official statistics in developed countries.

Big data sources are often the results of non probability sampling processes but, at the same time, they offer very rich data sets: the data can be classified by geographical domains and/or also cross-classified by social and demographic domains (such as gender, educational level for individuals or economic activities for enterprises). Anyway the "nature" itself of the data, as collected without a probability scheme, opens the door to possible selection bias, even at domain level.

Although, there is a trend to modernize official statistics through a more extensive use of big data, and non-probability samples in general, making reliable inferences from a non-probability sample alone is very challenging and a naive use of these data can lead to biased estimates as affected by selection bias and measurement error [5].

So inference from big data sources/domain level data needs to be rethought and selection bias adjustments introduced.

In this context Small area estimation (SAE) methods can contribute as a useful tool to integrate data from probability and non-probability sources. Usually, small area techniques provide official statistics at domain of study level using probability surveys and other sources of available information from which the estimators can borrow strength.

In this work, we assume that we have access to a non-probability sample and a probability survey sample from the same finite population and that the target variable is observed only in the big data source. This situation, tend to be very common in practice and very interesting for future use of big data sources.

2 Effect of the selecion bias when the study variable is not observed in the probability sample

We consider a population U of size N divided into m non-overlapping subsets (domains of study or areas) U_i of size N_i, $i = 1,\ldots,m$. Let y_{ij} denote the value of the target variable for the unit j belonging to the area i. We assume to have two samples referred to this population of interest: a non-probability sample and a probability sample. Moreover, we assume that the study variable is observed only in the non-probability sample.

A non-probability sample, denoted by B, is available for the target population, with $B \subset U$. We assume that the non-probability sample is available in each area of interest: B_i is the non-probability sample in the area i, $B_i \subset U_i$. We denote the inclusion indicator in B_i as δ_{ij}; in other words, $\delta_{ij} = 1$ if $j \in B_i$, $\delta_{ij} = 0$ otherwise; therefore $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$. The study variable y_{ij} is observed only when $\delta_{ij} = 1$. The non-probability contains other auxiliary variables, denoted by $\mathbf{x}$.

A survey data of size n, denoted by A, is available; A_i is a subset of U_i drawn randomly. The survey data do not contain the variable of interest but contain only auxiliary variables $\mathbf{x}$. The area-specific samples A_i are available in each area, but the number of sample units in each area, $n_i > 0$, is limited. Therefore, the areas of interest can be denoted as "small areas". In general, a domain (or area) is regarded as "small" if the domain-specific sample size is not large enough to obtain direct estimates with acceptable statistical significance [7]. These areas can be geographic areas, such as provinces or municipalities and other sub-populations, such as the firms belonging to a industry subdivision. In these cases, SAE techniques need to be employed.

In summary, the available data can be denoted by $\{(y_{ij}, x_{ij}), i \in B\}$ and $\{(x_{ij}), i \in A\}$.

The quantities of interest are the area means $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\ldots,m$.

By using the non-probability sample we can estimate $\bar{Y}_i$ by:

$$\bar{Y}_{B_i} = N_{B_i}^{-1} \sum_{j \in B_i} y_{ij},$$

where $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$ and y_{ij} is the jth observation in the area i. Because of the selection bias and the measurement error, the sample mean $\bar{Y}_{B_i}$ from the non-probability sample is biased. Indeed, non-probability samples have unknown selection/inclusion mechanisms and are typically biased, and they do not represent the target population [3, 8]. Thus, a non-probability sampling design makes the analysis results subject to selection bias.

Therefore, we propose a techniques in order to make valid inference from big data sources when the aim is to provide reliable estimates at small area level.

3 Reducing selection bias in big data sources: a data integration approach using Small Area Estimation methods

Data integration represents a quite new research area aimed at combining information from two independent surveys on the same target population [4].

Using multiple data sources is common in SAE; indeed, small area methods combine the data from a survey with predictions from a regression model using covariates from the administrative or census data. The SAE models are classified into two categories according to the available data on the target variable: (i) area level models and (ii) unit levels model. The *standard* SAE models for the mean use hierarchical model in which the deviation of an area mean from the overall mean is represented by a random effect.

If information at unit level is available, the standard unit-level small area model proposed by [1] may be used. In this case, the hierarchical model used for the individual response of the survey individual j in area i is:

$$y_{ij} = \mathbf{x}_{ij}^T \beta + u_i + e_{ij}, \tag{1}$$

where the area-specific random effects u_i and individual level errors e_{ij} are assumed to be normally distributed with mean 0 and variance σ_u^2 and σ_e^2, respectively.

We suppose that the quantities of interest are the area means, so it possible to express the mean in terms of linear combination between observed and unobserved units as follows

$$\theta_i = N_i^{-1} \left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} (\mathbf{x}_{ij}^T \hat{\beta} + \hat{u}_i) \right], \tag{2}$$

where $\hat{\beta}$ and $\hat{u}_i$ are the BLUE for β and u_d and s_i is the set of the sampled units in area i and r_i is the set of the non-sampled units in area i. Replacing the variance components by their estimators we obtain the Empirical Best Linear Unbiased Predictor (EBLUP).

Being assisted by unit level approach, we propose a new method to producing statistics at local level when the variable of interest has been recorded only in the non-probability sample. In particular, we consider a data integration method for combining probability and non-probability samples (i.e. big data sample) assisted by unit level small area model, following the approach of [3], in order to reduce the bias.

We consider the case in which the survey data and the big data are available in each small area of interest. We also assume that the selection mechanism for the big data is non-informative :

$$P(\delta_{ij} = 1 | \mathbf{x}_{ij}, y_{ij}; u_i) = P(\delta_{ij} = 1 | \mathbf{x}_{ij}; u_i)$$

where u_i is an area-specific random effect characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$.

Moreover, we can observe δ_{ij}, the big data sample inclusion indicator, from the sample A. In other words, among the elements in sample A, it is possible to obtain the membership information from the big data sample B.

We can use the data $\{(\delta_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the for the participation probabilities or propensity scores ($P(\delta_{ij} = 1|\mathbf{x}_{ij}) = p(\mathbf{x}, \lambda)$) in sample B based on the missing at random (MAR). Usually, a logistic regression model for the binary variable δ_{ij} can be used in order to obtain estimators $\hat{p}_{ij}$ in sample B.

In order to take in to account the hierarchical structure of the data, we consider the following generalized linear random intercept model for the propensity scores:

$$\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i) = g^{-1}(\mathbf{x}_{ij}^T \hat{\lambda} + \hat{u}_i),$$

where $g(\cdot)$ is a logit link function; $\hat{\lambda}$ and $\hat{u}_i$ are the ML estimates of λ and u_i.

In order to develop our estimator we suppose that the following working population model holds for sample B:

$$E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T \beta + \gamma_i\right), \tag{3}$$

where $h(\cdot)$ is the link function, assumed to be known and invertible, γ_i is the area-specific random effect for area i characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$. It should be noted that the covariates used here could be different from those used to fit the propensity model. Model in equation (3) includes three important special cases: the linear model obtained with $h(\cdot)$ equal to the identity function and y_{ij} is a continuous variable; logistic generalized linear random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized linear random intercept model where $h(\cdot)$ is the log link function and the individual y_{ij} values are taken to be independent Poisson random variable.

Using data from the big data sample B, assuming the model is correctly specified, we obtain an estimator of $\hat{\beta}$ which is consistent for β [6].

Then a doubly robust (DR) estimator of the mean is given by:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i)}(y_{ij} - \hat{\mu}_{ij}) + \frac{N_i}{n_i}\sum_{j \in A_i} \hat{\mu}_{ij}\right\}, \tag{4}$$

where $\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}\hat{\beta} + \hat{\gamma}_i\right)$ and $\hat{\beta}$ and $\hat{\gamma}_i$ are respectively the estimated regression coefficients and the random effects based on the big data sample.

The estimator in Eq. (4) is DR in the sense that it is consistent if both the model for propensity scores and the model for the study variable are correctly specified [3, 6].

4 Application Setting: Estimating Online-Based Enterprise Characteristics

Let us consider a setting in which the Big Data source is represented by the websites of enterprises that are accessed as a result of a web scraping procedure. Starting from a set of URLs (i.e. addresses identifying the enterprise websites), the procedure accesses URLs, extracts texts from the sites and stores such texts for subsequent analyses. In particular, text analyses can be performed to estimate the so-called Online-based Enterprise Characteristics (OBEC), i.e. some characteristics of businesses that are available on their own websites. In this specific setting, we assume to start from the Italian Statistical Business Register and being U the universe of enterprises with equal or more than 10 employees, we select the subset S having a (valid) URL available. The Big Data sample B is accessed starting from S and will consist of all the texts of scraped websites. Notice that, assuming that URLs are all valid, the cardinality of S is equal to cardinality of B, i.e. B is the online representation of enterprises in S. By using B, we would like to compute a Yes/No indicator Y, considering if the enterprise is sensitive or not to Sustainable Development Goals of the 2030 Agenda. The indicator, named SDG enterprise sensitiveness, can be computed by analyzing B and looking for the presence of a set of pre-defined SDG related words on each website. B and S share a set of X variables that include Vat Code, Name of the Enterprise, Address, Municipality, Province, Zip Code, NACE code and Number of employees. In addition, X variables are also common to specific survey data A; in this application, we will use data of the " ICT usage in enterprises" survey. Considering A and B, let us observe that we can consider a specific variable that denote enterprises present in A but not in B; the variable reports if an enterprise has a known website, i.e. a URL is available, or not. Figure 1 reports a visual representation of the application setting.

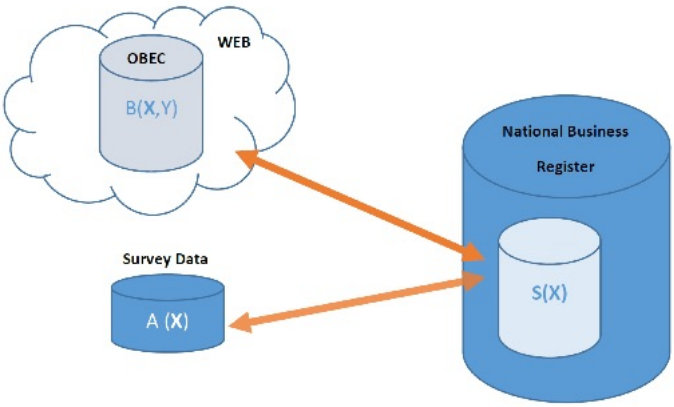


Fig. 1 Application Setting

In summary, as illustrated in this example, big data sources are a treasure of information that runs the risk of being underestimated as not connected with existing official data. They offer data affected by selection bias, as already stated in many scientific papers (see, among others, 2, 6) and adjusting for this selection bias in big data is an important and urgent problem. The effect of selection bias is likely to be even more serious at domain level when the domains are defined by socio-demographic groups. Age groups, gender, educational level, zone of residence, geography in general are often highly correlated with digital divide. This last is often the factor explaining self-selection bias and the presence/absence in big data sources of individuals, households and firms.

In this work we dealt with the problem of making reliable inference for small domains when the target variable is stored in a non-probability sample (big data sample) which is assumed to be available in each area and the number of units in each area is quite large. In particular, we propose a method based on the integration of a probability and a non-probability sample in order to reduce the selection bias associated with big data when the aim is to predict statistics at the local level.

References

[1] Battese, G.E., Harter, R.M., Fuller, W.A.: An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association **83**, 28–36 (1988)

[2] Beaumont, J.-F.: Are probability surveys bound to disappear for the production of official statistics?. Survey Methodology **46**, 1–28 (2020)

[3] Kim, J.K., Wang, Z.: Sampling techniques for big data analysis. International Statistical Review **87**, S177–S191 (2019)

[4] Lohr, S.L., Raghunathan, T.E: Combining survey data with other data sources. Statistical Science **32**, 293–312 (2017)

[5] Meng, X.-L.: Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics **12**, 685–726 (2018)

[6] Rao, J.N.K.: On making valid inferences by integrating data from surveys and other sources. Sankhya B **83**, 242–272 (2021)

[7] Rao, J.N.K., Molina, I.: Small area estimation. John Wiley & Sons, New York (2015)

[8] Yang, S., Kim, J.K.: Statistical data integration in survey sampling: A review. Japanese Journal of Statistics and Data Science (2020) doi: 10.1007/s42081-020-00093-w

Statistical methods and models for Sports Analytics

The 'hot shoe' in soccer penalty shootouts

La 'scarpa calda' nei calci di rigore

Andreas Groll and Marius Ötting

Abstract We propose a modeling framework for dealing with a large amount of covariates in hidden Markov models (HMMs) by considering a LASSO penalty. This modeling framework is, for example, useful in sports for analyzing a potential hot hand effect, as several existing studies on the hot hand consider HMMs. However, with most studies analyzing data from basketball or baseball, there are several confounding factors which have to be taken into account, leading to a potential large number of covariates. Hence, in those settings regularization methods are suitable to allow for implicit variable selection. As a case study we investigate a potential "hot shoe" effect among penalty-takers.

Abstract *Nel presente contributo si propone una modellistica per gestire una grande quantità di covariate nei c.d. hidden Markov models (HMM) considerando una penalità LASSO. Questa modellistica è, per esempio, utile nello sport per analizzare un potenziale effetto mano calda, come diversi studi esistenti sulla mano calda che considerano gli HMM. Tuttavia, come per la maggior parte degli studi che analizzano dati di basketball o baseball, ci sono diversi fattori di confusione che devono essere presi in considerazione e che portano a un numero potenzialmente elevato di covariate. In tali situazioni, i metodi di regolarizzazione sono adatti per consentire una selezione implicita delle variabili. Come caso di studio, si indaga un potenziale effetto "hot shoe" nel tiro dei calci di rigore.*

Key words: hidden Markov model; LASSO; hot hand; sports analytics; soccer.

Andreas Groll
TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany, e-mail: groll@statistik.tu-dortmund.de

Marius Ötting
Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany e-mail: marius.oetting@uni-bielefeld.de

1 Introduction

An often discussed phenomenon in different sports is the "hot hand", meaning that players may enter a state where they experience extraordinary success. This phenomenon is also discussed in the media, where commentators and journalists — e.g. in soccer — commonly refer to players as being "on fire" when they score in consecutive matches. Academic research on the hot hand started by Gilovich et al. (1985). In their seminal paper, they analyzed basketball free-throw data and found no evidence for the hot hand, arguing that people tend to belief in the hot hand due to memory bias.

More recent studies challenge the findings of Gilovich et al. (1985), often by analyzing data from basketball or baseball with regard to a hot hand effect, e.g, Miller and Sanjurjo (2018). In addition, these studies often consider hidden Markov models (HMMs), which constitute a natural modelling approach for the hot hand as they accommodate the idea that players potentially may enter a state where they experience extraordinary success. However, when modelling a potential hot hand effect, there is hardly any sport where no potential confounding factors exist, such as weather conditions in baseball or the performance of opponents in basketball. Accounting for those factors leads to a large number of covariates, and often multicollinearity issues occur, making model fitting and interpretation of parameters difficult. To tackle these problems and to obtain sparse and interpretable models, we propose to conduct variable selection in HMMs by considering a LASSO penalization approach (see Tibshirani, 1996).

First, the performance of LASSO-HMMs is investigated in a short simulation study. Next, as a case study, we investigate a potential "hot shoe" effect of penalty takers in the German Bundesliga ($n = 3,482$ penalties). Figure 1 shows all penalties taken by Bayern Munich's attacker Gerd Müller, indicating that there are periods (e.g. between 1975 and 1976) where he scored several penalties in a row, but also periods (e.g. around 1971) where he missed a few consecutive penalties.

2 Methods

In HMMs, the observations y_t are assumed to be driven by an underlying state process s_t, in a sense that the y_t are generated by one of N distributions according to

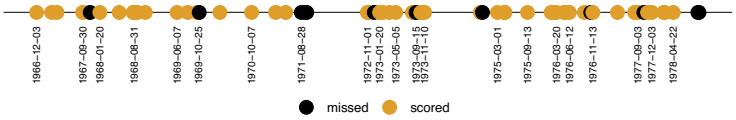


Fig. 1 Penalty history over time of the player Gerd Müller for the time period from 1964 until 1979 (successful penalties in yellow, failures in black).

the Markov chain. In our application, the state process s_t serves for the underlying varying form of a player. State switching is modelled by the transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$, with $\gamma_{ij} = \Pr(s_t = j | s_{t-1} = i)$, $i, j = 1, \ldots, N$. We further allow for additional covariates at time t, $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{Kt})^\top$, each of which assumed to have the same effect in each state, whereas the intercept is assumed to vary across the states, leading to the following linear state-dependent predictor:

$$\eta_t^{(s_t)} = \beta_0^{(s_t)} + \beta_1 x_{1t} + \ldots + \beta_k x_{Kt}\,.$$

For our response variable y_t, indicating whether the penalty attempt t was successful or not, we assume $y_t \sim \text{Bern}(\pi_t^{(s_t)})$ and link $\pi_t^{(s_t)}$ to our state-dependent linear predictor $\eta_t^{(s_t)}$ using the logit link function, i.e. $\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)}$. Defining an $N \times N$ diagonal matrix $\mathbf{P}(y_t)$ with i–th diagonal element being equal to $\Pr(y_t | s_t = i)$, and assuming that the initial distribution $\boldsymbol{\delta}$ of a player is equal to the stationary distribution, i.e. the solution to $\boldsymbol{\Gamma}\boldsymbol{\delta} = \boldsymbol{\delta}$ subject to $\sum_{i=1}^{N} \delta_i = 1$, the likelihood for a single player p is given by

$$L_p(\boldsymbol{\alpha}) = \boldsymbol{\delta}\mathbf{P}(y_{p1})\boldsymbol{\Gamma}\mathbf{P}(y_{p2})\ldots\boldsymbol{\Gamma}\mathbf{P}(y_{pT_p})\mathbf{1}\,,$$

with vector $\boldsymbol{\alpha} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1N}, \ldots, \gamma_{NN}, \beta_0^{(1)}, \ldots, \beta_0^{(N)}, \beta_1, \ldots, \beta_k)^\top$ collecting all unknown parameters, and column vector $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^N$ (see Zucchini et al., 2016). To obtain the likelihood for the complete data set, i.e. for multiple players, we assume independence between the observations of different players (here: $p = 310$), so that the likelihood is given by the product of the individual likelihoods:

$$L(\boldsymbol{\alpha}) = \prod_{p=1}^{310} L_p(\boldsymbol{\alpha}) = \prod_{p=1}^{310} \boldsymbol{\delta}\mathbf{P}(y_{p1})\boldsymbol{\Gamma}\mathbf{P}(y_{p2})\ldots\boldsymbol{\Gamma}\mathbf{P}(y_{pT_p})\mathbf{1}\,.$$

Parameter estimation is done by maximizing the likelihood numerically using `nlm()` in R. However, considering a large amount of covariates leads to a rather complex model, which is hard to interpret and, in addition, multicollinearity issues might occur. Hence, we propose to employ a penalized likelihood approach based on a LASSO penalty.

The basic idea is to maximize a penalized version of the log-likelihood $\ell(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha}))$. More precisely, one maximizes the penalized log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha})) - \lambda \sum_{k=1}^{K} |\beta_k|\,, \tag{1}$$

where λ represents a tuning parameter, which controls the strength of the penalization. To fully incorporate the LASSO penalty in our setting, the non-differentiable L_1 norm $|\beta_k|$ in (1) is approximated as suggested by Oelker and Tutz (2017). Specifically, $|\beta_k|$ is approximated by $\sqrt{(\beta_k + c)^2}$, where c is a small positive number (say c = 10^{-5}). Practically, a coefficient is then selected if $|\hat{\beta}_k| \geq 0.001$. The optimal value for the tuning parameter λ is chosen by model selection criteria such as AIC

and BIC. To estimate the required effective degrees of freedom, we consider all parameters in the model which are unequal to zero, i.e. all entries of the t.p.m., all state-dependent intercepts, and all selected β_j's.

3 Simulation study

We consider a simulation scenario similar to our real-data application, with a Bernoulli-distributed response variable, an underlying 2-state Markov chain and 50 covariates, 47 of which being noise covariates:

$$y_t \sim \text{Bern}(\pi_t^{(s_t)}), \quad \text{with}$$

$$\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)} = \beta_0^{(s_t)} + 0.5 \cdot x_{1t} + 0.7 \cdot x_{2t} - 0.8 \cdot x_{3t} + \sum_{j=4}^{47} 0 \cdot x_{jt} \,.$$

We further set $\beta_0^{(1)} = \text{logit}(0.75)$ and $\beta_0^{(2)} = \text{logit}(0.35)$. The performance of three different fitting schemes is investigated, namely HMMs without penalisation (i.e. $\lambda = 0$) and the LASSO-HMM with λ selected by AIC and BIC, respectively. The fitting schemes are compared by the mean squared error (MSE) of the β_j (see Figure 2). The results of the simulation study suggest that, in terms of MSE, the LASSO-HMM with λ selected by BIC performs worst, with the MSE being higher than for the HMM without penalisation. The LASSO-HMM with λ selected by AIC outperforms the other fitting schemes considered in terms of MSE.

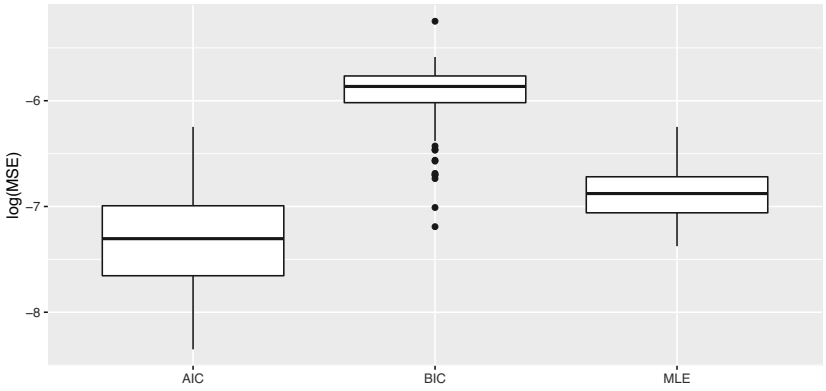


Fig. 2 Boxplots of the MSE obtained in 100 simulation runs. "AIC" and "BIC" denote the LASSO-HMM fitting schemes with λ chosen by AIC/BIC. "MLE" denotes unpenalised HMM.

4 Application

As the LASSO-HMM with λ selected by the AIC showed the most promising results in the simulations, we use this fitting scheme in the following. For modelling the hot shoe, we account for several factors potentially affecting the outcome of a penalty kick, namely a dummy indicating whether the match was played at home, the matchday, the minute of play the penalty was taken, the experience of both the penalty taker and the goalkeeper (quantified by the number of years the player played for a professional team), and the current match score difference. In addition, to account for player-specific abilities, we include dummy variables for all penalty takers and goalkeepers. This results in 656 covariates in total.

The parameter estimates obtained (on the logit scale) indicate that the baseline level for scoring a penalty is higher in the model's state 1 than in state 2 ($\hat{\beta}_0^{(1)} = 1.422 > -14.50 = \hat{\beta}_0^{(2)}$), thus indicating evidence for a hot shoe effect. State 1, hence, can be interpreted as a hot state, whereas state 2 refers to a cold state. In addition, with the t.p.m. estimated as

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.978 & 0.022 \\ 0.680 & 0.320 \end{pmatrix},$$

there is high persistence in state 1, i.e. in the hot state. However, when being in state 2 (cold state) switching to state 1 is most likely. Additionally, the model is slightly favoured by the AIC over a 1-state model, i.e. a standard logit model without a potential hot shoe effect ($\text{AIC}_{\text{hotshoe}} = 3664$, $\text{AIC}_{\text{1-state-model}} = 3670$). The coefficient paths of our model are shown in Figure 3. Out of the 656 covariates included in our model, only a single covariate is selected according to the AIC, namely the ability of the goal kepper Jean-Marie Pfaff with $\hat{\beta}_{\text{Pfaff}} = -0.0015$. The negative effect indicates that the odds for scoring a penalty decrease if Jean-Marie Pfaff is the goalkeeper of the opposing team — in fact he saved remarkable 9 out of 14 penalty kicks during his career in the Bundesliga. To further illustrate our variable selection approach, Figure 3 additionally highlights the covariates which would be selected next, namely the abilities of Günther Herrmann (outfield player) and Rudolf Kargus (goalkeeper). As several existing studies provide evidence for a home advantage in soccer, we also highlight in Figure 3 the corresponding coefficient path of the dummy variable indicating whether a match was played at home (but note that it is also not selected here). For more detailed results of the application and for results of the simulation study, see Ötting and Groll (2021).

5 Outlook

Further research could focus on additional penalties to conduct variable selection within HMMs, such as the ridge penalty or the elastic net. In the case of multicollinearity, especially the elastic net may show a superior performance compared

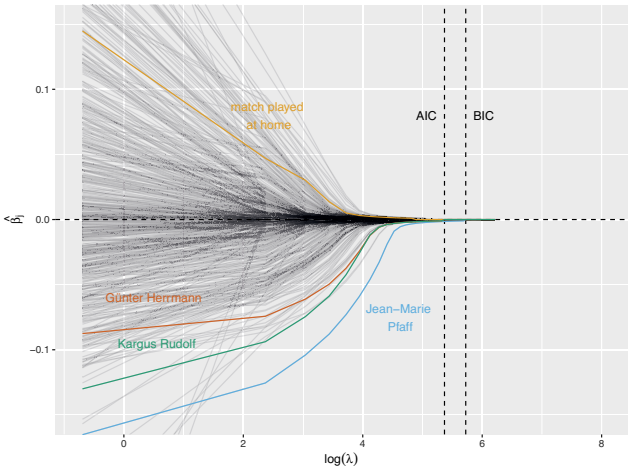


Fig. 3 Coefficient paths of all covariates considered in the LASSO-HMM models. Dashed vertical lines indicate the penalty parameters λ as selected by AIC and BIC, respectively. For BIC, no covariates are selected, whereas for the AIC the player-specific effect of Jean-Marie Pfaff is selected. The player-specific abilities of Günter Herrmann and Rudolf Kargus would be selected next.

to the LASSO. Moreover, modifications of the standard LASSO such as the relaxed-LASSO could be considered.

Acknowledgements We want to thank the group of researchers B. Bornkamp, A. Fritsch, L. Geppert, P. Gnändinger, K. Ickstadt, and O. Kuss for providing the German Bundesliga penalty data set. The hot shoe dataset is available via the R package `footballpenaltiesBL` from CRAN (Geppert et al., 2021).

References

1. Geppert, L.N., Gnändinger, P., Ickstadt, K., Bornkamp, B., Fritsch, A., Kuß, O.: `footballpenaltiesBL`: Penalties in the German Men's Football Bundesliga, R package version 1.0.0, 2021
2. Gilovich, T., Vallone, R., and Tversky, A.: The hot hand in basketball: On the misperception of random sequences. Cognitive psychology. **17**, 295–314 (1985)
3. Miller, J.B., Sanjurjo, A.: Surprised by the hot hand fallacy? A truth in the law of small numbers. Econometrica. **86**(6), 2019–2047 (2018)
4. Oelker, M.R., Tutz, G.: A uniform framework for the combination of penalties in generalized structured models. Advances in Data Analysis and Classification. **11**, 97–120 (2017)
5. Ötting, M., Groll, A.: A regularized hidden Markov model for analyzing the 'hot shoe' in football. Statistical Modelling, online first. (2021) doi: https://doi.org/10.1177/1471082X211008014
6. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. **58**, 267–288 (1996)
7. Zucchini, W., MacDonald, I.L., Langrock, R.: Hidden Markov Models for Time Series: An Introduction Using R. Chapman and Hall-CRC, Boca Raton (2016)

G-RAPM: revisiting player contributions in regularized adjusted plus-minus models for basketball analytics

G-RAPM: una rivisitazione del modello RAPM per i dati della pallacanestro

Luca Grassetti

Abstract Identification and analysis of players ranking have a central role in the sports analytics. An essential tool in this framework is the Regularized Adjusted Plus-Minus (RAPM) model. When player and lineup effects are included simultaneously, the interpretation of the RAPM model results can be cumbersome. The present work aims at estimating a modified version of the RAPM model, adopting a one-sided assumption for player effects. The proposed specification allows for a direct performance interpretation. The model can be estimated feasibly within the Bayesian framework, allowing for straightforward generalisations.

Abstract *L'identificazione e l'analisi delle performance dei giocatori hanno un ruolo centrale nello Sports analytics. Il modello Regularized Adjusted Plus Minus (RAPM) rappresenta una soluzione modellistica a questo problema. Quando gli effetti dei giocatori e delle formazioni sono introdotti contemporaneamente nel modello, l'interpretazione dei risultati di stima risulta complicata. Con l'obiettivo di rendere diretta l'interpretabilità dei risultati, Il presente lavoro propone un modello RAPM modificato che consideri gli effetti dei giocatori come parametri positivi. Tale modello può essere stimato con un approccio Bayesiano rendendo semplice l'eventuale generalizzazione dello stesso.*

Key words: RAPM, Basketball, Sports Management, Player Performances and Lineup Synergies

1 Introduction

The calculation of the performance of players considering their on-field statistics is a very relevant topic in the sports management framework [11]. As reported in [4],

Luca Grassetti
Department of Economics and Statistics – University of Udine, Via Tomadini, 30/a, Udine, Italy, e-mail: luca.grassetti@uniud.it

coaches and team managers can make decisions based on players ratings, which can be estimated in different ways. Among others, Regularized Adjusted Plus-Minus (RAPM) models [1, 8] can be considered as one of the best solutions because the model-based approach (first introduced by [7]) allows computing the efficiency of players while accounting for the value of the opposing lineup and including also some exogenous variables to mitigate the presence of confounding. When considering a model-based approach, most existing proposals suggest treating the player effects as real-valued parameters defining the contribution of each individual to the team score. This solution is really efficient and allows to compare the players, but from the sports management point of view, the interpretation of these results is not always straightforward. This shortcoming is even more relevant when the RAPM model is generalised to the simultaneous presence of lineup and player effects. A modification to the model adopted in [3] (called G-RAPM) is proposed aiming at the straightforward interpretation of the results. In particular, a production frontier-like approach (see [5] for further details) is considered in the specification of a RAMP model where the players are treated as cumulative inputs determining a naïve performance of the lineup. An additional real-valued lineup effect is finally introduced to adjust the sum of player effects, which accounts for their positive or negative synergies (interactions, in statistical terms). In this way, the player effects can be directly interpreted as contribution to the team performance, net of their interactions.

The present paper is developed in the basketball data analysis framework, but its generalisation to other team sports (such as ice hockey, soccer, and volleyball) is straightforward. The following analyses are based on the play by play data regarding Euroleague 2018/19 season. All the results are obtained using R ([6]), adopting the Full Bayes approach to RAPM model estimation. `rstan` library ([10]) and `cmdstanr` package ([2]) are used to estimate the model and visualize the diagnostic output, respectively.

2 A frontier like model for the scores

The idea motivating the present work is that by specifying models for the home and away scores separately, the role of players and lineups on the performance of the team can be assessed. The home and away scores (computed following the idea introduced in [9]) show a peculiar empirical distribution and, in particular, they present a positive skewness. In order to account for this asymmetry, one can define a model where the effects of interest (players and/or lineups) present a skewed distribution. For instance, the model for the home scores can be

$$\mathbf{y}^H = \mathbf{X}\beta^H + \mathbf{Z}^{(Hl)}\mu + \mathbf{Z}^{(Hp)}\gamma + \varepsilon^H, \tag{1}$$

where $\mathbf{X}$ is the design matrix for the model covariates (such as period of the game, that are generic for home and away data), $\mathbf{y}^H$, $\mathbf{Z}^{(Hl)}$ and $\mathbf{Z}^{(Hp)}$ are the home-specific response vector and the design matrices for lineup and player effects, respectively.

The model specification can be finalised considering $\mu \sim N(0, \sigma_\mu^2)$ and $\gamma \sim Exp(\lambda_\gamma)$ in the likelihood specification for the effects of lineups and players, respectively. The idiosyncratic error term is $\varepsilon^H \sim N(0, \sigma_{\varepsilon^H}^2)$. An analogous model can be defined for the away team by replacing H with A in superscripts. Lastly, the joint model for the difference between home and away scores results in

$$\mathbf{y} = \mathbf{y^H} - \mathbf{y^A} = \mathbf{X}\beta + \mathbf{Z}^{(l)}\mu + \mathbf{Z}^{(p)}\gamma + \varepsilon,$$

where $\mathbf{Z}^{(l)} = \mathbf{Z}^{(Hl)} - \mathbf{Z}^{(Al)}$, $\mathbf{Z}^{(p)} = \mathbf{Z}^{(Hp)} - \mathbf{Z}^{(Ap)}$, and $\beta = \beta^H - \beta^A$.

3 The empirical analysis

While the proposed model estimation results are fully comparable with those of a standard RAPM model (comparison is omitted here for space reason), their improved interpretability can be crucial for the data-driven management of a team. The information associated with the estimated effects is practically equivalent in the two frameworks. The lineup effects estimated under the Gaussian assumption are slightly larger than those involved in the G-RAPM model. The estimated player effects are strongly correlated between the models (the ranks correlation is > 0.95), but their size is quite different, due to distributional assumptions.

The most relevant G-RAPM model improvement regards the interpretability of estimated effects. In fact, the player performances, estimated assuming the one-sided distribution, are fully interpretable as individual contributions to the total performance of the lineups. Moreover, the lineup effect can be studied in-depth to evaluate the interaction among the individuals. Plots in Figure 1 show the results for the worst and best five lineups for the Milan team. The worst five lineups (on the left panel) show a counter-productive interaction among players. The plotted total lineup effects (identified with the white bars) are always lower than the simple sum of players estimated performances. On the contrary, for the best five lineups (on the right panel), the total lineup performance is always higher than the sum of players effects. The individuals involved in the latter lineups exhibit a positive synergy.

4 Conclusions

The novelty in the proposed specification is directly connected with a more natural concept of single-player contribution to the performance of the lineup. Each player in the proposed model specification corresponds to a positive effect whose size is directly associated with the performance evaluation. These effects can be used for a ranking, as usual, and summed up to define the potential performance of a lineup, which is then corrected considering the actual lineup effect, measuring the interaction among individuals. Consequently, the model interpretation is immediate, and

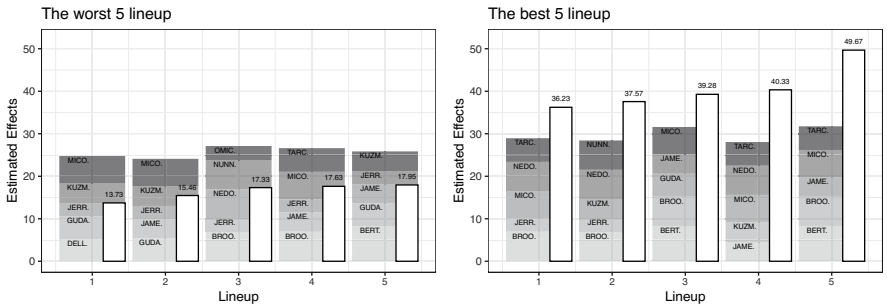


Fig. 1 The composition of total lineup effects (identified with the white bar in the figures) for the worst (left-panel) and the best (right-panel) five lineups in Milan team, including player effects and negative (or positive) synergies due to lineup composition.

the recognition of positive or negative synergies among players can be directly used to choose the best-five men units. Further discussion on this proposal can focus on alternative assumptions on player effects and the model generalisation considering home and away-specific effects for individuals and lineups.

References

1. Engelmann, J.: Possession-based player performance analysis in basketball (adjusted +/– and related concepts). In: Handbook of statistical methods and analyses in sports, pp. 231-244. Chapman and Hall/CRC (2017)
2. Gabry, J. and Češnovar, R.: cmdstanr: R Interface to 'CmdStan'. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org (2021)
3. Grassetti, L., Bellio, R., Gaspero, L., Fonseca, G., and Vidoni, P.: An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. IMA Journal of Management Mathematics 32, no. 4, pp. 385-409 (2021)
4. Hvattum, L.: A comprehensive review of plus-minus ratings for evaluating individual players in team sports. International Journal of Computer Science in Sport (2019) doi: 10.2478/ijcss-2019-0001
5. Kumbhakar, S.C., Parmeter, C.F., and Zelenyuk, V.: Stochastic Frontier Analysis: Foundations and Advances I. In: Ray S.C., Chambers R., Kumbhakar S.C. (eds) Handbook of Production Economics. Springer, Singapore (2021) doi: 10.1007/978-981-10-3450-3_9-2
6. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2022)
7. Rosenbaum, D.: Measuring how NBA players help their teams win. http://www.82games.com/comm30.htm (2004)
8. Sill, J.: Improved NBA adjusted +/- using regularization and out-of-sample testing. In Proceedings of the 2010 MIT Sloan Sports Analytics Conference. (2010)
9. Sisneros, R., and Van Moer, M.: Expanding plus-minus for visual and statistical analysis of NBA box-score data. In The 1st workshop on sports data visualization. IEEE. (2013)
10. Stan Development Team: RStan: the R interface to Stan. R package version 2.21.2. http://mc-stan.org/ (2020)
11. Tiedemann, T., Francksen, T., and Latacz-Lohmann, U.: Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. Central European Journal of Operations Research **19, no. 4**, 571-587 (2011)

Formative vs Reflective constructs: a CTA-PLS approach on a goalkeepers' performance model

Costrutti Formativi vs Riflessivi: un approccio CTA-PLS su un modello di performance dei portieri

Mattia Cefis and Eugenio Brentari

Abstract Nowadays, PLS-SEM is a trend-topic, whereas football is moving towards a data-driven approach; by combining these two worlds, we aim to show a new way for measuring football goalkeepers' performance, by using data provided from EA Sports experts and available on the Kaggle data science platform. Furthermore, another objective is to refine the model, supporting football experts from a statistical point of view. For this purpose, we adopt a confirmatory tetrad analysis (CTA-PLS) to validate and evaluate the nature (e.g. formative or reflective) of each latent variable. Then, a second-order PLS-SEM model is built. We validate and compare this new indicator with a benchmark (the EA *overall*). The final goal is to prove the CTA approach on a real case study and to refine a composite performance indicator for helping football policy makers taking strategic decisions.

Abstract *Al giorno d'oggi, il PLS-SEM è un argomento di tendenza mentre il calcio si sta muovendo verso un approccio data-driven; combinando questi due mondi, vogliamo mostrare un nuovo modo per misurare le abilità dei portieri, utilizzando i dati definiti dagli esperti EA e disponibili sulla piattaforma Kaggle. Come secondo obiettivo vogliamo supportare gli esperti grazie ad un approccio statistico. Con questo fine, applicheremo un'analisi CTA-PLS per valutare la natura (e.g. formativa o riflessiva) di ogni variabile latente. In seguito abbiamo implementato un modello PLS-SEM di secondo ordine. Abbiamo poi confrontato questo nuovo indicatore con un indice di riferimento (l'EA overall). L'obiettivo ultimo è quello di testare la CTA analisi su un reale caso di studio e offrire un indicatore composito di performance per aiutare gli addetti ai lavori a prendere decisioni strategiche.*

Key words: CTA-PLS, PLS-SEM, Latent variables, Football, Performance.

Mattia Cefis
University of Brescia, Department of Economics and Management, e-mail: mattia.cefis@unibs.it

Eugenio Brentari
University of Brescia, Department of Economics and Management, e-mail: eugenio.brentari@unibs.it

1 Introduction

The latest developments in sports research are moving towards a data-driven approach. In particular, focused on football (i.e. soccer for Americans), players' performance measure is becoming a strategic key for football coaches and policy makers, in order to evaluate players impartially. The majority of papers on performance evaluation are focused just on movement players (i.e. defenders, midfielders and forwards, [5]): by this research we want to focalize attention on a singular role, the goalkeepers. We are inspired by Electronic Arts (EA)[1] experts: in their opinion, goalkeepers' performance can be thought as a multidimensional construct made up of 7 performance composite indicators (i.e. the same 6 used for movement players plus a specific one for goalkeepers), each one made up of several specific skills, which combined form an *overall* index that sums up the performance; then, a statistical support is required [2, 4]. Using data provided by the Kaggle data science platform, our goal is to propose the use of an innovative confirmatory tetrad analysis applied in the PLS context (CTA-PLS) to support experts from a statistical point of view regarding the nature of each construct, as formative or reflective. Following the CTA-PLS output, we will build a second order Partial Least Squares - Structural Equation Model (PLS-SEM) model, in order to build a refined composite indicator dedicated to goalkeepers and comparing it with the well-known EA *overall*.

2 Literature overview and data employed

Existing literature focused on players' performance [2, 4] includes different approaches: for example Carpita et al [3] adopted an unsupervised method to classify different area of performance, Cefis and Carpita [5] already proposed a PLS-SEM model considering only movement roles, but without a CTA approach. The aim of this research is to focalize attention on the evaluation of goalkeepers' performance, exploring key performance indices (KPIs), in order to evaluate some different strategic latent variables (LVs) and their theoretical nature (i.e. formative or reflective).

For this application has been used data from EA experts and available on the Kaggle[2] data science platform; in particular, we will focus on all goalkeepers' stats from the top 5 European Leagues (e.g., Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This dataset contains 31 variables (e.g. KPIs), with periodic players' performance on a 0-100 scale with respect to different abilities, classified by *sofifa* experts into 6 latent traits: *attacking, skill, movement, power, mentality* and *goalkeeper features*; note that, after a preliminary check, we did not take into account the *defending* block for this model, since its skills are strictly related with movement players. Note that a block is a group of MVs forming a LV: for example the *skill* block is composed by dribbling, curve, fk

[1] www.easports.com

[2] www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

accuracy, long passing and ball control. The classification provided by *sofifa* experts is available online[3]. For our purpose we have chosen to take into account data relying the beginning of the season 2019/2020, so the dataset was composed by stats about 331 goalkeepers.

2.1 The PLS-SEM model and the CTA-PLS approach

PLS-SEM [15], also called PLS-PM, is a tool that offers a valid alternative as compared to the well-known covariance-based model [10]. Its goal is to measure causality relation between concepts (e.g. LVs), starting from some manifest variables (MVs), by an exploratory approach: the explained variance of the endogenous latent variables is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression. Additionally, PLS-SEM does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. In our framework, PLS-SEM estimates simultaneously two models: a measurement (outer) and a structural (inner). In particular, for what concern the measurement model, PLS-SEM allows two types of constructs, respectively reflective and formative: the first one implies that the *q-th* LV exists independently from the measures used (1) (i.e. causality from construct to items, where λ_{pq} is the loading connecting LV q with its MV p, by a simple linear regression, estimated by OLS), whereas the second is determined as a combination of its own indicators (2) (i.e. causality from items to construct, each latent variable ξ_q is considered to be formed by its own MVs following a multiple regression, where the weights are estimated by least squares).

$$x_{pq} = \lambda_{pq}\xi_q + \varepsilon_q \tag{1}$$

$$\xi_q = \sum_{p=1}^{p_q} w_{pq}x_{pq} + \delta_q \tag{2}$$

But there is a lack: while for reflective constructs exist several tests to assess their reliability, for what concern formative constructs researchers are just basing on theory and experts opinion, causing possible measurement misspecifications. As consequence, this can lead a bias in the inner model estimation and lead to incorrect assessments of relationships in PLS-SEM [8]. In order to overstep those limits, some researchers applied the confirmatory tetrad analysis (CTA, [1]) for drawing conclusions about the appropriateness of using formative measurement models as compared to reflective ones [8]. In brief, a tetrad τ is the difference between the product of two pairs of covariances; for instance, the six covariances of a block composed by four MVs permit the formation of three tetrads:

[3] https://sofifa.com/player/192985/kevin-de-bruyne/220030/

$$\begin{aligned}\tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} \\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} \\ \tau_{1423} &= \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43}\end{aligned} \tag{3}$$

Note that all tetrads for each block of LV must be tested using a bootstrap procedure (CTA-PLS uses the bias corrected bootstrap by a Bonferroni -nonparametric- approach [8]). If all tetrads confidence intervals (CIs) for that specific LV contain zero (i.e. vanishing tetrads) then the construct can be considered as reflective, otherwise it is formative [8, 1].

Starting from the output of the CTA-PLS, we have built a second order PLS-SEM model, as hierarchical model [12]. In this framework we can include LVs that represent a "higher-order" of abstraction (HOC). In fact, for our purpose, we will assume goalkeepers' macro-composite performance as extra-latent construct of second order, influenced directly from the others 6 lower order constructs (LOCs). Since the HOC is without any apparent MVs, literature suggested us a recent technique in order to modelling this framework: a mixed two-step approach [6]. In the first step we computed the classical repeated-indicators approach, while in the second one we applied the classical PLS-SEM using the computed scores (of LOCs) as MVs for the HOC. For what concern the structural (inner) model, in our framework it links all $R = 6$ LVs (LOCs) with the HOC, by a linear model (4), where the path coefficients (β_{rq}) are estimated by a factorial scheme (i.e. the correlation between the endogenous and the exogenous LV [11]).

$$\xi_q = \sum_{r=1}^{R} \beta_{rq}\xi_r + \zeta_q \tag{4}$$

For this project the *smartPLS*[4] software and the R software package *seminr* [13] have been used; we carried out a bootstrap validation (i.e. 5000 resampling) for the model in order to assess the path significance. In the next section, preliminary results are shown.

3 Results and discussion

Preliminary CTA-PLS output suggests us the following classification for the LOCs:

- Reflective constructs (i.e. all vanishing tetrads in each block): *attacking, mentality* and *power*.
- Formative constructs (i.e. at least one tetrad does not vanish in each block): *gk_features, movement* and *skill*.

At this point we run the model following the CTA-PLS advice and then we assessed each LV removing problematic MVs [14]:

[4] www.smartpls.com

- Reflective constructs: we removed some MVs with reliability problems (i.e. loadings < 0.7), in particular crossing, heading accuracy and short passing that refers to the *attacking* LV, aggression, vision and penalties relying *mentality*, and jumping, strength and long shot for *power*.
- Formative constructs: here we removed MVs with collinearity problems (i.e. VIF> 5) or outer weights non-significant; agility relying the *movement* construct, whereas diving, positioning and speed for the *gk_features* block.

The final model is showed in Fig. 1: in the light blue circle there are formative constructs, whereas in the light blue rectangles there are reflective constructs; finally, in the white circle there is the HOC. We can see how *GK_Features* (as we expected) have the strongest impact on the macro-composite indicator (i.e. beta coefficient significant and equal to 0.28 for the inner model). It's interesting to note how for each LV the strongest MV (i.e. with highest weight or loading) is a typical variable strictly related with the goalkeepers ability [9], for example: long passing for *skill*, reaction for *movement*, shot power for *power*, positioning for *mentality*, short passing for *attacking*. Other comforting results derived from the GoF index, that is 0.792 (i.e. the geometric mean between the inner and the outer model performances) and from the SRMR (standardized root mean square residual, the difference between the observed correlations and the model-implied correlation matrix), equals to 0.096 (i.e. under the threshold of 0.10) [14].

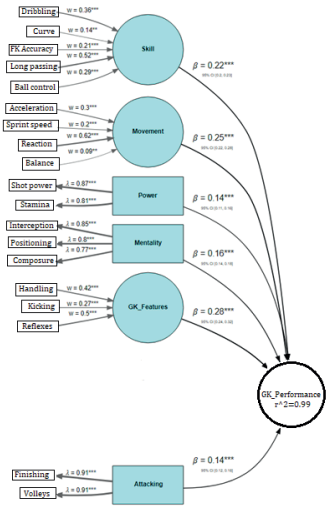


Fig. 1 PLS-SEM GK performance model after 5000 bootstrap resampling.

In order to check the concurrent validity, we compared our scores with some criteria measures (Tab. 1), such as the EA *overall*, wage and players' market value, with interesting results: all medium-high correlations and significant (no one CI 95% contains the zero), the highest between our indicator and the EA *overall*.

Table 1 Correlations of the GK Performance Indicators with three criterion variables.

	GK performance Sept. 2019	CI 95%
EA *overall* Sept. 2019	0.858	[0.826 – 0.884]
Wage Sept. 2019	0.605	[0.532 – 0.669]
Market Value Sept. 2019	0.585	[0.509 – 0.652]

Finally, this model seems to provide comforting results, and at this point for future projects it could be interesting to integrate it in some predictive modelling, such as the expected goal model used in football analytics [7], or to apply CTA-PLS also for movement roles [5]; it should be interesting to compare our model performance respect to a model that considers all constructs as formative or reflective, too.

References

1. Bollen, K.A., Ting, K.f.: A tetrad test for causal indicators. Psychological methods **5**(1), 3 (2000)
2. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the kaggle european soccer database. Statistical Modelling **19**(1), 74–101 (2019)
3. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research **156**(2), 815–830 (2021)
4. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the european soccer leagues. Journal of Applied Statistics **48**(9), 1696–1711 (2021)
5. Cefis, M., Carpita, M.: Football analytics: a higher-order pls-sem approach to evaluate players' performance. Book of Short Papers SIS 2021 pp. 508–513 (2021)
6. Crocetta, C., Antonucci, L., Cataldo, R., Galasso, R., Grassia, M.G., Lauro, C.N., Marino, M.: Higher-order pls-pm approach for different types of constructs. Social Indicators Research **154**(2), 725–754 (2021)
7. Green, S.: Assessing the performance of premier leauge goalscorers. OptaPro Blog (2012). URL http://www.optasportspro.com/about/optaproblog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/
8. Gudergan, S.P., Ringle, C.M., Wende, S., Will, A.: Confirmatory tetrad analysis in pls path modeling. Journal of business research **61**(12), 1238–1249 (2008)
9. Hughes, M.D., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C.: Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position (2012)
10. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika **43**(4), 443–477 (1978)
11. Lohmöller, J.B.: Predictive vs. structural modeling: Pls vs. ml. In: Latent variable path modeling with partial least squares, pp. 199–226. Springer (1989)
12. Sanchez, G.: Pls path modeling with r. Berkeley: Trowchez Editions **383**, 2013 (2013)
13. Shmueli, G., Ray, S., Estrada, J.M.V., Chatla, S.B.: The elephant in the room: Predictive performance of pls models. Journal of Business Research **69**(10), 4552–4564 (2016)
14. Tabet, S.M., Lambie, G.W., Jahani, S., Rasoolimanesh, S.M.: An analysis of the world health organization disability assessment schedule 2.0 measurement model using partial least squares–structural equation modeling. Assessment **27**(8), 1731–1747 (2020)
15. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York pp. 581–591 (1985)

Integrating available Data Sources for Official Statistics

The Use of Administrative Data for the Estimation of Italian Usually Resident Population

L'uso dei dati amministrativi per il conteggio della popolazione residente in Italia

Marco Caputi, Giampaolo De Matteis, Gerardo Gallo, Donatella Zindato

Abstract In the framework of the Permanent Population and Housing Census (PPHC), based on the combined use of survey and register data, Istat adopted a different methodology to produce the 2020 population census counts. Due to the Covid-19 pandemic and the subsequent withdrawal of the sample surveys foreseen by the design of the PPHC, Istat opted for the use of 'administrative signs of life' to estimate the coverage errors of the population register. This has been achieved through the use of classification criteria applied to statistical registers. This process has allowed to produce, solely on the basis of administrative data, municipal population counts by gender, age, citizenship and educational attainment with the same territorial detail of the integrated census data.

Abstract *Nell'ambito del Censimento Permanente della Popolazione e delle Abitazioni, basato sull'integrazione di indagini e dati amministrativi, l'Istat ha adottato una metodologia diversa per il conteggio della popolazione 2020. A causa della pandemia da Covid-19 e della cancellazione delle indagini campionarie previste dal Censimento Permanente, l'Istat ha optato per l'uso dei 'segnali di vita amministrativi' per stimare gli errori di copertura del registro della popolazione. Attraverso l'uso di criteri di classificazione applicati a registri statistici è stato possibile produrre, esclusivamente con i dati amministrativi, stime comunali della popolazione per genere, età, cittadinanza e grado di istruzione con lo stesso dettaglio territoriale dei dati prodotti con l'integrazione di indagini e registri.*

Key words: administrative signs of life, population Census, patterns of usual residence, integration of administrative data

[1] Marco Caputi, Istat; email: caputi@istat.it
Giampaolo De Matteis, Istat; email: dematteis@istat.it
Gerardo Gallo, Istat; email: gegallo@istat.it
Donatella Zindato, Istat; emil: zindato@istat.it

1 The Permanent Population and Housing Census

To replace the decennial census, in 2018 the Italian National Institute of Statistics (Istat) launched the Permanent Population and Housing Census (PPHC), based on the integration of administrative data with information collected from two sample surveys (Areal survey and List survey) conducted annually in self-representative municipalities and every four years, according to a rotation scheme, in non-self-representative municipalities (Falorsi, 2017).

In 2020, Istat achieved the goal of producing a count of the usually resident population by gender, age, citizenship and educational attainment, even if, due to the pandemic crisis, the field surveys were not carried out. To this aim, Istat took advantage of the progress in terms of quality and timeliness achieved by the Registers supporting the official statistical production. The availability of the population base register (hereinafter RBI), of the statistical base register of addresses and of thematic registers, such as those of Occupation and Education, as well as the use of administrative archives held by Public Bodies and Ministries (National Security System archives, Ministry of Economy and Finance archives, Real Estate Register, Pensioners' Register, etc.), made it possible to produce a population count through the integration of administrative data (Istat, 2021).

At the core of the PPHC is RBI, which constitutes an internal information environment supporting Istat statistical production processes, whose main source are the local population registers of Italian municipalities. In particular, RBI is the basis infrastructure for the production of official population statistics and the reference for the extraction of samples for the surveys planned for the Permanent Census and, more generally, for all household surveys. The RBI is updated on an annual basis with reference to the 31st December of each year through the integration of registered individual flows of demographic dynamics (births, deaths, moves of usual residence to and from another municipality or to and from abroad). The application of the MIDEA (Micro-DEmographic Accounting) demographic accounting model makes it possible to exploit the potential of the micro database (flows&stock) to produce more accurate and innovative indicators on demographic dynamics, taking into account the sequence of demographic events experienced by individuals (Istat, 2020).

As in 2018 and 2019, the purpose of the 2020 count was to correct the coverage errors of RBI, by identifying individuals recorded in RBI as usual residents but not found in the other administrative sources (over-coverage), on one hand, and individuals found in the administrative data as usually resident but not recorded as such in RBI (under-coverage) on the other hand. This correction was applied at the micro level, operating through the reclassification of individual records in the Register, defined or not as usually resident on the basis of administrative “signs of life” (Istat, 2021). This was indeed a significant methodological innovation, ensuring

the correspondence in terms of "head count" between the census count and the individual records of usual residents in RBI, differently from 2018 and 2019, when the coverage errors correction was achieved at the macro level, by applying weights to RBI usual residents' records (Istat, 2020; 2021).

2 The use of administrative "signs of life" for the 2020 population census count

For the purposes of the 2020 population count, lacking field survey data, Istat has set up an Integrated Data Base of Usual Residents (hereinafter AIDA). This database collects information from administrative sources other from the local population registers which, under compliance with the provisions of the law on confidentiality, are organized in the Integrated Microdata System (SIM) with the aim of supporting statistical production processes, both for social and economic statistics. The assignment of a unique and constant in time ID code makes it possible to identify each individual and economic unit within the different archives and to build relationships between the different sources, while at the same time guaranteeing the processing of data without making use of direct identifiers.
For the construction of AIDA, the sources relevant to the usually resident population have been selected within SIM and ordered hierarchically by thematic experts and methodologists, to the aim of observing the administrative "signs of life" of individuals usually resident in Italy (Gallo and Zindato, 2021). The AIDA database used for the 2020 population count integrates at the micro level, from the 1st of January 2019 to the 31st December 2020, RBI information with that of the Thematic Registers of Occupation and Education, of the tax returns and social security archives, as well as that of the real estate register.
Through this process, a definition of "signs of life" (better known in the international literature as administrative "signs of life") has been elaborated and has been included in the 2022 General Census Plan. Administrative signs of life" refer to activities carried out by individuals that can be deduced from administrative records and clearly identifiable with reference to the time (e.g. a year) and space (e.g. a municipality) in which they take place. Being self-employed or working for a company, being a civil servant, having a home lease, attending a school or university are examples of direct signs of administrative life. On the other hand, individuals' statuses or conditions, again deducible from administrative records, such as being a recipient of the 'basic income' subsidy or of an old-age pension, or being a dependent family member as a spouse, children or other relative in the tax declaration, are considered indirect signs of life. From this definition, it is also possible to deduce a hierarchical classification of the signs of life:

1. *direct (administrative) signs of life.* Work and study signals, as well as home leases or social welfare benefits from the National Security System are classifiable as direct signs of life with respect to being usual resident in Italy;

these records offer a considerable information detail, i.e. the duration of the activity, its location (municipality and address) and some specific attributes (employment contract, school/course attended, etc.) which are relevant in assessing the strength of the sign of life (Istat, 2021);

2. *indirect (administrative) signs of life.* Income tax return records (tax declarations, tax return filings, etc.) as well as owning a car according to the Cars Public Register or owning a property according to the cadastral archive provide indirect signals of usual residence in Italy. In fact, the 'dependent family member's box' of the tax return filings provide the main relationship between the 'spouse' or 'child/children' and the declaring relative Since these signs of presence in Italy are inferred 'indirectly' from the declaration of an income recipient, it has been decided to classify them as indirect signs; the same for possession of a car or a house, which is not a legal requirement;
3. *other types of indirect (administrative) signs of life* are those which refer to the relationship to the reference person within the household (according to RBI). In this case the relationships taken into account are those between the reference person and, respectively, the 'spouse' and the 'children.

By integrating the information available in SIM, it is possible to reconstruct the demographic profile of the usual residents in Italy, namely according to:

a) date of birth;
b) gender;
c) citizenship;
d) country of birth.

As for citizenship, if the information was not available in none of the archives used (RBI and Permits of stay), reference was made to the information on the country of birth. When, for an individual, information for the same variable was not consistent among the different sources, the one coming from the source hierarchically superior was taken. Generally speaking, RBI and the Tax Register are the most complete sources for what concerns the date of birth, the gender and the country of birth; while the latter is also very useful to the aim of determining the usual residence of individuals, thanks to the variable 'Municipality' of the fiscal domicile.

3 The continuity patterns of administrative "signs life" to identify the usual residents in Italy

The AIDA integration process involves the processing of data from more than forty administrative archives, each containing basic information on individuals' signs of life (events) and covering several years. For each administrative event, is recorded also the information on the location of the event itself, by means of the province and municipality codes.

This is important in light of the definition of usual residence adopted by the European Union Regulation 1260/2013, which defines the usual residence the place where the person has spent at least 12 months, before the reference date, or less than 12 months before the reference date but with the intention of staying there at least 12 months. On the basis of this definition, the period from the 1st of January 2019 to the 31st of December 2020 (which is the reference date of the 2020 Permanent Population Census) was chosen for the analysis of the signs of life (Gallo, Zindato, 2018).

The longitudinal observation of direct signals over two years allows us to capture specific profiles of presence of individuals on the territory. These profiles of presence in Italy in some cases clearly identify usual residents in Italy, while in others the administrative "signals of life" are of low intensity, or identify seasonal workers, i.e. (in both cases profiles that cannot be associated with the usual resident definition).

As shown in Table 1, each sign of life is associated with a specific individual and a specific place. For example, if in the period under consideration, for an individual identified by code '0000018' a record is found in the Occupation Register and another in a source related to the study, and both signs are located in the municipality of Agliè, we will have a single sign of life located in that specific municipality. This sign, however, is marked both by an attribute that allows us to trace the individual in both archives, and by an attribute related to the duration of the work and study activity. Therefore, the algorithm that processes the direct signs of work or study of AIDA produces a string that summarises the individual profile, showing (in red in the third column) a sign of work in the first position and a sign of study in position 9 and another string, in the last column of the table which shows that, for the specific case, the direct sign of life in the source of work or study is present for all months.

After the processing of the direct signs of life and the determination of the prevalent municipality where the study or work activity has been carried out, AIDA process integrates the individuals with direct signs of life with RBI. More precisely, for each municipality are identified all individuals with usual residence in Italy (i.e. with direct signs of life) who are not recorded in RBI and individuals recorded in RBI without direct signs of life. The next step consists in comparing the indirect signs of life (derived by the Tax Registry) related to "dependent family members" and owners of a car or of a real estate unit with the individuals recorded in RBI who were found to have no signs of life in the previous steps.

Table 1 – Example of individual profile of direct signs of life of work/study

Type of information	**Municipality code + individual code**	**Signs of life of work/study over the relevant period**	**Specific information**	**{month1--month24}**
Description	{Identifier}	{Sequence of the sources: every source has a specific position; 1= presence in the given source}	{Additional information}	{Monthly presence/absence }
Example	*Municipality of Agliè - Individual 18*	*UniEmens (pos.1)+Enrolled university (pos.9)*	*Permanent contract*	*Presence in every month*

Example data	001-001-0000018	1000000010	0----1-----	11111111111111111111

Source: Istat, 2021

Finally, the last step identifies individuals with neither direct nor indirect signs of life i.e. the over-coverage of RBI. On this sub-group of population, however, a further check is performed in order to identify "spouses" of reference persons who have direct signs of life. These, who would otherwise end up in the set of individuals classified as RBI over-coverage as lacking direct or indirect signs of life, as 'spouses' of individuals with direct signs of life are instead considered as usual residents. In fact, as extensively documented in international scientific literature, the approach followed by Istat has been not to limit the focus solely and exclusively on signs of life available from administrative sources, but to exploit the richness of administrative archives according to a *Knowledge Discovery from Databases* process (Chieppa et al, 2018); in short, by using a structured and iterative process, in which part of the variables to be analysed are constructed ongoing, as the processing of administrative signals proceeds.

4 An iterative process for the fine tuning of deterministic criteria

In the previous sections we have described how AIDA process was used in order to correct coverage errors of RBI, namely for identifying two population sub-groups:
1) individuals with direct signs of life of at least one year but not resident according to RBI at 31.12.2020, which represent the under-coverage of the municipal registers at the same date;
2) individuals usually resident according to RBI as of 31.12.2020 but without direct or indirect signs of life in other administrative sources, which represent the over-coverage of the municipal registers.
As mentioned above, based on the results of exploratory analyses conducted by integrating administrative and survey data of the 2018 and 2019 waves, deterministic criteria have been defined for classifying individuals belonging to specific subpopulations.
With reference to under-coverage, foreign citizens have been considered usual residents if having direct signs of life (i.e. having a valid permit of stay was not sufficient for being counted as usual resident). Furthermore, foreign citizens with a direct sign of life located in a border municipality have not been considered usually resident in Italy for obvious reasons connected to border movements for work or study. In any case, it is worth noting that individuals for whom the information on the place of the sign of life (place of work, study or of the rental contract), was not available were not counted as under-coverage, being not possible to assign them a municipality of usual residence
To the aim of identifying over-coverage, following the iterative approach and hierarchical logic, new variables at both individual and family level have been

constructed for individuals with no direct or indirect signs of life, in order to reinforce the absence of signs of life resulting from the AIDA output (and, therefore, classify them as over-coverage) or instead validate their presence in RBI (even in absence of signs of life in AIDA). As a result of this fine tuning, residents in RBI with no signs of life (neither direct nor indirect) have been confirmed as usual residents if members of a family nucleus whose reference person works or perceives a pension, of households with children aged less than 14 attending school in the same municipality, of households where a household member owns or rents a property, or if aged 68 or over, non-perceiving a pension and owning a car (Gallo, Zindato, 2021).

Finally, all individuals of RBI resident in very small municipalities have been confirmed as usual residents, since longitudinal indicators of maintenance show that in small municipalities municipal registers are quite accurate, as confirmed also by several exploratory analyses conducted by Istat researchers over the last 5 years.

The same was done for the elderly (people of at least 98 years of age), whose usual residence in RBI was confirmed due to the high quality of RBI data for this subpopulation, and for residents in institutional households (following a verification activity carried out by Istat in the first months of 2021, during which both the addresses and the corresponding population aggregates were extracted from RBI and submitted to the validation by Municipal Census Offices).

Following this approach, population counts were obtained on the basis of which the population census was calculated with reference date to 31 December 2020.

Table 2 shows the comparison between the individuals with administrative signs of life provided by the AIDA archive and the individuals registered as usual residents in the RBI population register. This makes it possible to compute the amount of population that can only be deduced from the AIDA archive, the over- and under-coverage population of the local population registers and the population census counts.

Table 2 – Population census counts and total population at 31 December 2020 as a result of AIDA versus RBI integration

Description of outcomes	Type of register or Archive	Total population counts	Population census counts
Population correctly placed in RBI	RBI vs AIDA	58,713,660	Yes
Under-coverage at national level	Only in AIDA	324,932	Yes
Over-coverage at national level	Only in RBI	1,005,908	No
Uncertain units	*Only in RBI*	*197,621*	*Yes*
Uncertain units	*Only in AIDA*	*288,211*	*No*
Under/over-coverage at local level	AIDA vs RBI	20,423	No
Population not entered in the count	AIDA with unusable signs	1,410,497	No
Usual resident population	AIDA vs RBI	59,236,231	Yes
Total population	AIDA	61,961,252	

Source: Istat, 2022

AIDA's independent archive identifies almost 62 million individuals with administrative life signs but of these only just over 59,2 million can be considered to be usually resident in Italy. According to the comparison with RBI, the population correctly registered in RBI amounts to 58,7 million. The national under-coverage of RBI is equal to almost 325,000, while the over-coverage of population registers is just over 1 million.
In the comparison with RBI, however, there is a population subgroup of just under 200,000 for which the administrative signs of life do not clearly identify whether these individuals are usually residents or not as the signs are very weak. As these people are registered in the population registry and given the uncertain signs provided by the administrative sources, the conservative criterion was chosen for counting purposes and therefore they were considered in the final population census counts. On the other hand, those who are not registered in the RBI and whose administrative signs are in any case uncertain have not been considered in the census count.
Finally, it should be noted that the AIDA archive identifies a population group of just over 1,4 million people for whom the administrative signs are very weak or not well localized and as such have been excluded from the census population count.
A limitation of the administrative signs of life is the identification of the misplacement error of population register (i.e. those individuals who are registered in a municipality but their usual residence turns out to be elsewhere). At present, Istat is acquiring energy consumption archives (smart meters data) that can provide very significant objective assessment elements with respect to the real place of usual residence (Albert, Rajagopal).

5 Conclusions

The PPHC has been designed according to Istat modernization program, which places the integrated system of statistical registers at the core of statistical production. The role of field surveys in this system is to support registers, in the broad sense of assessing their quality and to add information that is missing, incomplete or of insufficient quality. This allows the yearly availability of detailed census statistics.
At the core of the PPHC is the population register, while two sample surveys are conducted annually to evaluate and correct the coverage errors of RBI and collect the data needed to produce Census outputs.
During the first two waves (2018 and 2019), due to fieldwork quality issues, administrative ‘signs of life’ classified according to duration patterns, type and reliability of the source were integrated in the estimation process to correct the survey under-coverage. Individuals in RBI who had not been enumerated were thus considered usually resident if associated with strong administrative ‘signs of life’.
The use of administrative data has been further accelerated because of the cancelation of the field surveys for the 2020 wave due to the pandemic. In order to

predict population counts at municipal level for age, sex and citizenship, a process integrating available data from the past waves and 'signs of life' was set up to establish deterministic criteria applied to individual records in RBI.
This obliged push towards a larger use of administrative data for a rethinking of the statistical framework for the quality assessment of the estimation processes of the PPHC. To this aim the processing of 2021 Census data, currently ongoing, will be of great importance, given the availability of both survey data and administrative ones. Comparisons among different estimation models, the integration of administrative and survey data, the evaluation of fieldwork quality are all important areas of investigation to improve the design of the future PPHC cycles.
As to the second cycle of the PPHC, starting from 2022, the field surveys will continue to play a crucial role for assessing the quality of administrative sources but, unlike the first cycle, the two surveys will have completely different purposes. The Areal survey will be aimed at measuring the quality of population counts produced with "signs of administrative life " and at providing data for the improvement of deterministic criteria for the use of "administrative signs of life", while the List survey will continue to provide data for information that in registers is missing, incomplete or of insufficient quality.

References

1. Albert, A., Rajagopal, R.: Smart Meter Drive Segmentation: What Your Consumption Says About You. IEEE Transactions on power systems (2013): 4019-4030.
2. Chieppa, A., Gallo, G., Tomeo, V., Borrelli, F., Di Domenico, S.: Knowledge discovery for inferring the usually resident population from administrative registers. In MPS (2018), DOI: 10.1080/08898480.2017.1418114 To link to this article: https://doi.org/10.1080/08898480.2017.1418114
3. Falorsi, S.: The Italian experience on the Population and Housing Census: the Master Sample. Presentation at UNECE Meeting, October 4-6 (2017) at the following link: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/Day2_1130_Italy_falorsi_presentation.ppt__1_.pdf
4. Gallo, G., Zindato, D.: Annex H. Italy case study, in UNECE (2018), Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses, Geneva (2018) at the following link: https://unece.org/guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0.
5. Gallo, G., Zindato, D.: Italy: The combined use of survey and register data for the Italian Permanent Population Census count in UNECE, Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses (endorsed by the 69th plenary session of the Conference of European Statisticians), October (2021) at the following link: https://unece.org/statistics/publications/CensusAdminQuality.
6. Istat: Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età cittadinanza, grado di istruzione e condizione professionale per gli anni 2018 e 2019 (2020) at https://www.istat.it/it/files/2020/12/NOTA-TECNICA-CENSIPOP.pdf
   Istat: Nota tecnica sulla produzione dei dati del Censimento Permanente: la popolazione residente per genere, età, cittadinanza e grado di istruzione al 31.12.2020 (2020) at https://www.istat.it/it/files//2021/12/NOTA-TECNICA-CENSIMENTO-POPOLAZIONE_2020.pdf

New frontiers for the analysis of the territorial economic phenomena

An empirical tool to classify industries by regional concentration and spatial polarization

Un metodo empirico per la classificazione delle industrie in base al grado di concentrazione regionale e di polarizzazione spaziale

Diego Giuliani, Maria Michela Dickson, Flavio Santi, Giuseppe Espa

Abstract Traditional measures of geographical concentration of industries based on regional data (such as Gini, Herfindhal, and Ellison-Glaeser indices) do not consider the information about the spatial positions of regions. This implies their insensitivity to regions' spatial order and inability to account for neighboring effects. As an attempt to cope with this limitation, a recent stream of literature [7, 4, 9, among others] has focused on developing measures that quantify the degree of concentration of an industry while adjusting for spatial connections among regions. Following the idea that a single measure cannot fully describe the characteristics of an industry in terms of both concentration and spatial interactions, this paper proposes an alternative approach that measures the two dimensions jointly and allows for the classification of economic sectors into meaningful types of geographical configurations.

Abstract *Le misure tradizionali di concentrazione geografica delle industrie basate su dati regionali (quali gli indici di Gini, di Herfindhal e di Ellison-Glaeser) non considerano l'informazione riguardante le posizioni spaziali delle regioni. Ciò implica che siano invarianti rispetto all'ordine spaziale delle regioni e, dunque, non tengano conto degli effetti di vicinato. Alla luce di tale limitazione, una letteratura recente [7, 4, 9, tra gli altri] si è occupata di sviluppare indici che consentano di misurare il grado di concentrazione di un'industria controllando per le connessioni*

Diego Giuliani
Department of Economics and Management, University of Trento, via Inama 5, Trento
e-mail: diego.giuliani@unitn.it

Maria Michela Dickson
Department of Economics and Management, University of Trento, via Inama 5, Trento
e-mail: mariamichela.dickson@unitn.it

Flavio Santi
Department of Economics, University of Verona, via Cantarane 24, Verona
e-mail: flavio.santi@univr.it

Giuseppe Espa
Department of Economics and Management, University of Trento, via Inama 5, Trento
e-mail: giuseppe.espa@unitn.it

spaziali tra le regioni. In base all'idea che una singola misura non possa fornire una descrizione completa delle caratteristiche di un'industria in termini sia di concentrazione e sia di interazioni spaziali, questo lavoro propone un approccio alternativo che misura le due dimensioni congiuntamente e permette di classificare i settori economici in categorie rilevanti di configurazione geografica.

Key words: Geographical concentration, Spatial polarization, Bivariate confidence regions

1 Classifying industries considering regional concentration and spatial polarization of economic activities

To develop an approach to classify industries into valid geographical configurations, we follow the idea formalized by [1] and reprised by [10, 6, 11] of grouping industries according to the combination of the values of a geographical concentration index and a spatial autocorrelation index. More specifically, let $\mathbf{s} = [s_1, s_2, \ldots, s_n]$ represent the vector containing the regional shares of employment in a given industry, where n is the total number of regions in the economy of interest. Then, let $C(\mathbf{s})$ denote a proper index of geographical concentration of $\mathbf{s}$ characterized by a benchmark value, B_c, such that $C(\mathbf{s}) > B_c$ if the industry is *geographically concentrated* while $C(\mathbf{s}) < B_c$ if the industry is *geographically dispersed*. Moreover, let $A(\mathbf{s})$ indicate a proper index of spatial autocorrelation of $\mathbf{s}$ characterized by a benchmark value, B_a, such that $A(\mathbf{s}) > B_a$ if the industry is *positively spatially autocorrelated* while $A(\mathbf{s}) < B_a$ if the industry is *negatively spatially autocorrelated*.

By considering both indices together, seven different notable geographical configurations can be readily identified:

1. When the employment of an industry is both geographically concentrated and positively spatially autocorrelated, that industry can be classified as a *strongly polarized concentrated* industry (*sP-Con*) as it is characterized by concentration within regions but also by strong polarization of regions in which the industry is concentrated.
2. An industry that tends to agglomerate mainly by concentrating within regions without sprawling over a relatively large number of neighboring regions can be described as a *mildly polarized concentrated* industry (*mP-Con*). This geographical configuration implies that industry employment is geographically concentrated but not spatially autocorrelated.
3. An industry that is overrepresented in a few isolated regions, as evidenced by geographically concentrated and negatively spatially autocorrelated employment, can be appropriately denoted as a *weakly polarized concentrated* industry (*wP-Con*).

4. An industry that is geographically dispersed and positively spatially autocorrelated may be categorized as a *strongly polarized dispersed* industry (*sP-Dis*) as it is depicted by employment dispersion in neighboring regions.
5. An industry that is geographically dispersed but not spatially autocorrelated may be labeled as a *mildly polarized dispersed* industry (*mP-Dis*) as it shows employment dispersion within regions that are not spatially polarized.
6. An industry that is geographically dispersed and negatively spatially autocorrelated could be treated as a *weakly polarized dispersed* industry (*wP-Dis*) as it shows employment dispersion in a few isolated regions.
7. Finally, a *randomly located* industry (*Rand*) is an industry in which employment does not follow any notable geographical pattern as it is randomly spatially distributed.

Table 1 and Figure 1 summarize the relationship between the geographical configurations and the values of $C(\mathbf{s})$ and $A(\mathbf{s})$.

Table 1 Geographical configurations and corresponding values of $C(\mathbf{s})$ and $A(\mathbf{s})$

Geographical configuration	$C(\mathbf{s})$	$A(\mathbf{s})$
sP-Con	$> B_c$	$> B_a$
mP-Con	$> B_c$	$\approx B_a$
wP-Con	$> B_c$	$< B_a$
sP-Dis	$< B_c$	$> B_a$
mP-Dis	$< B_c$	$\approx B_a$
wP-Dis	$< B_c$	$< B_a$
Rand	$\approx B_c$	$\approx B_a$

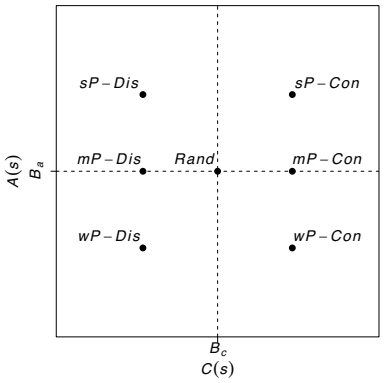


Fig. 1 Graphical representation of the relationship between the geographical configurations and corresponding values of $C(\mathbf{s})$ and $A(\mathbf{s})$

1.1 Index of geographical concentration

To classify industries according to the proposed framework, a good candidate as an index of geographical concentration, $C(\mathbf{s})$, is the Ellison and Glaeser (EG) index [3]. In the class of spatial concentration indices based on regionally aggregated data, EG is the one with properties closest to those of an "ideal" index; indeed, it controls for overall agglomeration of manufacturing, it is robust to industrial concentration from small numbers of firms and allows comparability across industries and levels of spatial aggregation of data [2]. For a given industry, it can be expressed as

$$\gamma_{EG} = \frac{G - H(1 - \mathbf{x}'\mathbf{x})}{(1-H)(1-\mathbf{x}'\mathbf{x})},$$

where H is a Herfindahl index measuring concentration of the industry firm's employment distribution[1], $G = (\mathbf{s}-\mathbf{x})'(\mathbf{s}-\mathbf{x})$ is the Gini statistic measuring the raw industry geographical concentration and $\mathbf{x}' = [x_1, x_2, \ldots, x_n]$ is a vector containing the regional shares of total employment.

The benchmark value of γ_{EG} representing spatial randomness is 0. Positive (negative) values measure industry employment concentration (dispersion) beyond that due to randomness.

1.2 Index of spatial autocorrelaton

The natural choice for an index $A(\mathbf{s})$ of spatial autocorrelation of regional employment is the popular Moran's I index [8]. For a given industry, it can be expressed as

$$I = \frac{\mathbf{q}'\mathbf{W}\mathbf{q}}{\mathbf{q}'\mathbf{q}},$$

where $\mathbf{q} = \mathbf{s} - \bar{\mathbf{s}}$ and $\mathbf{W} = (w_{ij})$ is a nonnegative row-standardized spatial weight matrix such that w_{ij} indicates how close is region j to region i; in particular, a large value of w_{ij} means that j is a neighbor of i.

The benchmark value of I representing the absence of spatial autocorrelation is $-1/(n-1)$. Values higher (lower) than $-1/(n-1)$ measure positive (negative) spatial autocorrelation.

[1] Therefore, $H = \mathbf{z}'\mathbf{z}$, where $\mathbf{z} = [z_1, z_2, \ldots, z_K]$ is a vector containing the K firms' shares of industry's total employment.

1.3 Bivariate confidence regions

To assign an industry to one of the seven geographical configurations, one must look at how far the EG and Moran's I indices are from their respective benchmarks. Verifying if they are *far enough* requires taking into account the statistical significance of the estimated γ_{EG}- and I-values. However, since the indices are correlated, separate hypothesis tests may not be appropriate. To take into account that γ_{EG} and I are related, we suggest a procedure to obtain their bivariate confidence region. However, since the joint sampling distribution of the two indices is unknown, and their covariance is analytically intractable, the confidence region can be derived by the bootstrap method. In particular, the joint bootstrap distribution can be obtained through block-wise resampling, defining the industry-region combinations as blocks. For each replication, a sample of firms is drawn randomly and independently with replacement from each block.

2 An illustrative application

To verify that the proposed approach provides reasonable results, we apply it to a simulated dataset which is meant to emulate the geographical configurations postulated in Section 1. The details of how we generated the data are as follows.

We specified $n = 49$ regions, as cells of a 7-by-7 lattice, and 7 industries, each with a total employment of 5000 and a total number of firms of 200 units. For each industry, the values of the vector $\mathbf{z}$ of the 200 firms' employment shares in the industry's total employment have been generated from a Beta distribution, $\mathrm{B}(\alpha_1, \beta_1)$, and then scaled by their sum to add to one. Therefore, $5000\mathbf{z}$ provides the number of employees for each of the 200 firms in the industry. Secondly, for each industry, the 200 firms are randomly assigned to the 49 regions (say the cells) with probabilities generated from a Beta distribution, $\mathrm{B}(\alpha_2, \beta_2)$, and then scaled by their sum to add to one. By setting specific values for α_1, β_1, α_2 and β_2, it is possible to obtain regionally distributed firm-level employment data characterized by levels of geographical concentration that are consistent with the seven geographical configurations. Thirdly, for some industries, to make $\mathbf{s}$ characterized by a prescribed spatial autocorrelation measured by a given value of the Moran's I index, I^*, the cells have been randomly swapped according to the cell-swapping algorithm by [5].

By setting properly the parameters, the data for the 7 industries have been generated to recreate the seven archetypical geographical configurations. In particular,

- *Industry 1*(randomly located): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 2$, $\beta_2 = 2$.
- *Industry 2*(strongly polarized dispersed): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 1$, $\beta_2 = 1$, $I^* \geq 0.4$.
- *Industry 3*(strongly polarized concentrated): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 2$, $\beta_2 = 6$, $I^* \geq 0.4$.
- *Industry 4*(mildly polarized dispersed): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 1$, $\beta_2 = 1$.
- *Industry 5*(mildly polarized concentrated): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 1$, $\beta_2 = 6$.
- *Industry 6*(weakly polarized dispersed): $\alpha_1 = 20$, $\beta_1 = 20$, $\alpha_2 = 1$, $\beta_2 = 1$, $I^* \leq -0.4$.
- *Industry 7*(weakly polarized concentrated): $\alpha_1 = 30$, $\beta_1 = 30$, $\alpha_2 = 2$, $\beta_2 = 6$, $I^* \leq -0.4$.

Figure 2 depicts the 95% and 99% bootstrap confidence regions of the estimated γ_{EG}- and I-values for the seven industries based on 9999 replications. The graph shows that the procedure assigns the industries to the correct geographical configurations.

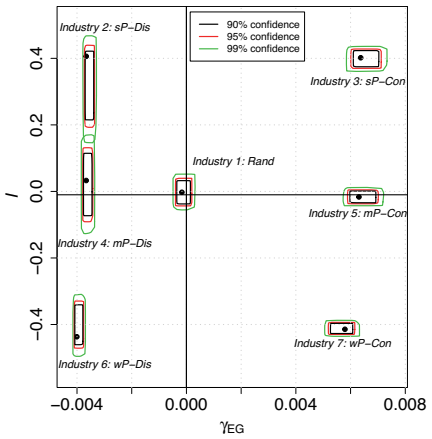


Fig. 2 Bivariate confidence regions to classify industries into geographical configurations

References

1. Arbia, G.: The role of spatial effects in the empirical analysis of regional concentration. J. Geogra. Syst. **3**(3), 271 – 281 (2001)
2. Duranton, G., Overman, H.G.: Testing for Localization Using Micro-Geographic Data. Rev. Econ. Stud. **72**(4), 1077–1106 (2005)
3. Ellison, G., Glaeser, E.L.: Geographic concentration in u.s. manufacturing industries: A dartboard approach. J. Political Econ. **105**(5), 889–927 (1997)
4. Ferrante, M., Magno, G.L.L., Cantis, S.D., Hewings, G.J.: Measuring spatial concentration: A transportation problem approach. Pap. Reg. Sci. **99**(3), 663–682 (2020)
5. Goodchild, M.F.: Algorithm 9: Simulation of autocorrelation for aggregate data. Environ. and Plan. A: Econ. and Space **12**(9), 1073–1081 (1980)
6. Guillain, R., Gallo, J.L.: Agglomeration and dispersion of economic activities in and around paris: An exploratory spatial data analysis. Environ. and Plan. B: Plan. and Des. **37**(6), 961–981 (2010)
7. Guimarães, P., Figueiredo, O., Woodward, D.: Accounting for neighboring effects in measures of spatial concentration. J. Reg. Sci. **51**(4), 678–693 (2011)
8. Moran, P.: Notes on continuous stochastic phenomena. Biometrika **37**(1/2), 17–23 (1950)
9. Panzera, D., Cartone, A., Postiglione, P.: New evidence on measuring the geographical concentration of economic activities. Pap. Reg. Sci. **101**(1), 59–79 (2022)
10. Sohn, J.: Information technology in the 1990s: More footloose or more location-bound? Pap. in Reg. Sci. **83**(2), 467 – 485 (2004)
11. Sohn, J.: Industry classification considering spatial distribution of manufacturing activities. Area **46**(1), 101–110 (2014)

Comparing Non-Compensatory Composite Indicators: A Case Study Based on SDG for Mediterranean Countries

Un confronto tra indicatori compositi non compensativi: un'applicazione agli SDG per i Paesi del Mediterraneo

Francesca Mariani, Mariateresa Ciommi, Maria Cristina Recchioni, Giuseppe Ricciardo Lamonica, and Francesco Maria Chelli

Abstract Composite indicators provide a summary picture of multidimensional phenomena and the corresponding rankings facilitate evaluations and comparisons over time and space. Standard composite indicators often assume compensability among the indicators. We argue that the compensability hypothesis needs to be restricted, especially when analysing economic, social, and environmental aspects.
In this paper, we start with the simplest non-compensatory index, namely the geometric mean, and we introduce a new aggregation method. The method, called the weighted penalized geometric mean, is a generalization of the penalized geometric mean used to consider weights. The method introduces a penalty in the weighted geometric mean in terms of the (horizontal) variability of the normalized indicators transformed via the zero-order Box-Cox function. To illustrate the appeal of our proposal, we compare the above-mentioned non-compensatory approaches by proposing an application to selected targets of the Sustainable Development Goals (SDGs) for Mediterranean countries.

Francesca Mariani
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: f.mariani@univpm.it

Mariateresa Ciommi
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: m.ciommi@univpm.it

Maria Cristina Recchioni
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: m.c.recchioni@univpm.it

Giuseppe Ricciardo Lamonica
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: g.ricciardo@univpm.it

Francesco Maria Chelli
Department of Economics and Social Sciences, Università Politecnica delle Marche, Ancona, Italy
e-mail: f.chelli@univpm.it

Abstract *Gli indicatori compositi forniscono un quadro riassuntivo dei fenomeni multidimensionali e le rispettive classifiche facilitano le valutazioni e i confronti sia nel tempo che nello spazio. Gli indicatori compositi standard spesso presuppongono l'ipotesi di compensabilità tra gli indicatori. Tuttavia, riteniamo che l'ipotesi di compensabilità debba essere limitata soprattutto quando si analizzano aspetti economici, sociali e ambientali.*
In questo lavoro, partendo dal più semplice indice non compensativo, ovvero la media geometrica, introduciamo un nuovo metodo di aggregazione. Il metodo, detto media geometrica penalizzata ponderata, è una modifica della media geometrica penalizzata tramite l'introduzione di pesi. Il metodo si basa su una penalizzazione per le unità con valori sbilanciati degli indicatori, misurata in termini di variabilità (orizzontale) degli indicatori normalizzati e opportunamente scalati e trasformati tramite la funzione Box-Cox di ordine zero. Per illustrare la nostra proposta, confrontiamo gli approcci non compensativi sopra menzionati tramite alcuni targets selezionati degli Obiettivi di Sviluppo Sostenibile (SDGs) per i Paesi del Mediterraneo.

Key words: Composite indicators, Geometric mean, Non-compensatory approach, Penalty, Weights, SDG.

1 Introduction

In this paper, we introduce the weighted version of the penalized geometric mean defined in Mariani and Ciommi (2022) [7]. In analogy with the Mazziotta-Pareto composite indicator (Mazziotta and Pareto, 2016 [8]), the weighted penalized geometric mean is obtained by penalizing the weighted geometric mean through a factor that measures the information loss that occurs when we use the weighted geometric mean of the variables instead of the variables themselves. The penalty factor is the weighted version of the penalty factor used in Mariani and Ciommi (2022) [7] to penalize the geometric mean.

The aim of this paper is to investigate the role of weights in the aggregation and their effect on the ranking. Moreover, a comparison with geometric and weighted geometric mean approaches is presented. To illustrate the different role played by weights, we first compute the geometric mean (GM) and the penalized geometric mean (pGM) as introduced in Mariani and Ciommi (2022) [7] and we define a weighted version for both methods, namely the weighted geometric mean (wGM) and the weighed penalized geometric mean ($wpGM$).

We apply the above-mentioned non-compensatory approaches to 11 indicators belonging to the Sustainable Development Goals (SDGs) and related to agro-food aspects, as discussed in Casini et al. (2019) [3]. Motivated by the fact that environmental, economic, and social issues are a severe challenge for the sustainability of the agro-food system (Sachs et al., 2019 [9]) and that Mediterranean countries

are not on track for achieving the SDGs goals according to an analysis of overall country ranking[1], we decide to focus on 17 Mediterranean countries.

The rest of the paper is organized as follows. Section 2 illustrates the methodology. Section 3 describes the results of the application of the four methods and Section 4 contains the conclusion.

2 The weighted penalized geometric mean approach

In this section, we illustrate the two aggregation methods used below. The two methods are penalized versions of the geometric mean and are derived in analogy with the Mazziotta-Pareto approach (Mazziotta and Pareto, 2016 [8]). The penalized geometric mean is introduced in Mariani and Ciommi (2022) [7] and here we extend it to include the weights.

We consider m normalized variables and n units. For each unit i, $i = 1,2,\ldots,n$, and use z_{ij} to denote the value of normalized variable j for unit i. We use wGM to denote the composite indicator aggregating the normalized variables through the weighted geometric mean. The value of the composite indicator associated with the $i-$th unit is then given by

$$wGM_i = \prod_{j=1}^{m} z_{ij}^{w_{ij}}, \quad i = 1,2,\ldots,n, \tag{1}$$

where $w_{ij} \in [0,1]$ is the weight attributed to the j-th indicator for the i-th unit, $i = 1,2,\ldots,n$ and such that $\sum_{j=1}^{m} w_{ij} = 1$.
For $i = 1,2,\ldots,n$ the composite indicator wGM_i is the solution to the following optimization problem (Berger and Casella, 1992 [1]) :

$$\min_{\xi \in \mathbb{R}_+} F(\xi) = \sum_{j=1}^{m} w_{ij}(\ln z_{ij} - \ln \xi)^2, \tag{2}$$

where $\ln \xi$ is the zero-order Box-Cox transformation (Box and Cox, 1964) [2]. That is, $\ln wGM_i$ is the best least-squares fit of the normalized variables transformed via the logarithm function $\ln(z_{i1})$, $\ln(z_{i,2})$, ..., $\ln(z_{im})$, weighting the contribution of each transformed variable $\ln(z_{i1})$ with the corresponding weight w_{ij}, $j = 1,2,\ldots,m$.

Function F in (2) is the sum of the squared residuals obtained by approximating the transformed variables $\ln z_{i1}$, $\ln z_{i2}$, ..., $\ln z_{im}$ by $\ln \xi$. The value of function F in (2) at the optimum wGM_i measures the loss of information that occurs when the composite indicator wGM is used rather than a set of the individual indicators $\{z_j\}_{j=1}^{m}$. This information loss is measured by the (horizontal) variability of the normalized variables transformed via the logarithm function, that is,

[1] See https://dashboards.sdgindex.org/rankings.

$$wS_{0,i}^2 = F(wGM_i) = \sum_{j=1}^{m} w_{ij}(\ln z_{ij} - \ln wGM_i)^2, \quad i = 1,2,\ldots,n. \tag{3}$$

The size of (3) determines the reliability of the estimate $\ln wGM_i$ and, as a consequence, the weighted geometric mean wGM_i.

Therefore, for equal weighted geometric means, the units should be ranked according to the value of (3). Specifically, we expect units i_1 and i_2 with the same value of weighted geometric mean $wGM_{i_1} = wGM_{i_2}$, with $wS_{0,i_1}^2 > wS_{0,i_2}^2$, to be assigned with a different value of the composite indicator, which is larger for unit i_2 in the case of positive polarity and smaller otherwise. In other words, a good composite indicator should distinguish among units with same weighted geometric mean, penalizing the units with greater loss of information.

To consider the loss of information summarized in equation (3), we follow what was done in Mariani and Ciommi (2022) [7] for the penalized geometric mean and we define the penalized weighted geometric mean (pwGM) relative to the i-th unit as follows:

$$pwGM_i^{\pm} = wGM_i \exp\left\{\pm wS_{0,i}^2\right\}, \quad i = 1,2,\ldots,n. \tag{4}$$

In (4), we choose the sign $+$ and $-$ for negative and positive polarity, respectively. When we set $w_{ij} = 1/m$, $j = 1,2,\ldots,m$, in (3),(4), we have the penalized geometric mean (pGM) of Mariani and Ciommi (2022) [7].

3 Application to SDGs

To illustrate the appeal of our proposal, we describe the results obtained by computing the geometric mean and its weighted version, namely, *GM* and *wGM*, respectively, and the penalized version of the geometric mean as proposed by Mariani and Ciommi (2022) [7] and its weighed version, namely *pGM* and *pwGM*, respectively.

Since the purpose is only illustrative, we use data collected by Casini et al. (2019) [3] to focus not on data collection, but on the aggregation and, in particular, on weighting. Thus, the data refer to 11 SDGs (see Casini et al. 2019 [3], Table 8 and Appendix A) related to four SDG domains concerning agro-food sustainability: Food security and Sustainable Agriculture (SDG2), Clean Water and Sanitation (SDG6), Sustainable Consumption and Production Patterns (SDG12), and Sustainable Management of Terrestrial Ecosystems (SDG15).[2]

[2] List of variables: 1) Overweight population; 2) Land use; 3) GHG emissions (total) per sq. km; 4) Cereal yield; 5) Agriculture value added; 6) Fertilizer consumption; 7) Crop water productivity; 8) Annual freshwater withdrawal for agriculture; 9) Population using safely managed water services (rural); 10) Population using safely managed sanitation services (rural); 11) Research and development expenditure.

The data are used to compare the performance of 17 Mediterranean countries, namely Algeria, Croatia, Cyprus, Egypt, France, Greece, Israel, Italy, Jordan, Lebanon, Malta, Morocco, Portugal, Slovenia, Spain, Tunisia, and Turkey.

3.1 Two computational issues: normalization and weights

To ensure comparability of the data across the selected indicators, a normalization step has been proposed. Since the starting point is the geometric mean, classical max-min methods cannot be applied in their original form since they produce at least one zero element for each indicator, which could lead to multiple zero values in the aggregate index. Therefore, following de la Cruz and Kreft (2018) [6], the data are first normalized in the interval $[0,1]$ applying the max-min methods (after considering the polarity in order to get all the indicators positively related to the phenomenon under analysis). A 1 is then added to the final values in order to avoid zeros. Finally, after computing the geometric mean of this shifted data, 1 is subtracted to yield the final results.

To capture the vertical variability, we add weights. Since a more unequal distribution of an elementary indicator among countries implies a greater weight for that indicator (Chelli et al. 2015 [4]), we follow Ciommi et al. (2017) [5] by weighting according to the Gini index of elementary indicators across countries.

3.2 Results

Table 1 reports the results of the four methods and the corresponding rankings. The results show that European Mediterranean countries (MCs) generally tend to perform better in the overall index compared to non-European Mediterranean countries, which occupy the lowest positions. Moreover, the results obtained by applying the geometric mean and its modifications are in line with the results of Casini et al. (2019) [3].

Finally, to compare the role of the weights, we compute the relative contribution of the weights on *GM* and *pGM* in terms of the relative differences between *wGM* and *GM* and *pwGM* and *pGM*, respectively, as follows:

$$eff_{GM} = \frac{wGM - GM}{GM} \qquad eff_{pGM} = \frac{pwGM - pGM}{pGM} \tag{5}$$

Figure 2 shows the differences in geometric mean and penalized geometric mean values due to the introduction of the weights. Figure 2 shows that penalization amplifies the effects of introducing the weights. In fact, only for two countries (i.e., Slovenia and France) the relative difference eff_{pGM} is smaller than eff_{GM}. This is not unexpected, since in (4) we can see that the weights doubly affect the values of $pwGM_i^{\pm}$, once through the term wGM_i and again through the term $\exp\{\pm wS_{0,i}^2\}$.

Table 1 Values and corresponding rankings

Country	GM	rank GM	wGM	rank wGM	pGM	rank pGM	pwGM	rank pwGM
Algeria	0.3138	15	0.2407	15	0.0389	14	0.1088	14
Croatia	0.5584	5	0.5355	5	0.1857	5	0.2058	5
Cyprus	0.4075	10	0.3547	11	0.0820	10	0.1454	10
Egypt	0.3170	14	0.4358	8	0.0384	15	0.1079	15
France	0.7332	2	0.7481	2	0.3401	2	0.2776	2
Greece	0.4781	8	0.4149	10	0.1256	8	0.1739	8
Israel	0.5924	4	0.6049	4	0.2136	4	0.2188	4
Italy	0.6425	3	0.6103	3	0.2587	3	0.2406	3
Jordan	0.2527	17	0.2256	16	0.0185	17	0.0853	17
Lebanon	0.3657	13	0.3433	12	0.0626	13	0.1282	13
Malta	0.4213	9	0.4851	6	0.0923	9	0.1487	9
Morocco	0.2647	16	0.1511	17	0.0215	16	0.0905	16
Portugal	0.5165	6	0.4477	7	0.1530	6	0.1897	6
Slovenia	0.7801	1	0.7540	1	0.3862	1	0.2987	1
Spain	0.5060	7	0.4223	9	0.1472	7	0.1855	7
Tunisia	0.3690	12	0.2928	14	0.0627	12	0.1303	12
Turkey	0.3692	11	0.3010	13	0.0631	11	0.1304	11

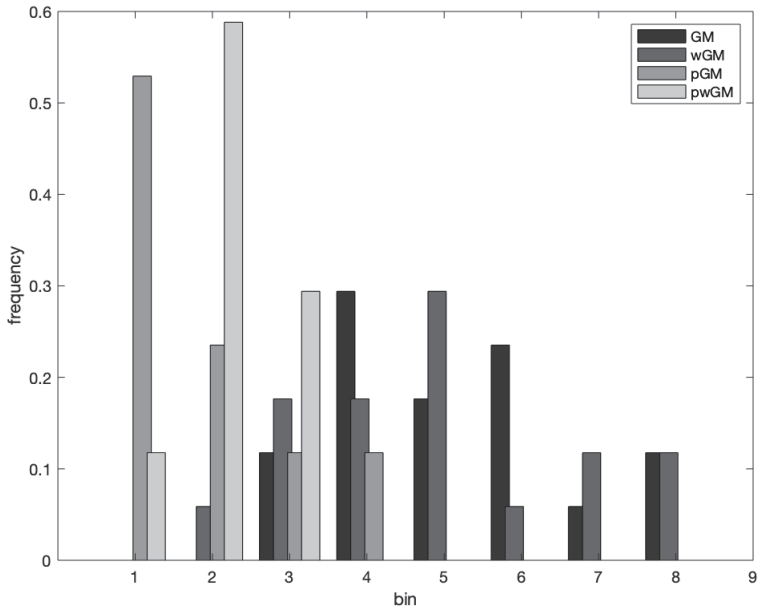


Fig. 1 Distribution comparisons

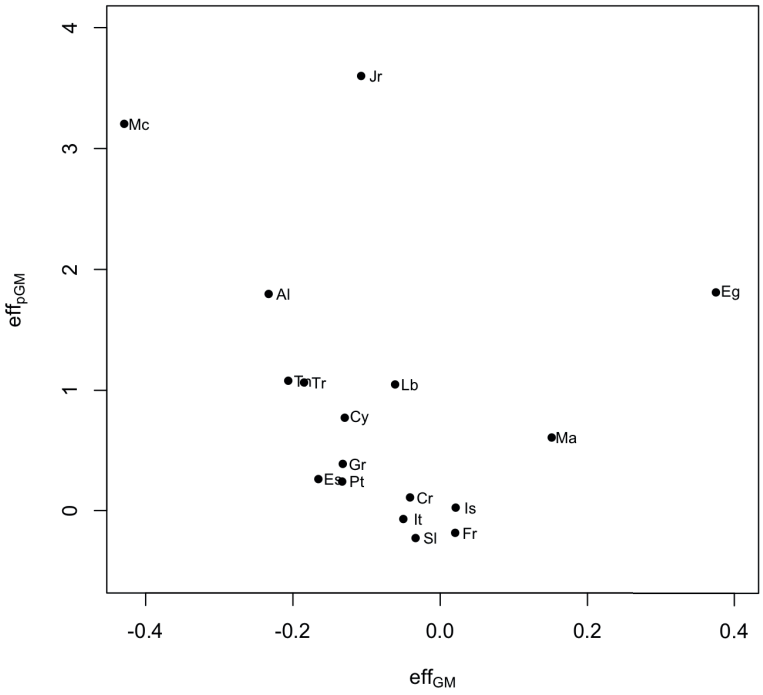


Fig. 2 Comparison of the role of weights

However, it is worth noting that high relative differences between *wpGM* and *pGM* do not correspond to differences in the ranking (see Table 1).

4 Conclusions

The empirical results of different methods applied to the 17 Mediterranean countries according to 11 SDGs indicators related to the agro-food sustainability show that northern Mediterranean countries better perform compared to southern and eastern countries.

The analysis of weights reveals that their introduction in the geometric mean has a great impact with respect to introducing them in the penalized version of the geometric mean. This could be interpreted as a strength for the penalized geometric mean. In fact, the simple geometric mean requires to be weighted to account for inequality, while the penalized method already accounts for inequality across in-

dicators. That is, the ranking deriving from penalised method remains unchanged when weights are included to account for inequality across countries.

References

1. Berger, R.L., Casella, G. (1992). Deriving Generalized Means as Least Squares and Maximum Likelihood Estimates. The American Statistician, 46, 279–282.
2. Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26 (2), 211–252.
3. Casini, M., Bastianoni, S., Gagliardi, F., Gigliotti, M., Riccaboni, A., Betti, G. (2019). Sustainable Development Goals indicators: A methodological proposal for a Multidimensional Fuzzy Index in the Mediterranean area. Sustainability, 11(4), 1198.
4. Chelli, F.M., Ciommi, M., Emili, A., Gigliarano, C., Taralli, S. (2015). Comparing equitable and sustainable well-being (Bes) across the Italian Provinces. a factor analysis-based approach. Rivista Italiana di Economia Demografia e Statistica, 69 (3) 61–72.
5. Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., Chelli, F. M. (2017). A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces. Ecological Indicators, 76, 281–296.
6. De la Cruz, R., Kreft, J. U. (2018). Geometric mean extension for data sets with zeros. arXiv preprint arXiv:1806.06403.
7. Mariani, F., Ciommi, M. (2022). Aggregating composite indicators through the geometric mean: a penalization approach, submitted for publication.
8. Mazziotta, M., Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socioeconomic phenomena. Social indicators research, 127(3), 983–1003.
9. Sachs, J., Schmidt-Traub, G., Pulselli, R.M., Gigliotti, M., Cresti, S., Riccaboni, A. (2019). Sustainable Development Report 2019 - Mediterranean Countries Edition. Siena: Sustainable Development Solutions Network Mediterranean (SDSN Mediterranean).

Evaluating the determinants of innovation from a spatio-temporal perspective. The GWPR approach

Una prospettiva spazio-temporale per lo studio delle determinanti dell'innovazione. L'approccio GWPR

Gaetano Musella, Giorgia Rivieccio, Emma Bruno

Abstract Innovation is one of the main leverages of regional economic development. It has been previously studied through classical methods (e.g., OLS) without considering the potential spatial heterogeneity influence. Local regression methods, such as geographically weighted regression (GWR), might describe the phenomenon more appropriately. The geographically weighted panel regression (GWPR) combines GWR with panel estimation controlling for spatial and individual heterogeneity as a methodological enhancement. This paper compares the estimates of GWPR, GWR and global models using data on 287 NUTS-2 European regions in 2014-2021. The results confirm that GWPR estimations significantly differ from GWR and global models, potentially producing new patterns and findings.

Abstract L'innovazione è una delle principali leve dello sviluppo economico regionale. Gli studi precedenti hanno analizzato il fenomeno utilizzando modelli classici (ad esempio, OLS) senza considerare la potenziale influenza dell'eterogeneità spaziale. Il fenomeno potrebbe essere descritto in modo più appropriato dai metodi di regressione locale, come la geographically weighted regression (GWR). La geographically weighted panel regression (GWPR) rappresenta un avanzamento metodologico combinando la GWR con i modelli panel. Il presente lavoro confronta la GWPR con i modelli classici e con la GWR utilizzando dati su 287 regioni europee nel 2014-2021. L'analisi evidenzia come la GWPR produca risultati significativamente diversi dalla GWR e dai modelli globali.

Keywords: Local regression models, GWR, GWPR, Panel, Innovation

[1] Gaetano Musella, University of Naples Parthenope, gaetano.musella@uniparthenope.it
Giorgia Rivieccio, University of Naples Parthenope, giorgia.rivieccio@uniparthenope.it
Emma Bruno, University of Naples Parthenope, emma.bruno@uniparthenope.it

1 Introduction

During the last years, innovation has claimed the interest of scholars around the world. Ahmad and Zheng (2022) highlighted the leading role played by innovation as an engine driver for economic growth, dynamism, and competitiveness. This interest has led to several European policies aimed to foster the innovation performance of firms and territories. For example, the European Union established in early 2002s the 'Lisbon Strategy' proposing a multitude of guidelines to improve the Member States' economic development. Enhancing the knowledge-based economy, a pillar of good innovation performance, was considered a cornerstone of the EU strategy to make the Union most competitive and dynamic over a decade (European Communities, 2009).

It is not surprising that many researchers have aimed to identify the factors that encourage or hinder companies or territories in developing and adopting innovations. One of the starting points of previous research was investigating the relationship between the output side of innovation – which can be proxied by several variables such as patents or designs – and the more intuitive input side, namely research and development (R&D) expenditure. The R&D has empirically proved its fostering action in different periods and territories (Park, (2005); Kim et al., (2012)). However, Shefer and Frenkel (2005) highlighted that the innovation-R&D relationship is related, albeit with different degrees, to firm size, organisational structure, ownership type, industrial branch, and location. What emerged from their study is that large firms tend to invest more in R&D than the small ones, and the pivotal role of urban areas composition since R&D tends to be concentrated in large urban areas. In other words, there is a spatially varying impact of R&D since it plays a more significant role in creating innovation in central than peripheral areas. Many other drivers of innovation exist, with the empirical and theoretical literature that has ranged its interest from human capital (Rodríguez-Pose and Wilkie, (2019)), to the composition of the workforce (Lopes et al., (2021)), to scientific collaborations (Ganau and Grandinetti, (2021)). A spatially varying relationship with innovation might be present for each of them.

Studies considering the territorial distribution of innovation determinants are still scarce despite many contributions. The expected relationship might differ in different territories since regions' development is uneven, and within the same territory, the time dimension deserves the proper attention. In other words, the relationship between innovation and its drivers presented in the most existing literature is essentially a global estimate, as the relationship applies invariantly over space. Such estimates might be informative at a large spatial scale but might be misleading for regional development programmes. Promoting regional development requires analysing the regional disparities. Studies considering the spatial dimension in the innovation generating process exist. However, they lack an empirical framework to explore the hypothesis that driving factors have a different impact on innovation performance in different territories. For example, Moreno et al. (2005) examined the spatial distribution of innovative activity in European regions. They pointed out the relevance of R&D and agglomeration economies for local development. Ganau and Grandinetti (2021) tested the role of innovation inputs in a

regional heterogeneity perspective. The authors find that public and business R&D expenditure factors do not work unconditionally and everywhere. While the scholars aimed to analyse the spatial heterogeneity of innovation enhancing factors, their work was based on an average relationship estimated through a Probit model.

To overcome this lack in spatial econometrics models, geographically weighted regression (GWR) was proposed (Brundson et al., (1996); Fotheringam et al., (1997)). This local spatial approach allows constructing local models and estimating local regression coefficients. As the main advantage, GWR coefficients vary across the space, allowing to explore spatial heterogeneity explicitly. While GWR is a useful exploratory technique for studying phenomena where spatial non-stationarity is suspected, it suffers drawbacks, such as potential coefficients' multicollinearity (Bruna and Yu, (2013)). Moreover, in the GWR, local models capture the geographic space information through cross-sectional data, not exploring the possibility that relationships are potentially varying also in temporal space. A first attempt to combine geographic space with temporal space was by Yu (2010), who proposed geographically weighted panel regression (GWPR) by combining GWR with the panel data model. As the main methodological advancement, GWPR allows studying local responses and detecting the presence of specific space-time patterns in the data.

This paper presents GWPR in the context of innovation studies seeking to contribute to the literature in two ways. First, to our best knowledge, this is the first research to examine how the relationships between innovation and its determinants vary locally. Second, we evaluate whether new previously hidden insights in the dataset arise by considering the temporal space in local models. For this purpose, by resorting to an innovation panel data from 2014 to 2021 for European regions (NUTS-2 of Eurostat classification), we compare the GWR results (estimated on 2014 and 2021 data) and GWPR estimations (on the whole period).

This article is structured as follows. In Section 2, we present the local models' framework. Section 3 offers the methodological details, while Section 4 presents the dataset used. In Section 5, the results for different models are compared and analysed. Section 6 concludes.

2 The path of spatio-temporal analysis

The ordinary least square (OLS) regression has always been one of the most useful methods to investigate the relationships among variables. It can, however, easily produce biased or inefficient estimations when the assumptions necessary for its implementation are no longer valid. Specifically, when dealing with spatial data, the dependency between nearby observations could break the assumption of uncorrelated residuals. The spatial proximity influences the relationships between phenomena or objects: observations are related to one another, but closest observations are more related than those further away. Moreover, empirical evidence shows that the assumption of stationarity over space may be unrealistic since non-stationarity often concerns spatial data (Fotheringham et al., (1997); Leung et al., (2000)). So, the occurrence of spatial non-stationarity, i.e., the influence of explanatory variables on the dependent variable varies with the location of the

observations, needs modelling strategies that take it into account (Fotheringham et al., (2003)).

Geographically Weighted Regression (GWR) is a local exploratory technique investigating heterogeneity in data relationships across space. It suits situations when the global (stationary) model does not properly describe spatial relationships and a localised fit is needed. The model, pioneered by Brunsdon et al. (1996), extends the OLS regression framework by allowing local rather than global parameters to be estimated for each relationship in the model. By repeating the estimation procedure at each point in space, GWR estimates as many coefficients as local areas, thereby better reflecting the spatially varying relationships between dependent and explanatory variables.

Yu (2010) took another step forward in exploring spatial heterogeneity by combining GWR and panel data analysis. Geographically Weighted Panel Regression (GWRP) involves the time dimension in the GWR model assessing the time series of observations at a specific area as a realisation of a smooth spatio-temporal process (Bruna and Yu, (2013)). Such a spatiotemporal process is based on the idea that closer observations, either in space or time, are more related than distant ones. This approach addresses two issues: *i)* it takes the spatial structure of the data and non-stationary variables into account, extending the classical linear regression to local spatial models providing specific parameters for each local area; *ii)* it also considers the time dimension, allowing for more accurate results than the pooled models. The enlarged sample size gives more degrees of freedom and reduces the collinearity among explanatory variables, thus improving the efficiency of econometric estimates (Wooldridge, (2002)).

3 Methodology

This paper investigates the determinants of innovation and the spatial non-stationarity of relationships across European regions. Following the procedure suggested by Yu (2010), we perform the analysis by using the GWPR. A fixed or random effects model can be applied to obtain the spatially varying parameters. Since we resorted to the fixed effects model, we present this specification. For a set of locations indexed by i = 1, 2, ..., N observed throughout the study period t = 1, 2, ..., T, the GWPR with fixed effects can be written as (Yu, (2010)):

$$y_{it} = \beta_0(u_{it}, v_{it}) + \sum_{k=1}^{p} \beta_k(u_{it}, v_{it}) x_{itk} + \varepsilon_{it}; \quad i = 1,2,\dots,N; t = 1,2,\dots,T \qquad (1)$$

where u_{it}, v_{it} are the geographical coordinates for the *i-th* location at time t; y_{it}, x_{itk}, and ε_{it} are, respectively, the dependent variable, the *k-th* explanatory variable, and the error term at the *i-th* location; p is the number of explanatory variables. $\beta_k(u_{it}, v_{it})$ is the coefficient of the *k-th* variable for the *i-th* unit, while $\beta_0(u_{it}, v_{it})$ is the intercept that denotes the time-invariant fixed effects. The Weighted Least Squares approach estimates the parameters in the GWPR model. Based on the assumption that for each regression point (i), closer observations have more influence in estimating parameters than more remote observations, the weight system (W) is defined as a function of the distance. More specifically, W is calculated with the bi-square kernel

function, which assigns the observations a decreasing weight with distance, and this weight is zero above a specific distance (bandwidth) (Bruna and Yu, (2013)):

$$w_{ij} = \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2 \text{ if } d_{ij} < h_i, 0 \; otherwise \tag{2}$$

where d_{ij} is the Euclidean distance between observations at locations *i* and *j*, while h_i is the adaptive bandwidth for the *i-th* location: each unit has its proper bandwidth selected so that the same number of neighbours is considered for all the regression points. The optimum bandwidth is defined by calibrating the GWPR model through the Cross-Validation (CV) criterion, which accounts for model prediction accuracy, defined as follows (Yu, (2010)):

$$CV = \sum_{i=1}^{n} \left(\bar{y}_i - \hat{\bar{y}}_{\neq i}(h_i)\right)^2 \tag{3}$$

where $\bar{y}_i$ is the average over time of the dependent variable at the location *i*, $\hat{\bar{y}}_{\neq i}(h_i)$ is the fitted value of y_i with bandwidth h_i when calibrating the model with all the observations except y_i.

3 Data

The GWPR and GWR models are estimated on official data covering 2014-2021. The units of analysis are 287 regions of Europe. We have excluded the regions presenting missing data from the analysis. The European regions (NUTS-2 of Eurostat classification) as the units of analysis represent the finest territorial level for data availability. The regional data are drawn from the 2021 edition of the Regional Innovation Scoreboard (RIS) by the European Commission (Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs).

Moreover, The European Commission proposes the Regional Innovation Index (RII). The RII is a composite indicator calculated as the unweighted average of the scores of RIS variables. It combines the output side of innovation (e.g., the number of patent applications per billion GDP) and input variables (e.g., the R&D expenditure). Since the RII is a mixture of innovation's input and output side, it is not suitable for regression analysis (Edquist et al., (2018)). In this light, we split the RII's information into a composite indicator (the dependent variable) to capture the innovation capabilities of European regions and into a set of innovation drivers used as regressors. Notably, all RIS variables are normalised, ranging from 0 to 1.

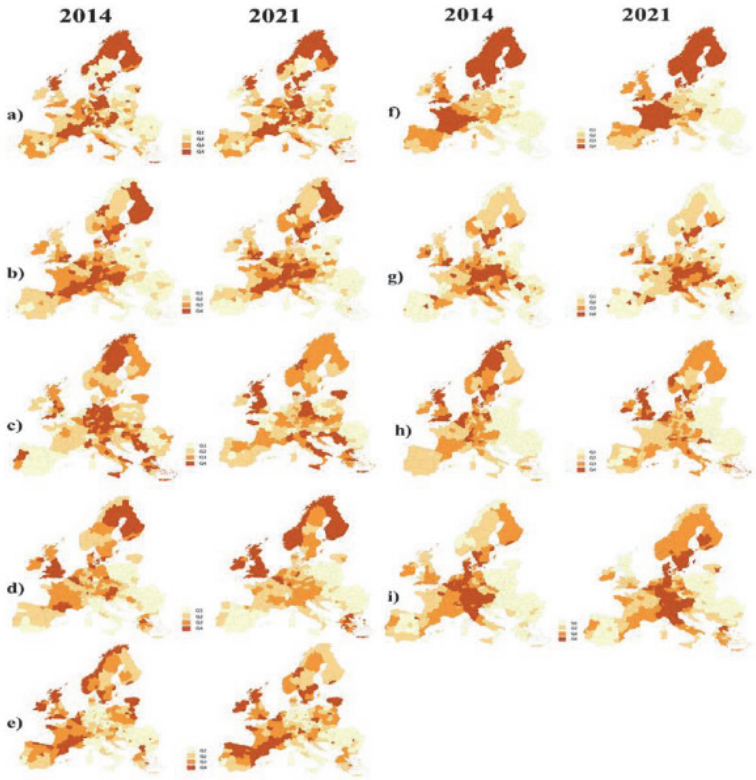
Based on the above, the dependent variable is a composite indicator obtained as the average of five elementary variables (Hollanders et al., (2019)). The elementary variables are listed in **Table 1** (section 'Innovation Output'). The patent, trademark, and design variables measure the final or intermediate step of the innovation process due to large firms and/or service sectors (Edquist et al., (2018)). The SMEs' innovation and Sales of new-to-market and new-to-firm innovations variables capture the innovation due to small and medium firms (Edquist et al., (2018)). As well as the elementary variables, the dependent variable is normalised, and it ranges 0-1. We have controlled for a set of explanatory variables as suggested by the

innovation-related empirical literature. The explanatory variables are listed in **Table 1** (section 'Innovation Input'). Finally, **Figure 1** shows the territorial distribution of variables.

Table 1: Definition of variables

Variable	*Definition*	*References*
Innovation Output		
Patent Applications	Number of patents applied for at the EPO (by year of filing and inventor's address) per billion regional GDP in PPS	Braunerhjelm et al., (2020))
Trademark Applications	Number of trademarks applied for at the EUIPO per billion regional GDP in PPS	Ganau and Grandinetti, (2021)
Design Applications	Number of designs applied for at the EUIPO per billion regional GDP in PPS	Hollanders et al., (2019)
SMEs' innovation	Number of SMEs introducing a product, process, marketing or organisational innovation as a percentage of total SMEs	Lopes et al., (2021)
Sales of new-to-market and new-to-firm innovations	Sum of the total turnover of new or significantly improved products for SMEs as a percentage of SMEs' total turnover	Hollanders et al., (2019)
Innovation Input		
Public R&D	Public expenditure dedicated to developing technological innovations and new products as a share of GDP	Moreno et al., (2005)
Business R&D	Expenditure in the business sector dedicated to developing technological innovations and new products as a share of GDP	Moreno et al., (2005)
Non-R&D innov. expenditure	Total innovation expenditure for SMEs as a percentage of SMEs' total turnover (excluding intramural and extramural R&D expenditures)	Hollanders et al. (2019),
SME collab. innov.	Number of SMEs with innovation co-operation activities (co-operation agreements on innovation activities with other enterprises or institutions) as a percentage of total SMEs	Lopes et al., (2021)
Education	Persons aged 30–34 years with some form of post-secondary education as a percentage of the total population aged 30–34 years	Rodríguez-Pose and Wilkie, (2019)
Lifelong learning	Persons in private households aged 25–64 years who have participated in the four weeks preceding the interview in any education or training as a percentage of the total population aged 25–64 years	Ganau and Grandinetti, (2021)
Employment knowledge	Employed persons in knowledge-intensive services sectors as a percentage of the total workforce	Hollanders et al. (2019),
Scientific research	Number of scientific publications among the top-10% most cited publications worldwide as a percentage of total scientific publications in the region	Ganau and Grandinetti, (2021)

Figure 1: Quantile maps of variables, 2014 and 2021



Note: *a)* Public R&D; *b)* Business R&D; *c)* Non-R&D innovation expenditure; *d)* SME collaborating for innovation; *e)* Education; *f)* Lifelong learning; *g)* Employment knowledge; *h)* Scientific research; *i)* Innovation output

4 Empirical results

The paper focuses on the GWR extension to panel data and its differences with in-average models and cross-sectional GWR. To emphasise the differences between global regressions (cross-section and panel) and local regressions, we present the results of several models, namely cross-section in 2014 and 2021, panel data with fixed effects, GWR in 2014 and 2021, and GWPR with fixed effects in 2014-2021.

Table 2 shows the global models' estimations. Regarding cross-sectional estimates, a relatively higher innovation outcome is mainly associated with a higher endowment of business R&D expenditure, non-R&D expenditure for innovation, scientific research, and employee in knowledge-related sectors. In particular, the results confirm the pivotal role of investment in research and development. On the one side, the business R&D might be related to large firms' activities leading their innovation activities (Moreno et al., (2005)); on the other side, non-R&D

investments – such as the acquisition of machinery, market research, or feasibility studies – are suitable in explaining innovation in smaller entrepreneurship where in-house R&D activities are lacking (Thomä and Zimmermann, (2020); Baumol, (2005)). Notably, public R&D is statistically significant only in the 2021 model. Scientific research is another main innovation driving factor. According to De Rassenfosse and de la Potterie (2009), an explanation might be that academic contributions could incorporate market-oriented initiatives overcoming the boundaries of classic scientific research. More surprising are the results of the education variable since the coefficients show a negative impact on innovation. Although the result might sound strange, other evidence exists on the negative effects of human capital on innovation. For example, Roper and Hewitt-Dundas (2015) found this relationship relatively to process innovation activities. Ganau and Grandinetti (2021) used a composite indicator (similar to that used in this analysis) to measure the innovation activities finding a negative value for the human capital's coefficient.

Regarding the panel data global model, we resort to a fixed-effects model following the result of the Hausmann test (see **Table 2**). Some interesting insights emerge since the estimation differs from the cross-sectional ones. First, only business R&D and scientific research remain statistically significant. The relevant role of the collaboration between SMEs and lifelong training programs emerges from introducing time dimensions. In particular, SMEs can use collaborative agreements to share know-how and exploit opportunities by interacting with similar agents (Hervás-Oliver et al., (2021)). However, knowledge sharing is time-consuming; this could explain why this variable becomes significant in the panel model. Similarly, lifelong learning programs need time to recalibrate and reskill the workforce to provide the technical competence and mastery of analytic tools that could stimulate creative thinking and facilitate its utilisation (Baumol, (2005)).

Table 2: Global regression and Monte Carlo test (2014; 2021), global panel regression (2014-2021)

	2014		***2021***		***Fixed effects***
Variable	***Coeff.***	***Monte Carlo test***	***Coeff.***	***Monte Carlo test***	***Coeff.***
Intercept	0.162*** (0.023)	0.00	.259*** (0.028)	0.00	0.286*** (0.021)
Public R&D	-0.003 (0.028)	0.38	0.187*** (0.034)	0.00	0.012 (0.020)
Business R&D	0.183*** (0.029)	0.47	0.253*** (0.039)	0.90	0.041* (0.021)
Non-R&D innov. expenditure	0.121*** (0.037)	0.97	0.079* (0.046)	0.31	0.010 (0.009)
SME collab. innov.	0.001 (0.029)	0.00	0.044 (0.037)	0.00	0.184*** (0.008)
Education	-0.091*** (0.028)	0.00	-0.168*** (0.033)	0.03	0.029 (0.018)
Lifelong learning	-0.004 (0.030)	0.00	0.052 (0.034)	0.00	0.063** (0.029)
Employment knowledge	0.187*** (0.029)	0.10	0.069* (0.038)	0.02	0.027 (0.017)

Scientific research	0.301*** (0.029)	0.00	0.087** (0.044)	0.00	0.054*** (0.013)
R^2 Adjusted	0.701		0.528		0.121
N	287		287		2,296
Breusch-Pagan LM test	-		-		4,348.8 (*p-value*:0.00)
Hausman test	-		-		145.2 (*p-value*:0.00)

Note: ***; **; *: Significance level at 1 %, 5 %, 10 %. Standard errors in brackets. Values for Monte Carlo test columns are p-values.

To explore the coefficients' spatial heterogeneity, we estimate GWR (for 2014 and 2021) and GWPR with fixed effects models. As a first step, we define the optimal kernel bandwidth by minimising the cross-validation (CV) criterion. The procedure suggests using the adaptative bi-square kernel with 93 nearest neighbours[1]. Once the optimal kernel bandwidth is defined, we test the spatial non-stationarity of parameters through the Monte Carlo significance test[2]. The results of the Monte Carlo test (**Table 2**) show that the associations between innovation and its determinants are deemed mostly non-stationary in European regions. Notably, exceptions exist. In particular, for 2014, the coefficients of the following variables are stationary: public and business R&D, non-R&D innovation expenditure, and employment in knowledge sectors. In 2021 the scenario changed significantly since only Business R&D and non-R&D innovation expenditure failed the non-stationary test. On the one hand, this emphasises the need for local fitting techniques to improve estimates' accuracy and provide more suitable analysis; on the other hand, a remarkable change in regional innovation determinants over time emerges. On this basis, it is clear how conducting a cross-sectional study would lead to a partial representation of the driving forces of innovation in European regions. Finally, we perform the Hausmann local tests to evaluate which panel estimation is more appropriate (random vs fixed effects) for GWPR. The results favour GWPR with fixed effects since we reject the null hypothesis in 245 out of 287 regions.

Figure 2(a-h) shows quantile maps of local cross-sectional coefficients and local fixed effects panel estimates. The coefficients not statistically significant are shadowed. **Figure 2(i)** shows the local adjusted R^2. Some interesting observations emerge. First, comparing GWPR and cross-sectional GWR models appears a general

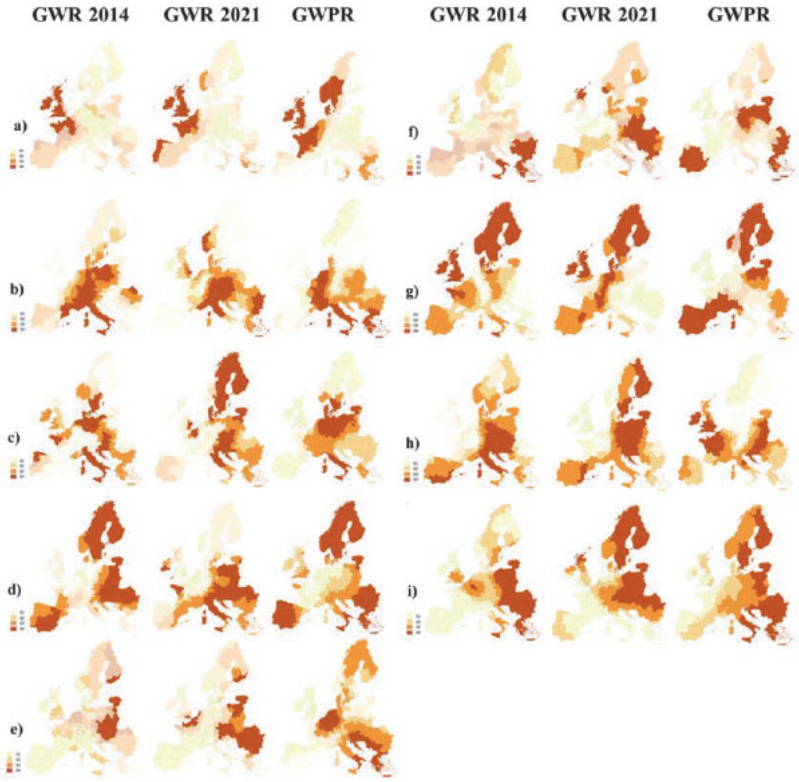
[1] Notably, for the three models (GWR 2014 and 2021, and GWPR) the optimal bandwidth procedure converges towards adaptative bi-square kernel but it highlights three different nearest neighbours: 85 (GWR 2014), 62 (GWR 2021), and 93 (GWPR). This is not surprising since CV procedure is based on the value of dependent and independent variables. We adopt the larger bandwidth for sake of comparability between models. However, the estimations with different adaptative bi-square kernels show very similar patterns (respect to those reported in the paper). We do not report here for conciseness but are available upon request.

[2] We estimate the GWR and GWPR models through R software. Unfortunately, the Monte Carlo test has not implemented in GWPR routine yet. For this test, we only refer to GWR. The spatial variability of GWPR local parameters can be evaluated only through the F test (at least one coefficient is spatially varying) and the local t tests.

change in coefficients' quantile distribution and statistical significance. For example, the public R&D is the only investment-related variable spatially varying (just in 2021), highlighting that regional-specific relationships do not exist with innovation activities. This consideration seems to change in the panel analysis since clear clusters of regions emerge. The regions of northern Europe (almost all of the UK and Ireland, many areas of France, Belgium, the Netherlands, Sweden and Norway) are characterised by a high impact of public R&D on innovation. The same occurs for Grecian regions. In east Europe and some Italian regions, the relationship is very weak. In all other regions, there is no effect. This result is in contrast with previous works that pointed out the leading role of public R&D not only in average-based studies on the whole sample but also in research based on a regional split of European territory (Ganau and Grandinetti, (2021); Lopes et al., (2021)). This might be because the previous empirical analyses were conducted through average estimation methods within the sub-sample identified.

Local regressions show even more noticeable improvement in estimates for collaborating SMEs for the innovation variable. While the coefficients are not significant in the global models, the local regression analyses prove the pivotal role of the SMEs' collaborating activities in enhancing the innovation performance of some regions. However, the full impact of collaboration emerges only in the GWPR model since the spatio-temporal patterns suggest the existence of relevant information hidden in local cross-sectional estimations. First, the GWPR leads to a considerable improvement in coefficients' significativity with respect to GWR. Second, GWPR highlights how it is a crucial driver in Mediterranean countries, east Europe, and the Scandinavian peninsula. This pattern does not arise in the GWR models (for example, the estimates fail to capture the role of the variable in Italy and Greece (2014) and Spain (2021)). However, this shall not come as a surprise considering that the flow of knowledge between enterprises requires time, and this feature is rather obscured in local cross-sectional analysis. Moreover, regional specific characteristics emerge. For example, the Scandinavian and Greek regions feature a significantly higher SME collaboration performance than the whole EU, i.e., their regions dominate the list of the top 40 European best-performing regions (Hollanders et al., (2019)). Finally, the local estimations significantly improve the goodness of fit, especially in the GWPR case. Indeed, in GWPR, the values of local $R^2_{adjusted}$ ranging 0.007-0.461 (average=0.181; median=0.164; third quartile= 0.272), increasing respect to the 0.121 of the global model.

Figure 2: Coefficients generated with GWR (2014 and 2021) and GWPR by quantiles.



Note: *a)* Public R&D; *b)* Business R&D; *c)* Non-R&D innovation expenditure; *d)* SME collaborating for innovation; *e)* Education; *f)* Lifelong learning; *g)* Employment knowledge; *h)* Scientific research; *i)* Local $R^2_{adjusted}$. The coefficients not statistically significant are shadowed.

5 Conclusions

This work presents the GWPR method as a procedure able to fill the gap between GWR literature and panel data literature. The main originality of GWPR is that it allows studying potential spatial heterogeneity in models controlling for individual heterogeneity. We compared the GWPR with global regressions (2014, 2021, and 2014-2021) and cross-sectional GWR (2014, and 2021). Some interesting results emerge. First, the local estimations accurately describe the relationship between innovation and its determinants regarding the global average models. Second, the local estimates are somewhat different when introducing the time dimension. Third, GWPR leads to an improvement in coefficients' statistical significance.

From an empirical point of view, future research developments might include the introduction of other potentially relevant regressors and finer spatial data (e.g., provincial level). Moreover, introducing a new option in the software routine may also allow evaluating the spatial variability in GWPR (i.e., Monte Carlo simulation) and the local multicollinearity (i.e., local VIF).

References

1. Ahmad, M., Zheng, J.: The Cyclical and Nonlinear Impact of R&D and Innovation Activities on Economic Growth in OECD Economies: a New Perspective. Journal of the Knowledge Economy, 1-50. (2022).
2. Baumol, W. J.: Education for innovation: Entrepreneurial breakthroughs versus corporate incremental improvements. Innovation policy and the economy 5: 33-56 (2005).
3. Braunerhjelm, P., Ding, D., Thulin, P.: Labour market mobility, knowledge diffusion and innovation. European Economic Review 123, 103386. (2020).
4. Bruna, F., and Yu, D.: Geographically weighted panel regression. XI Congreso Galego de Estatística e Investigación de Operacións. http://xisgapeio. udc. es. (2013).
5. Brunsdon, C., Fotheringham, A. S., Charlton, M. E.: Geographically weighted regression: a method for exploring spatial non-stationarity. Geographical analysis, 28(4), 281-298. (1996).
6. Commission of the European Communities: Communication From the Commission to the Council, the European Parliament, the European Economic and Social Committee, and the Committee of the Regions: A Mid-Term Assessment of Implementing the EC Biodiversity Action Plan. Journal of International Wildlife Law & Policy, 12(1-2), 108-120. (2009).
7. De Rassenfosse, G., de la Potterie, B. V. P.: A policy insight into the R&D–patent relationship. Research Policy 38.5: 779-792 (2009).
8. Edquist, C., Zabala-Iturriagagoitia, J. M., Barbero, J., Zofío, J. L.: On the meaning of innovation performance: Is the synthetic indicator of the Innovation Union Scoreboard flawed?. Research Evaluation, 27(3), 196-211. (2018).
9. Fotheringham, A. S., Brunsdon, C., Charlton, M.: Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons. (2003).
10. Fotheringham, A. S., Charlton, M.E., Brunsdon, C.: Measuring spatial variations in relationships with geographically weighted regression. Recent developments in spatial analysis. Springer, Berlin, Heidelberg, 60-82 (1997).
11. Ganau, R., Grandinetti, R.: Disentangling regional innovation capability: what really matters?. Industry and Innovation, 28(6), 749-772. (2021).
12. Hervás-Oliver, J. L., Parrilli, M. D., Rodríguez-Pose, A., Sempere-Ripoll, F.: The drivers of SME innovation in the regions of the EU. Research Policy 50.9: 104316 (2021)
13. Hollanders, H., Es-Sadki, N., Merkelbach, I.: Regional innovation scoreboard 2019. (2019).
14. Kim, Y. K., Lee, K., Park, W. G., Choo, K.: Appropriate intellectual property protection and economic growth in countries at different levels of development. Research policy, 41(2), 358-375. (2012).
15. Leung, Y., Mei, C. L., Zhang, W. X.: Statistical tests for spatial non-stationarity based on the geographically weighted regression model." Environment and Planning A 32.1, 9-32 (2000).
16. Lopes, J. M., Silveira, P., Farinha, L., Oliveira, M., Oliveira, J.: Analyzing the root of regional innovation performance in the European territory. International Journal of Innovation Science. (2021).
17. Moreno, R., Paci, R., Usai, S.: Spatial spillovers and innovation activity in European regions. Environment and planning A, 37(10), 1793-1812. (2005).
18. Park, W. G.: Do intellectual property rights stimulate R&D and productivity growth? Evidence from cross-national and manufacturing industries data. Intellectual Property and Innovation in the Knowledge-Based Economy, Industry Canada, Ottawa, 9, 1-9. (2005).
19. Rodríguez-Pose, A., Wilkie, C.: Innovating in less developed regions: What drives patenting in the lagging regions of Europe and North America. Growth and Change, 50(1), 4-37. (2019).
20. Roper, S., Hewitt-Dundas, N.: Knowledge stocks, knowledge flows and innovation: Evidence from matched patents and innovation panel data. Research Policy, 44(7), 1327-1340. (2015).
21. Shefer, D., Frenkel, A.: R&D, firm size and innovation: an empirical analysis. Technovation, 25(1), 25-32. (2005).
22. Thomä, J., Zimmermann, V.: Interactive learning—The key to innovation in non-R&D-intensive SMEs? A cluster analysis approach. Journal of Small Business Management 58.4 :747-776 (2020)
23. Wooldridge, J. M.: Econometric analysis of cross section and panel data MIT press. Cambridge, MA 108.2, 245-254 (2002).
24. Yu, D.: Exploring spatiotemporally varying regressed relationships: the geographically weighted panel regression analysis. The international archives of the photogrammetry, remote sensing and spatial information sciences, 38(Part II), 134-139. (2010).

Dimension Reduction for complex data

Discrimination and clustering via principal components

Discriminazione e clustering tramite componenti principali

N. Trendafilov and V. Simonacci

Abstract In many modern data, the number of variables is much higher than the number of observations and the within-group scatter matrix is singular. This work proposes a way to circumvent this problem by doing LDA in a low-dimensional space formed by the first few principal components (PCs) of the original data. Two approaches are considered to improve their discrimination abilities in this low-dimensional space. Specifically, the original PCs are rotated to maximize the LDA criterion, or penalized PCs are produced to achieve simultaneous dimension reduction and maximization of the LDA criterion. Both approaches are illustrated and compared on a well known data set. In addition, these procedures are extended to clustering.

Abstract *In molti dataset moderni il numero di variabili è molto più alto del numero di osservazioni e la matrice di dispersione entro i gruppi è singolare. Questo lavoro propone un modo per aggirare questo problema svolgendo un LDA in uno spazio a dimensione ridotta formato soltanto da poche componenti principali (PC) dei dati originali. In particolare, si propongono due approcci per migliorare la capacità di discriminazione nello spazio a dimensioni ridotte. Una prima opzione si basa sulla rotazione delle PC originali per massimizzare il criterio LDA. Un altro modo consiste nel produrre PC penalizzate che contemporaneamente ottengono una riduzione dimensionalee massimizzano il criterio LDA. Entrambi gli approcci sono illustrati e confrontati usando un noto dataset. Inoltre, queste procedure sono estese al clustering.*

Key words: Dimension reduction, orthogonal rotations, penalized PCA.

N. Trendafilov
University of Naples "L'Orientale", Italy, e-mail: ntrendafilov@unior.it

V. Simonacci
University of Naples Federico II, Italy, e-mail: violetta.simonacci@unina.it

1 Introduction

Many modern data $X \in \mathbb{R}^{n\times p}$ have much more variables than observations, $p \gg n$. Then, the Fisher's linear discriminant analysis (LDA) cannot be applied because the within-group scatter matrix S_W is singular. There exists a great number of approaches to circumvent this problem [3, Ch 7.4]. This work proposes to do LDA in a low-dimensional space formed by the first few principal components (PCs) of X. Two approaches are considered to improve their discrimination abilities in this low-dimensional space. Specifically, the original PCs are rotated to maximize the LDA criterion, or penalized PCs are produced to achieve simultaneous dimension reduction and maximization of the LDA criterion. Both approaches are illustrated and compared on a well known data set. It is shown how they can be extended to clustering.

2 Revisiting PCA

Let X be an $n \times p$ data matrix which columns are centered, i.e. $1_n^\top X = 0_p^\top$, and have unit lengths, i.e. $\mathrm{diag}(X^\top X) = I_p$. For short, such X is called whitened. Principal component analysis (PCA) of X is performed by its singular value decomposition (SVD). Assuming that the rank of X is $r \leq \min\{n, p\}$, the SVD can take the form $X = FDA^\top$, where $F^\top F = A^\top A = I_r$ and $D \in \mathbb{R}^{r\times r}$ is a positive definite diagonal matrix which diagonal entries are arranged in decreasing order. Note that $1_n^\top F = 0_r^\top$.

In PCA applications, we are looking for some $s \leq r(\leq \min\{n, p\})$ and replace the original data matrix X by its truncated SVD of the form $X_s = F_s D_s A_s^\top$, where F_s and A_s denote the first s columns of F and A respectively, and D_s is a diagonal matrix with the first (largest) s singular values of X. The matrices F_s and A_s contain the component scores and loadings respectively, and D_s^2 contains the variances of the first s PCs. Usually, we are interested in considerable dimension reduction, i.e. $s \ll r$ and in many cases we even set $s = 2$.

The PCA interpretation is based on the component loadings A_s, which reveal the importance of the original variables. The component scores F_s are used to visualize the n observations into a s-dimensional space. It is worth noting that for any orthogonal matrix $Q \in \mathbb{R}^{s\times s}$ we have:

$$X_s = F_s QQ^\top D_s A_s^\top = (F_s Q)(A_s D_s Q)^\top = F_s D_s QQ^\top A_s^\top = (F_s D_s Q)(A_s Q)^\top . \quad (1)$$

Now, right multiplication of X by (the projector) $F_s F_s^\top$ shows that PCA can be expressed as a least-squares (LS) projection of the data onto the s-dimensional subspace in $\mathbb{R}^n$ spanned by the columns of F_s, i.e. by the component scores. Then, PCA of X can be rewritten as

$$\min_{F^\top F = I_s} \|X - FF^\top X\|_E , \quad (2)$$

where $\|A\|_E^2 = \text{trace}(A^\top A)$ is the Euclidean (Frobenius) norm of $A \in \mathbb{R}^{n\times p}$. Thus, $X_s = F_s F_s^\top X$ is the best rank s (LS) approximation to X in $\mathbb{R}^n$.

If, in addition, the observations of X are divided into g groups, then the component scores F_s visualize also the g groups in a s-dimensional space. However, such a projection does not take into account the group structure of the data and thus, is not optimal. Linear discriminant analysis (LDA) overcomes this weakness by finding a low-dimensional space where the groups are best separated.

3 LDA of F_s

As above, let X be an $n \times p$ whitened data matrix with observations divided into g groups and membership defined by an $n \times g$ indicator matrix G with $\{0,1\}$ elements, such that the matrix of group means is given by $\bar{X} = (G^\top G)^{-1} G^\top X$. Then, we have:

$$S_W = S_T - S_B = X^\top X - \bar{X}^\top (G^\top G)\bar{X} = X^\top X - X^\top H X \ , \tag{3}$$

where $H = G(G^\top G)^{-1} G^\top$ and S_B and S_W are the between- and within-groups scatter matrices of X [3, Ch 7.1]. The purpose of the Fisher's LDA is to find a transformation matrix $A \in \mathbb{R}^{p\times s}$, such that the *a-priori* groups are better separated in the transformed data $Y = XA$ than with respect to any of the original variables [1]. This is achieved by solving the following generalized eigenvalue problem

$$S_B A = S_W A \Lambda \ , \tag{4}$$

where Λ is the $s \times s$ diagonal matrix of the s largest eigenvalues of $S_W^{-1} S_B$ ordered in decreasing order. This is possible if S_W^{-1} exists. We stress that there are *at most* $\min\{p, g-1\}$ non-zero eigenvalues in Λ, which is the rank of S_B.

3.1 LDA with rotated component scores

Our purpose is to improve the discrimination features of the component scores. We assume that PCA is already performed and F_s is available, keeping in mind that $s \le \min\{p, g-1\}$. For short we simply write F. As we want to do LDA on F, our data matrix X in (3) becomes $F \in \mathbb{R}^{n\times s}$, which is a centered orthonormal matrix. This simplifies (3) to:

$$S_W = I_s - F^\top H F = F^\top (I_n - H) F \ . \tag{5}$$

Then, the Fisher's LDA of F requires the solution of:

$$F^\top H F Q = Q \Lambda \ , \tag{6}$$

which is a symmetric eigenvalue problem for $F^\top HF$ or a singular value problem for HF. Equivalently, we can express (6) as the optimization problem:

$$\max_{Q^\top Q=I_s} \|HFQ\|_E \ , \tag{7}$$

meaning that LDA of F, in fact, finds an orthogonal rotation Q, such that the rotated component scores FQ maximize trace(S_B), the between-group sum of squares.

The problem with this solution is that it "adjusts" the group memberships of the objects with respect to group means obtained by the initial PCA of X. It will be seen, that this procedure is more suitable for clustering, when both the group memberships and the centroids are adjusted at every step.

3.2 LDA with penalized component scores

In the previous Section 3.1 the dimension reduction and the discrimination are performed one after another. Here, we want to put them together into a single procedure. For this reason, we enhance PCA with discriminatory features by considering a joint minimization of the PCA objective function (2) and maximization of trace(S_B), the between-groups sum of squares of the component scores F. This results in the following problem:

$$\min_{F^\top F=I_s} \|(I_n - FF^\top)X\|_E - \|HF\|_E \ , \tag{8}$$

which is equivalent to:

$$\max_{F^\top F=I_s} \text{trace} F^\top (XX^\top + H)F \ , \tag{9}$$

and is solved as truncated EVD of $XX^\top + H$.

3.3 Example: Fisher's Iris data

The famous Fisher's Iris data contains observation of three $(g = 3)$ Iris species on $n = 150$ flowers by measuring $p = 4$ variables. It is well known that one of the species is very well separated, but the other two are very close and difficult to split. The number of misclassified flowers by the classical LDA solution is 5 (3.33%), while for PCA they are 27 (18%).

Figure 1 depicts the projections of the flowers and the three groups on the rotated and on the penalized PCs. As expected, the rotated PCs (left) do not make considerable improvement: the number of misclassified observations is reduced by two to 25 (16.67%). However, the penalized PCs (right) achieve quite clear separation of the groups and 12 (8%) misclassified observations.

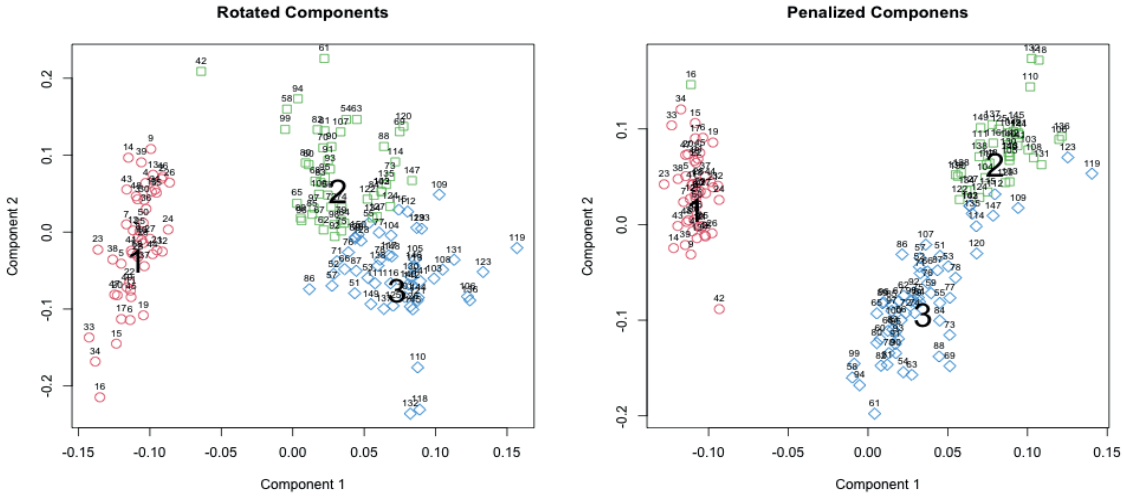


Fig. 1 Iris data: plots of the flowers on rotated (left) and penalized PCs.

4 Clustering of component scores

The procedures outlined in Section 3.1 and Section 3.2 can be readily adapted to perform sequential and simultaneous dimension reduction and clustering.

In clustering problems the labels of the data points are unknown. We circumvent this obstacle by combining K-means-like updates of cluster centroids and membership, with updates of the component scores F at each iteration. The whole procedure is summarized in Algorithm 0.1.

4.1 Example: Fisher's Iris data (continued)

First, we apply K-means clustering to the (whitened) Iris data. The `R` function `kmeans` from `library(cluster)` produces 25 (16.67%) misclassified flowers [2].

The clustering of the rotated component scores is depicted in Figure 2. The allocation to a group is measured by Euclidean (left) and Manhattan (right) distances and the number of misclassified flowers are: 29 (19.34%) and 12 (8%) respectively.

The clustering of the penalized component scores produces well separated clusters. The allocation to a cluster is measured by Euclidean and Manhattan distance. The number of misclassified flowers is 28 (18.67%) and 21 (14%) respectively.

Algorithm 0.1 Clustering of component scores.

set number k of clusters
set random indicator matrix G
initial component scores F by SVD of $X = FDA$
centroids $C \leftarrow (G^\top G)^{-1} G^\top F$
$H \leftarrow G(G^\top G)^{-1} G^\top$
$f_0 \leftarrow \|HF^\top\|_E$, $f \leftarrow 0$
while $|f_0 - f| > 10^{-6}$ **do**
  update G by finding closest (in Euclidean, Manhattan, etc sense) scores to centroids
  $C \leftarrow (G^\top G)^{-1} G^\top F$
  $H \leftarrow G(G^\top G)^{-1} G^\top$
  if rotated scores **then**
    $F \leftarrow FQ$ with Q from (7)
  else
    update F from (9)
  end if
  $f \leftarrow \|HF^\top\|_E$
end while

Fig. 2 Iris data: clustering rotated PCs, with Euclidean (left) and Manhattan allocation.

References

1. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**, 179–188 (1936)
2. Pison, P., Struyf, A., Rousseeuw, P. J.: Displaying a clustering with CLUSPLOT. Comput. Statist. Data Anal. **30**, 381–392 (1999)
3. Trendafilov, N., Gallo, M.: Multivariate Data Analysis on Matrix Manifolds (with Manopt), Springer, New York, NY (2021)

Exploratory graph analysis for configural invariance assessment

Analisi esplorativa delle reti psicometriche per la valutazione dell'invarianza configurazionale

Sara Fontanella, Alex Cucco and Nicola Pronello

Abstract Within the framework of graph theory, we discuss an exploratory approach to evaluate the configural invariance of a test. Networks embedding, coupled with the theory of Gaussian graphical models, provides a flexible approach to verify if the latent structure has the same pattern across different groups. Through a simulation study, we demonstrate that the proposed method is able to identify the differences.
Abstract *Ricorrendo alla teoria dei grafi, in questo lavoro viene presentato un approccio esplorativo per valutare l'invarianza configurazionale di un test. In tale contesto, networks embedding, congiuntamente con la teoria dei modelli grafici Gaussiani, rappresenta uno strumento utile per la verifica dell'equivalenza di struttura degli strumenti che vengono utilizzati per confrontare diversi gruppi. Attraverso uno studio di simulazione, viene dimostrato che il metodo proposto riesce ad evidenziare correttamente i diversi aspetti delle strutture latenti e le differenze esistenti.*

Key words: Configural invariance, psychometric networks, Bayesian statistics, sparse modelling, dimensionality reduction

Sara Fontanella
National Heart and Lung Institute, Imperial College London, London, UK e-mail: s.fontanella@imperial.ac.uk

Alex Cucco
National Heart and Lung Institute, Imperial College London, London, UK e-mail: a.cucco20@imperial.ac.uk

Nicola Pronello
Department of Neurosciences, Imaging and Clinical Sciences, University of Chieti-Pescara, Chieti, Italy e-mail: nicola.pronello@studenti.unich.it

1 Introduction

Self-report survey instruments are frequently used to investigate differences between groups of respondents, such as citizens of different nations in cross-country comparative analyses. A main methodological problem with this kind of comparative research is that the measurement instrument may not function invariantly across the groups. Measurement invariance pertains to the extent to which respondents across groups perceive and interpret the content of the survey instrument and can be broadly defined as stable measurement parameters across multiple groups. Reflective latent variable models are the standard models used in measurement theory. In these models, observable indicators that measure a given construct are thought to co-occur because of an underlying latent variable that causes the covariation between the manifest variables. There are three distinct and hierarchically ordered levels of measurement invariance, and each level is defined by the parameters constrained to be equal across groups [13]. Here, we consider configural invariance, or weak factorial invariance, which holds if there is the same number of factors and an invariant factor loading pattern in all the groups. In the last decade, mutualism models have been proposed to study the relations between observed indicators. These models considers that observable variables co-occur not because of latent causes but because they are causally coupled. Under this data generation hypothesis, factors emerge from their constituent causal connections rather than cause them. Mutualism models rely on correlation-based network, also known as psychometric networks [4], where nodes represent variables and edges represent the association between two nodes conditioned on all other nodes. Recent research [12] has shown that latent variable and network models can be mathematically equated, despite their differing representations and hypotheses about the cause of the co-occurrence between observable variables. Golino et al. [9] propose exploratory graph analysis (EGA). Given a latent variable model as the true underlying causal model, indicators in a network model will feature strongly connected clusters for each latent variable. Under this condition, in EGA the correlation matrix of the observable variables is firstly estimated, then the graphical LASSO procedure is used to obtain the sparse inverse covariance matrix, and, finally, a community detection algorithm is applied to find the number of dense subgraphs of the partial correlation network. The number of clusters identified equals the number of latent factors.
In our work, considering a multi-group comparative analysis and measurement instruments consisting of ordered categorical indicators, we propose to use EGA to assess the instrument configural invariance. We assume that if the measurement instrument functions invariantly across the groups, the group specific correlation-based networks will be characterized by a similar structure. To estimate the sparse inverse correlation matrix we adopt a Bayesian approach with sparse inducing priors [5]. Principal component analysis on the space of labelled networks will be exploited to investigate the structure similarity in a simulation study.

2 Network estimation

Currently, the most common model for psychometric networks is the partial correlation network. Following common notation, the graph is then denoted by $\mathscr{G} = (V, E)$ and consists of nodes $V = \{1, \ldots, p\}$ as well as the edge set $E \subset V \times V$ that contains nodes $(y_i; y_j)$ that share a conditional relationship. In contrast, conditionally independent nodes are not included in E. Assuming jointly Gaussian variables $\boldsymbol{y} = (y_1, \ldots, y_p)'$, all marginal dependence information is contained in the covariance matrix $\boldsymbol{\Sigma}$, and all conditional independence information in its inverse, the precision matrix $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$. Partial correlation between variables y_i and y_j, after conditioning on all other variables in $\boldsymbol{y}$, $y_{-(i,j)}$, are obtained by normalising the precision matrix. The two random variables are conditionally independent given the rest if and only if the (i, j)-th entry, of the precision matrix is zero. Therefore estimating the graph for a Gaussian graphical model is equivalent to identifying zeros in the precision matrix. A way to obtain the partial correlation coefficients is by using node-wise regressions [10], where each variable is regressed on all the others

$$y_j = \beta_{j,0} + \sum_{i \neq j} \beta_{j,i} y_i + \varepsilon_j, \quad j = 1, \ldots, p.$$

The partial correlation coefficients are given by $\rho^{j,i} = Cor\left(y_i, y_j | y_{-(i,j)}\right) = \frac{\beta_{j,i}\sigma_{\varepsilon_i}}{\sigma_{\varepsilon_j}} = \frac{\beta_{i,j}\sigma_{\varepsilon_j}}{\sigma_{\varepsilon_i}}$. Partial correlation networks are usually estimated using regularization techniques [see and references therein [3]], which jointly perform parameter estimation and model-selection, leading to a sparse network structure. In our work, we recover the sparse network structure by imposing sparsity-inducing priors on the regression coefficients. More specifically, assuming that some elements of the partial correlation matrix are close to zero while others have larger values, a spike and slab prior can be specified for each regression coefficients. We consider the spike and slab prior defined by a two-component normal mixture model [7]

$$\beta_{j,i} | \zeta_{j,i} \sim (1 - \zeta_{j,i}) \mathscr{N}(0, \tau^2) + \zeta_{j,i} \mathscr{N}(0, g^2 \tau^2)$$

where τ is positive but small, such that $\beta_{j,i}$ is close to zero when $\zeta_{j,i} = 0$, and g is large enough to allow reasonable deviations from zero when $\zeta_{j,i} = 1$. In addition, the prior probability that there is a conditional dependence between y_j and y_i is $P(\zeta_{j,i} = 1) = 1 - P(\zeta_{j,i} = 0) = p_{j,i}$. The classification between zero and non zero coefficients can be based on the posterior probability of inclusion (*ppi*), given by $P(\zeta_{j,i} = 1 | \mathbf{y})$ [6]. Therefore, from the p regression models, a $p \times p$ sparse matrix is obtained that corresponds to the underlying structure of $\mathscr{G} = (V, E)$.

We apply this approach to detect the underlying structure of measurement instruments constituted by Likert-type scales. For ordinal items, we can assume that Gaussian y-variables underlie the observed ordinal measures. For a given subject n, the relation between the response to item j, measured on a C-point rating scale, and the underlying variable is given by the threshold model $x_{n,j} = c \quad if\ \gamma_{j,c-1} \leq y_{n,j} \leq$

$\gamma_{j,c},\ c = 1,\ldots,C$; $\gamma_{j,0} = -\infty, \gamma_{j,C} = \infty$. A uniform prior distribution is chosen for the threshold parameters, truncated to the region $\{\gamma_{j,c} \in \mathscr{R}, \gamma_{j,c-1} \leq \gamma_{j,c} \leq \gamma_{j,c+1}\}$, $c = 1,\ldots,C-1,\ \forall\, j$, to take account of the order constraints. The full conditional of most parameters can be specified in closed form, which allows for a Gibbs sampler, although Metropolis-Hastings steps are required to sample the ordered threshold parameters.

3 Networks embedding for exploratory analysis

To evaluate instrument configural invariance, one can adopt dimensionality reduction techniques to represent each partial correlation graph as a point in a reduced Euclidean space of principal components, which preserve and emphasize the distinctions between the structures in the different groups. Here, we exploit the framework of object-oriented data analysis and adopt a recently proposed method that extend classical principal component to samples of networks [11]. Accordingly, recalling that the partial correlation matrix can be expressed by means of an unweighted network, $\mathscr{G}$, with edges $E_i = \{\beta_{i,j} : \beta_{i,j} \in (0,1)\ i,j \leq p\}$, here we consider the graph $\mathscr{G}$ as a single observation in the sample of networks. Consequently, each of these networks can be represented through their Laplacian matrix $L = l_{i,j}$ defined as:

$$l_{i,j} = \begin{cases} -\beta_{i,j}, & \text{if } j \neq i \\ \sum_{j \neq i} \beta_{i,j}, & \text{if } j = i \end{cases}.$$

Considering the correspondence between a graph $\mathscr{G}$ and its Laplacian L, we can define the space of networks as the space of Laplacians $\mathscr{L}_p$. Since the space $\mathscr{L}_p$ is not an Euclidean space (i.e manifold with corners of dimension $p\frac{p-1}{2}$ [8]), statistical techniques need to be adapted to account for the geometry of that space. As statistical analysis on manifolds can be conducted in a tangent space of the original space, after defining a suitable metric, each original element can be mapped to a tangent space T_v in v. Here, by using the intrinsic metric in $\mathscr{L}_p$ - namely the Frobenious distance: $d(L_1,L_2) = ||L_1 - L_2|| = [\text{trace}(L_1 - L_2)^T (L_1 - L_2)]^{1/2}$) - and by noticing that $\mathscr{L}_p$ has curvature 0, it is straightforward to obtain the coordinates on the tangent plane T_v [11]. In particular, for any elements $L_k \in \mathscr{L}_p$, with $k = 1,...,n$, $\mathbf{v}_k = \text{vech*}(HL_kH^T)$, holds. Here, H is the Helemert matrix of dimension $(p-1) \times p$, and vech* is the half vectorization of a matrix (including the diagonal) with the off-diagonal elements multiplied by $\sqrt{2}$ [11]. Once the network data are projected in a suitable Euclidean space, it is possible to obtain the principal components in the tangent space via eigendecomposition of the empirical covariance matrix $S = \frac{1}{n}\sum_{k=1}^{n} v_k v_k^T$.

4 Simulation results

In order to evaluate the performance of the exploratory approach described above, we performed a simulation study. Exploiting the mathematical equivalence of latent variable and network model formulations, we generated the data under the multidimensional IRT formalism.
We considered 42 items measured by a 4-point rating scale, and defined several factorial structures by varying the dimensionality of the discrimination parameter matrix and the level of sparsity as described in [5]. To estimate the sparse structure, we focused on the median probability model ($ppi > 0.5$) [1], and we set $\tau = 0.01$ and $g = 100$ for the spike and slab prior. For each combination of parameters, we generated 15 graphs.
The results show that the low-dimensional representation of the networks is able to capture the different features of the simulated structures. The networks appeared to be clustered according to the different dimensionality of the discrimination matrices, levels of sparsity and relations between items and latent traits (Figure 1).

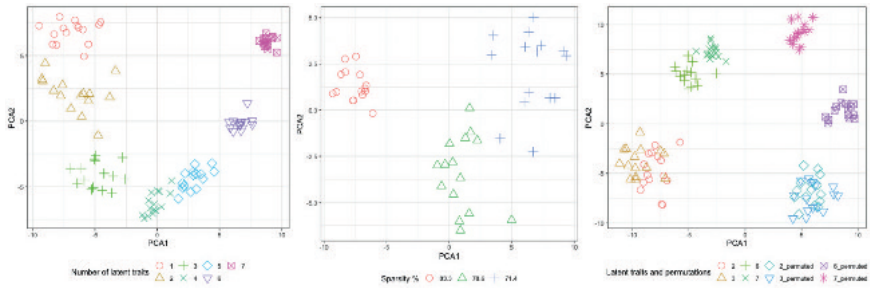


Fig. 1 Plot of the 2 first principal component scores for the simulated networks with varying numbers of latent traits, sparsity levels and item-latent traits structures.

5 Conclusion

In this paper, we discussed an exploratory analysis to investigate configural invariance of a test. Capitalising on the flexibility of Bayesian sparse modelling and the theory of graphical models, this approach provides a simple solution to explore the differences between groups of respondents. Through a simulation study, we demonstrated that this approach to EGA is able to capture differences in the latent structures.

References

1. Barbieri, M.M. and Berger, O.B.: Optimal Predictive Model Selection. The Annals of Statistics **32**(3):870-897 (2004). doi:10.1214/009053604000000238
2. Christensen, A.P. and Golino, H.: On the equivalency of factor and network loadings. Behavior Research Methods **53**:1563-1580 (2021). doi: 0.3758/s13428-020-01500-6.
3. Epskamp, S. and Fried, E.I.: A tutorial on regularized partial correlation networks. Psychological Methods **23**(4):617-634 (2018). doi:10.1037/met0000167.
4. Epskamp, S., Maris, G., Waldorp, L.J. and Borsboom, D.: Network psychometrics. In Irwing, P. , Booth, T. and Hughes, D.J.: The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development, 953–986. Wiley Blackwell (2018). doi10.1002/9781118489772.ch30
5. Fontanella, L., Fontanella, S.,Valentini, P. and Trendafilov, N.: Simple Structure Detection Through Bayesian Exploratory Multidimensional IRT Models. Multivariate Behavioral Research **54**(1):100-112(2019). doi:10.1080/00273171.2018.1496317.
6. Frühwirth-Schnatter, S. and Wagner, H.: Bayesian Variable Selection for Random Intercept Modeling of Gaussian and non-Gaussian Data. Bayesian Statistics 9. Oxford University Press (2010). doi: 10.1093/acprof:oso/9780199694587.003.0006.
7. George, E.I. and Mcculloch, R.E.: Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association **88**(423):881-889 (1993). doi:10.1080/01621459.1993.10476353.
8. Ginestet, C.E., Li, J. , Balachandran, P., Rosenberg, S. and Kolaczyk, E.D.: Hypothesis testing for network data in functional neuroimaging. The Annals of Applied Statistics **11**(2):725-750 (2017). doi:10.1214/16-AOAS1015
9. Golino, H.F. and Epskamp, S.: Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. PLoS ONE **12**(6)(2017). doi:10.1371/journal.pone.0174035.
10. Meinshausen, N. and Bühlmann, P.: High-dimensional graphs and variable selection with the lasso.Annals of Statistics **34**(3):1436 – 1462 (2006). doi:10.1214/009053606000000281.
11. Severn, K.E., Dryden, I.L. and Preston, S.P.: Manifold valued data analysis of samples of networks, with applications in corpus linguistics. The Annals of Applied Statistics **16**(1):368-390 (2022). doi:10.1214/21-AOAS1480
12. van Bork, R., Rhemtulla, M., Waldorp, L.J., Kruis, J., Rezvanifar, S. and Borsboom, D.: Latent Variable Models and Networks: Statistical Equivalence and Testability. Multivariate Behavioral Research **56**(2):175-198 (2021). doi:10.1080/00273171.2019.1672515.
13. Vandenberg, R.J. and Lance, C.E.: A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research.Organizational Research Methods **3**(1):4-70 (2000). doi:10.1177/109442810031002.

Penalized likelihood factor analysis

Analisi fattoriale di verosimiglianza penalizzata

Kei Hirose

Abstract In the factor analysis model, a penalized likelihood estimation has been recently used as an alternative to the rotation technique. This paper presents the relationship between penalized likelihood procedure and rotation techniques. Furthermore, two penalties related to conventional rotation criteria are described; minimax concave penalty (MCP) and *pr*oduct-based *e*lastic *net* (*prenet*) penalty.

Abstract *Nel modello di analisi dei fattori, una stima di verosimiglianza penalizzata è stata recentemente utilizzata come alternativa alla tecnica di rotazione. Questo articolo presenta la relazione tra la procedura di verosimiglianza penalizzata e le tecniche di rotazione. Inoltre, vengono descritte due penalità relative ai criteri di rotazione convenzionali; la penalità minimax concava (MCP) e la penalità prenet (product-based elastic net).*

Key words: Factor analysis, minimax concave penalty, penalization, product-based elastic net, rotation technique

1 Introduction

Factor analysis investigates the correlation structure of high-dimensional observed variables by constructing a small number of latent variables called common factors. Conventionally, a rotation technique has been used to find a simple structure of the loading matrix. Many rotation techniques have been proposed in the literature [1]. The main purpose of the rotation techniques is to get a good solution that is as simple as possible.

A problem with the rotation technique is that it cannot produce a sufficiently sparse solution in some cases because the loading matrix must be found among a

Kei Hirose
Institute of Mathematics for Industry, Kyushu University e-mail: hirose@imi.kyushu-u.ac.jp

set of unpenalized maximum likelihood estimates. We may employ a penalization method to obtain sparser solutions than the factor rotation. It is shown that the penalization is a generalization of the rotation techniques and can produce sparser solutions than the rotation methods [2]. Therefore, any rotation techniques can be extended to the penalization methods. For example, the L_1-type penalization, such as the lasso [3] and the minimax concave penalty (MCP) [4], is considered as an extension of one of the rotation criteria referred to as component loss criterion [5, 6]. The lasso and MCP have been widely used because they shrink some of the parameters toward exactly zero; in other words, parameters that need not to be modeled are automatically disregarded. Another penalty based on the rotation techniques is a *prenet* (*pr*oduct-based *e*lastic *net*) penalty [7], which is based on the product of a pair of parameters in each row of the loading matrix. The prenet penalty is considered as a generalization of the quartimin criterion [8], a widely-used oblique rotation method. A remarkable feature of the prenet is that a large amount of penalization leads to the perfect simple structure, a desirable structure in terms of the simplicity of the loading matrix. Furthermore, the perfect simple structure estimation via the prenet penalty is shown to be a generalization of the k-means clustering of variables.

In this paper, we briefly describe the penalized likelihood factor analysis based on the MCP and prenet penalties.

2 Penalized likelihood factor analysis

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional observed random vector with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The factor analysis model is

$$\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times m$ loading matrix, $\boldsymbol{F} = (F_1, \cdots, F_m)^T$ is a random vector of common factors, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)^T$ is a random vector of unique factors. It is assumed that $E(\boldsymbol{F}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \boldsymbol{I}_m$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi}$, and $E(\boldsymbol{F}\boldsymbol{\varepsilon}^T) = \boldsymbol{O}$, where $\boldsymbol{I}_m$ is an $m \times m$ identity matrix, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix. The diagonal elements of $\boldsymbol{\Psi}$ are referred to as unique variances. Under these assumptions, the covariance matrix of observed random vector $\boldsymbol{X}$ is $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$.

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be n observations and $\boldsymbol{S} = (s_{ij})$ be the corresponding sample covariance matrix. Let $\boldsymbol{\theta} = (\mathrm{vec}(\boldsymbol{\Lambda})^T, \mathrm{diag}(\boldsymbol{\Psi})^T)^T$ be a parameter vector. We estimate the model parameter by minimizing the penalized loss function $\ell_\rho(\boldsymbol{\theta})$

$$\ell_\rho(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \rho P(\boldsymbol{\Lambda}), \tag{2}$$

where $\ell(\boldsymbol{\theta})$ is a negative log-likelihood function expressed as

$$\ell_{\mathrm{DF}}(\boldsymbol{\theta}) = \frac{1}{2}\left\{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - \log|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}| - p\right\}, \tag{3}$$

$P(\boldsymbol{\Lambda})$ is a penalty function, and $\rho > 0$ is a regularization parameter.

3 Relationship with factor rotation

The model has a rotational indeterminacy; both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}\mathbf{T}$ generate the same covariance matrix $\boldsymbol{\Sigma}$, where $\mathbf{T}$ is an arbitrary orthogonal matrix. Thus, when $\rho = 0$, the solution that minimizes (2) is not uniquely determined. However, when $\rho > 0$, the solution may be uniquely determined except for the sign and permutation of columns of the loading matrix when an appropriate penalty $P(\boldsymbol{\Lambda})$ is chosen.

When $\rho = 0$, a rotation technique, such as the varimax method, has been widely used to find the matrix $\mathbf{T}$ that gives a meaningful relation between items and factors. Suppose that $Q(\boldsymbol{\Lambda})$ is an orthogonal rotation criterion at $\boldsymbol{\Lambda}$. The criterion is minimized over all orthogonal rotations with an initial loading matrix being $\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}$, i.e.,

$$\min_{\boldsymbol{\Lambda},\boldsymbol{\Psi}} Q(\boldsymbol{\Lambda}), \text{ subject to } \boldsymbol{\Lambda} = \hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}\mathbf{T} \text{ and } \mathbf{T}^T\mathbf{T} = \mathbf{I}_m. \tag{4}$$

Now we assume that the maximum likelihood estimates $\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}$ are unique if the indeterminacy of the rotation in $\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}$ is taken out. The problem (4) is then expressed as

$$\min_{\boldsymbol{\Lambda},\boldsymbol{\Psi}} Q(\boldsymbol{\Lambda}), \text{ subject to } \quad \ell(\boldsymbol{\Lambda},\boldsymbol{\Psi}) = \hat{\ell}, \tag{5}$$

where $\hat{\ell} = \ell(\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}, \hat{\boldsymbol{\Psi}}_{\mathrm{ML}})$.

The sparsity may be enhanced by modifying the problem (5) as follows:

$$\min_{\boldsymbol{\Lambda},\boldsymbol{\Psi}} Q(\boldsymbol{\Lambda}), \text{ subject to } \quad \ell(\boldsymbol{\Lambda},\boldsymbol{\Psi}) \leq \ell^*, \tag{6}$$

where ℓ^* ($\ell^* \geq \hat{\ell}$) is a constant value. The value ℓ^* controls the balance between the fitness of data and sparseness. When $\ell^* = \hat{\ell}$, the solution coincides with the maximum likelihood estimate. The estimate of $\boldsymbol{\Lambda}$ becomes sparse when ℓ^* is large. The problem in (6) can be solved by minimizing the following penalized log-likelihood function $\ell_\rho(\boldsymbol{\Lambda},\boldsymbol{\Psi})$:

$$\ell_\rho(\boldsymbol{\Lambda},\boldsymbol{\Psi}) = \ell(\boldsymbol{\Lambda},\boldsymbol{\Psi}) + \rho Q(\boldsymbol{\Lambda}), \tag{7}$$

where $\rho > 0$ is a regularization parameter.

Here, the rotation criterion $Q(\boldsymbol{\Lambda})$ can be viewed as a penalty function in the penalized maximum likelihood procedure, $P(\boldsymbol{\Lambda})$. The regularization parameter ρ controls the amount of shrinkage; that is, the larger the value of ρ, the greater the amount of shrinkage. When $\rho \to +0$, the solution in (7) becomes the maximum likelihood estimate with the rotation technique in (5). Thus, the penalized likelihood procedure can be viewed as a generalization of the maximum likelihood method with the rotation technique.

4 Two penalties

Because penalized likelihood procedure is a generalization of the maximum likelihood method with the rotation technique, any rotation techniques can be extended to the penalization methods. In this paper, we describe two penalties, MCP and prenet.

4.1 MCP

For ease of comprehension, we assume that the penalty term $P(\boldsymbol{\Lambda})$ is given by the component loss criterion, that is, $P(\boldsymbol{\Lambda}) = \sum_{i=1}^{p}\sum_{j=1}^{m} P(|\lambda_{ij}|)$ [5, 6]. An example of the component loss criterion is the lasso, which provides sparse solutions for some values of ρ. However, in our experience, the lasso estimates an overly dense model. To handle this issue, a nonconvex penalty can achieve sparser models than the lasso. In particular, the minimax concave penalty (MCP) [4] has been widely used:

$$\begin{aligned}\rho P(|\theta|;\rho;\gamma) &= \rho \int_0^{|\theta|} \left(1 - \frac{x}{\rho\gamma}\right)_+ dx \\ &= \rho \left(|\theta| - \frac{\theta^2}{2\rho\gamma}\right) I(|\theta| < \rho\gamma) + \frac{\rho^2\gamma}{2} I(|\theta| \geq \rho\gamma).\end{aligned}$$

For each value of $\rho > 0$, $\gamma \to \infty$ yields a soft threshold operator (i.e., lasso penalty) and $\gamma \to 1+$ produces a hard threshold operator.

4.2 Prenet

Another penalty based on the rotation criterion is the product-based elastic net (prenet) penalty:

$$P(\boldsymbol{\Lambda}) = \sum_{i=1}^{p}\sum_{j=1}^{m-1}\sum_{k>j}\left\{\gamma|\lambda_{ij}\lambda_{ik}| + \frac{1}{2}(1-\gamma)(\lambda_{ij}\lambda_{ik})^2\right\}, \tag{8}$$

where $\gamma \in (0,1]$ is a tuning parameter. The most significant feature of the prenet penalty is that it is based on the product of a pair of parameters. The prenet penalty is a generalization of the quartimin criterion [8]; setting $\gamma \to 0$ to the prenet penalty in (8) leads to the quartimin criterion

$$P_{\text{qmin}}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p}\sum_{j=1}^{m-1}\sum_{k>j}(\lambda_{ij}\lambda_{ik})^2.$$

The first term of the prenet penalty is to perform the sparse estimation of the loading matrix; with a sufficiently large ρ, some of the factor loadings are estimated to be exactly zero.

With the prenet penalization, we obtain the following proposition.

Proposition 1. *As $\rho \to \infty$, the estimated loading matrix possesses the perfect simple structure, that is, each row has at most one nonzero element.*

Proof. As $\rho \to \infty$, $P(\hat{\boldsymbol{\Lambda}})$ must satisfy $P(\hat{\boldsymbol{\Lambda}}) \to 0$. Otherwise, the second term of the right-hand side of (2) diverges. When $P(\hat{\boldsymbol{\Lambda}}) = 0$, $\hat{\lambda}_{ij}\hat{\lambda}_{ik} = 0$ for any $j \neq k$. Therefore, the ith row of $\boldsymbol{\Lambda}$ has at most one nonzero element.

The perfect simple structure is known as a desirable property in the factor analysis literature because it is easy to interpret the estimated loading matrix.

Furthermore, the perfect simple structure corresponds to variables clustering; variables that correspond to nonzero elements of the jth column of the loading matrix belong to the jth cluster. Thus, it would be interesting to investigate the relationship between prenet and conventional clustering methods. The following proposition shows the relationship between prenet and k-means clustering:

Proposition 2. *Assume that $\boldsymbol{\Psi} = \alpha \boldsymbol{I}_p$ and α is given. Suppose that $\boldsymbol{\Lambda}$ satisfies $\boldsymbol{\Lambda}^T\boldsymbol{\Lambda} = \boldsymbol{I}_m$. The prenet solution with $\rho \to \infty$ is obtained by an optimization problem that is a generalization of the k-means clustering.*

Proof. The proof appears in [7].

The proposition 2 shows that the prenet solution with $\rho \to \infty$ is a generalization of the k-means clustering of variables. We remark that the condition $\boldsymbol{\Psi} = \alpha \boldsymbol{I}_p$ in Proposition 2 implies the probabilistic principal component analysis (probabilistic PCA; [9]); thus, the penalized probabilistic PCA via the prenet is also a generalization of the k-means clustering of variables.

References

1. Browne, M. W.: An overview of analytic rotation in exploratory factor analysis. Multivariate behavioral research **36**(1), 111–150. (2001)
2. Hirose, K., Yamamoto, M.: Sparse estimation via nonconcave penalized likelihood in factor analysis model. Statistics and Computing **25**(5), 863–875. (2015)
3. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288. (1996)
4. Zhang, C. H.: Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics **38**(2), 894–942. (2010)
5. Jennrich, R. I.: Rotation to simple loadings using component loss functions: The orthogonal case. Psychometrika **69**(2), 257–273. (2004)
6. Jennrich, R. I.: Rotation to simple loadings using component loss functions: The oblique case. Psychometrika **71**(1), 173–191. (2006)
7. Hirose, K., Terada, Y.: Simple structure estimation via prenet penalization. arXiv preprint arXiv:1607.01145. (2016)

8. Carroll, J. B.: An analytical solution for approximating simple structure in factor analysis. Psychometrika **18**(1), 23–38. (1953)
9. Tipping, M. E., Bishop, C. M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622. (1999)

3 Solicited Sessions

Bayesian nonparametric modelling and learning

A regularized-entropy estimator to enhance cluster interpretability in Bayesian nonparametrics

Uno stimatore a entropia regolarizzata per migliorare l'interpretabilità dei cluster in bayesiana nonparametrica

Beatrice Franzolini, Giovanni Rebaudo

Abstract Bayesian nonparametric mixture models are widely used to cluster observations. However, one of the major drawbacks of the approach is that the estimated partition often presents only a few dominating clusters and a large number of sparsely-populated ones. This feature translates into results that are uninterpretable unless we accept to ignore a relevant number of observations and clusters. Here, we explain this phenomenon through the study of the cost functions involved in the estimation of the partition. Moreover, we propose a post-processing procedure to reduce the number of sparsely-populated clusters. The procedure takes the form of entropy-regularization of posterior cluster allocations. While being computationally convenient with respect to alternative strategies, it is also theoretically justified as a correction to the Bayesian loss function used for point estimation and, as such, can be applied to any posterior distribution of clusters, regardless of the specific Bayesian model used.

Abstract *I modelli Bayesiani nonparametrici con misture sono ampiamente utilizzati per effettuare cluster analysis. Tuttavia, uno dei principali limiti è il fatto che spesso identifichino un ampio numero di cluster poco popolati. Questa caratteristica si traduce in risultati di difficile interpretazione a meno che non si accetti di ignorare un numero di osservazioni e cluster. In questo lavoro, spieghiamo questo fenomeno attraverso lo studio delle funzioni di costo coinvolte nella stima della partizione. Inoltre, proponiamo una procedura di post-processing volta a ridurre il numero di cluster scarsamente popolati. La procedura prende la forma di una regolarizzazione dell'entropia dell'allocazione in cluster. La proposta appare computazionalmente conveniente rispetto a strategie alternative e trova giusticazione teorica in quanto correzione della funzione di perdita bayesiana impiegata nella stima puntuale, e, proprio per questa ragione, può essere adottata a prescindere dallo specifico modello utilizzato.*

Key words: Bayesian nonparametrics, Exchangeable partition probability function, Entropy, Clustering, Dirichlet process mixture

---

Beatrice Franzolini
Agency for Science, Technology and Research, Singapore, e-mail: franzolini@pm.me
Giovanni Rebaudo
Department of Statistics and Data Sciences, the University of Texas at Austin, USA,
e-mail: rebaudo.giovanni@gmail.com

1 Introduction

Clustering methods are used to detect patterns by partitioning observations into different groups. What are desirable characteristics of clusters depends on the specific applied problem at hand [see e.g., 13]. Nonetheless, clustering methods are typically motivated by the idea that observations are more similar within the same cluster than across clusters (accordingly to a certain definition of similarity).

Clustering has been proved useful in a large variety of fields including but not limited to image processing, bio-medicine, marketing, and natural language processing. Clustering methods are used not only to detect sub-groups of subjects, but also for dimensionality reduction [4, 23], outlier-detection [28, 21, 8], and data preprocessing [32]. Among clustering techniques, we can distinguish two main classes: model-based and non model-based.

Contrary to other popular clustering techniques, as k-means, model-based clustering methods allow us to perform inference via rigorous probabilistic assessments. Typically, model-based clustering frameworks are equivalent to the assumption that the observations $y_1, \dots y_n$ are extracted from an infinite population following a mixture

$$y_i \overset{iid}{\sim} \sum_{h=1}^{K} w_h k(\cdot;\theta_h) \qquad \text{i = 1,..., n,} \tag{1}$$

where the mixture components $k(\cdot;\theta_h)$ are probability kernels to be interpreted as distributions of distinct clusters in the infinite population, $(w_h, \theta_h)_{h=1}^{K}$ are unknown parameters that determine the relative proportion and the shape of such population clusters, and K is the total number of clusters in the population. K can be either a fixed value or an unknown parameter. However, the main goal of clustering techniques is to estimate a partition of the observed sample, more than the distribution of the whole ideal population in (1). The partition that one wants to estimate can be encoded using a sequence of subject-specific labels $(c_1, \dots, c_n)$ taking value in the set of natural numbers such that $c_i = c_j = c$ if and only if y_i and y_j belong to the same cluster and follow the same mixture component $k(\cdot;\theta_c)$, i.e. $y_i \mid c_i \overset{ind}{\sim} k(\cdot;\theta_{c_i})$ for $i = 1, \dots, n$. The indicators $(c_1, \dots, c_n)$, as just defined, are affected by the label switching problem [see, for instance, 29, 18, 10]. To overcome the issue, in the following, we assume them to be encoded in order of appearance. The likelihood for $\boldsymbol{c} = (c_1, \dots, c_n)$ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{K_n})$ is

$$\mathscr{L}(\boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) = \prod_{c=1}^{K_n} \prod_{i:c_i=c} k(y_i;\theta_c). \tag{2}$$

An important and typically unknown parameter is the number of clusters K_n observed in the sample, i.e., the number of occupied components. Obviously, $K_n \leq K$. For this reason, when we let n vary, finite fixed values for K are usually to be avoided and K is either fixed to $+\infty$ [e.g. in Dirichlet process mixtures, 7, 17] or it is estimated from the data [e.g. mixtures of finite mixtures, see 20, 1].

When K_n is unknown, the clustering labels in (2) cannot be estimated with a standard frequentist approach. In fact, when the maximum likelihood estimator (MLE) for (2) exists, it coincides with the vector of MLEs $(\hat{\theta}_1, \ldots, \hat{\theta}_n)$, where each $\hat{\theta}_i$ is obtained considering one observation at a time and the independent models $y_i \sim k(y_i \mid \theta_i)$, for $i = 1, \ldots, n$. Moreover, note that under typical mixture model assumptions for clustering, we have that $\hat{\theta}_1 \neq \ldots \neq \hat{\theta}_n$. For instance, when k is a multivariate Gaussian density and θ is the pair of mean vector and variance matrix of the Gaussian component, the MLE entails a number of clusters equal to the number of distinct observed values, that by model's assumptions equals n with probability 1. Thus, no information on clusters can ever be gained through MLE and overfitting is unavoidable unless one relies on strong restrictions of the parameter space. In this regard, note that maximizing (2) is not the same as computing the nonparametric maximum likelihood estimator [16, 26, 27] for the mixture model in (1).

Differently, Bayesian models, and in particular Bayesian nonparametric (BNP) models, are largely used for model-based clustering, since priors act as penalties shrinking the number of distinct clusters.

The content of the paper is organized as follows. Section 2 presents the study of the cost functions involved in BNP clustering models and explains a common drawback, i.e., the presence of noisy and sparsely populated clusters typically observed in the posterior estimates of these models. Then, a computationally convenient and theoretically justified solution to reduce the number of sparsely populated clusters is presented in Section 3 and showcased on simulated and real data, respectively in Sections 4 and 5.

2 Implied costs functions in Bayesian nonparametric clustering

The vast majority of Bayesian models for clustering rely on a prior for $\boldsymbol{c}$ and K_n defined through an exchangeable partition probability function (EPPF) [see, 24] and, independently, a prior P is used for the unique values $(\theta_1, \ldots, \theta_{K_n})$. Therefore, the corresponding posterior distribution is

$$p(K_n, \boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{c=1}^{K_n} \prod_{i:c_i=c} k(y_i; \theta_c) \times \text{EPPF}(n_1, \ldots, n_{K_n}) \times P(d\boldsymbol{\theta}), \tag{3}$$

which can be equivalently represented as the cost function $-\log(p(K_n, \boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{y}))$, i.e.

$$C(K_n, \boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) = C_{\text{lik}}(K_n, \boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) + C_{\text{part}}(K_n, \boldsymbol{c}; \alpha) + C_{\text{base}}(K_n, \boldsymbol{\theta}),$$

which is the sum of three terms, that in the following are named respectively likelihood cost, partition cost, and base cost.

As already mentioned, the minimum of the likelihood cost

$$C_{\text{lik}}(K_n, \boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) = -\sum_{c=1}^{K_n} \sum_{i:c_i=c}^{n} \log k(y_i; \theta_c)$$

typically corresponds to K_n equal to the number of distinct observed values. The remaining two costs are those defined by the prior of the model and their marginal behavior is described here below. Clearly, any inference result has to be derived based on the whole posterior distribution in (3), which is the result of the joint, and not marginal, effect of all three costs. Nonetheless considering one cost at a time allows us to gain insights regarding the estimation procedure and the frequentist penalties induced by the prior. A lot of attention in the literature has been devoted to the choice of the EPPF and many alternatives are available [see, for example, 14, 15, 6, 12], while, except for few cases [22, 31, 2], the role of the base cost appears partially overlooked within the Bayesian methodology literature.

However, when BNP clustering methods are applied in practice, the choice of an appropriate base distribution is known to be crucial. The most common choice is to use an independent prior on the unique values so that $\theta_c \overset{iid}{\sim} P_0$ and

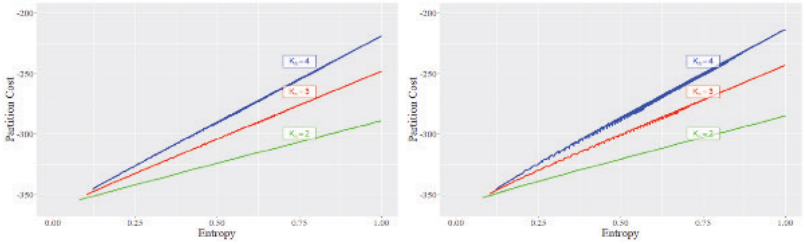
$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = -\sum_{c=1}^{K_n} \log P_0(d\theta_c),$$

where the variance of the distribution P_0 is known to play an important role in the estimation process and, typically, the higher the variance of P_0 the lower the number of clusters identified by the posterior [cfr., e.g. 9, p. 535]. This phenomenon can be explained by looking at the joint distribution induced by P_0 on the unique value. For instance, when P_0 is set to be a univariate normal distribution centered in μ and with variance σ^2, we have

$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = \frac{K_n}{2} \log(2\pi) + \frac{K_n}{2} \log \sigma^2 + \frac{1}{2} \sum_{c=1}^{K_n} \frac{(\theta_c - \mu)^2}{\sigma^2}.$$

When the variance is increased from σ^2 to λ^2, intuitively the base cost increases for those vectors $(\theta_1, \ldots, \theta_{K_n})$ whose components are similar and it decreases for vectors with more diverse components, thus ultimately favoring the variability of the unique values and penalizing many overlapping clusters. More formally, defining the K_n-sphere $\boldsymbol{\theta} \in \mathbb{R}^{K_n}$ such that $\sum_{c=1}^{K_n} (\theta_c - \mu)^2 = K_n \frac{\log(\lambda^2/\sigma^2)\sigma^2\lambda^2}{\lambda^2 - \sigma^2}$, we have that the cost increases for vectors $(\theta_1, \ldots, \theta_{K_n})$ corresponding to points inside the sphere and decreases for those vectors corresponding to points outside the sphere. In practice, P_0 is usually set to be a continuous scale mixture, where the mixed density is conjugate to the kernel k for computational convenience, while the mixing density is used to increase appropriately the marginal scale of the mixture P_0.

Finally, let us comment on the partition cost C_{part}. Its behavior is less straightforward and we consider here only two important and widely used cases: Dirichlet process mixtures (DPM) and Pitman-Yor process [25] mixtures (PYPM). With a DPM model, up to an additive constant, we have



(a) (b)

Fig. 1: Partition cost as function of entropy in a DPM model with $\alpha = 1$ (panel a) and in a PYPM model with $\alpha = 1$ and $\sigma = 0.5$ (panel b) for $n = 100$ observations clustered into 2 (blue line), 3 (red line), and 4 (green line) clusters.

$$C_{\text{part}}(K_n, \boldsymbol{c}; \alpha) = -K_n \log \alpha - \sum_{c=1}^{K_n} \log \Gamma(n_c),$$

where α is the concentration parameter of the Dirichlet Process. The DPM partition cost tends to favor parsimonious values of K_n (wrt to the likelihood cost that in general tends to favor $K_n = n$). However, contrary to the base cost, it depends also on clusters' frequencies.

Figure 1(a) showcases the partition cost of DPM for different values of what we refer henceforth to as the entropy of the frequencies $(n_1, \ldots, n_{K_n})$, i.e.

$$S(n_1, \ldots, n_{K_n}) = -\sum_{c=1}^{K_n} \frac{n_c}{n} \log_{K_n} \frac{n_c}{n}.$$

Overall the EPPF acts favoring frequencies $(n_1, \ldots, n_{K_n})$ with low entropy and thus, roughly speaking, higher sample variance of the frequencies. However, this feature ultimately results in two distinct effects: one acting on the total number of occupied clusters K_n and another acting on the variance of the clusters' frequencies $(n_1, \ldots, n_{K_n})$. Even though these two features both favor a reduced entropy, they entail very different scenarios in terms of estimated clustering structure, especially from an applied and practical point of view. Penalizing large numbers of clusters is typically desirable in applications because an elevated number of clusters may be difficult to interpret, however a partition with few dominating clusters and many sparsely populated clusters is highly undesirable because it is hard to interpret unless one decides to ignore all the information contained in the small clusters and focus only on the dominating ones. See also [11] for a study of the posterior entropy in the Dirichlet process mixture and [12] for more details on entropy in mixture of finite mixture models. In the case of a PYPM the partition cost, up to an additive constant, equals

$$C_{\text{part}}(K_n, \boldsymbol{c}; \alpha, \sigma) = -\sum_{c=1}^{K_n} \log(\alpha + \sigma(c-1)) - \sum_{c=1}^{K_n} \log \Gamma(n_c - \sigma) + K_n \log \Gamma(1-\sigma).$$

Despite that the EPPFs are different, Figure 1 shows in both processes a closely similar behavior in terms of entropy penalization.

Note that Figure 1 provides us with insights into the behavior of the EPPFs evaluated in a vector of clusters' frequencies $(n_1,\dots,n_{K_n})$, i.e., the probability of a specific clustering configuration with unordered frequencies $\{n_1,\dots,n_{K_n}\}$. Note that the vectors $(n_1,\dots,n_{K_n})$ are not in a one-to-one correspondence with the partitions and the number of partitions corresponding to certain frequencies varies across vectors. The same is true for other marginal quantities such as the number of clusters K_n. For instance, the number of possible partitions rapidly increases with K_n accordingly to Stirling numbers of the second kind. Importantly, this information must also be considered combined with the partition cost evaluated in a specific partition, represented in Figure 1, if we are interested in fully understanding the impact of the EPPF on prior and posterior distributions of functionals of the partition, e.g., on the marginal distribution of K_n. Note that combining the two features the typical partition cost strongly penalized too many clusters suggested by the likelihood costs, i.e. $K_n = n$, but favors a small number of clusters with respect to n that adaptively increases with the sample size n, [See e.g., 6]. Considering both aspects is also important if we want to understand the effect of the partition cost on a point estimate of the clustering that is different from the MAP (maximum a posteriori) of the partition, but minimizes the Bayesian risk, i.e., posterior expected loss, according to flexible loss as discussed in the next section.

3 Regularized-entropy estimator

Once the posterior distribution $\mathbb{P}(\boldsymbol{c} \mid y_{1:n})$ over the space of partitions is obtained, typically thanks to a Markov Chain Monte Carlo algorithm, a point estimate $\hat{\boldsymbol{c}}$ of the partition can be obtained accordingly to the decision-theoretic approach of Bayesian analysis. More precisely, $\hat{\boldsymbol{c}}$ is obtained by minimizing the Bayesian risk, i.e, the expected value of a loss function $L(\boldsymbol{c},\hat{\boldsymbol{c}})$ with respect to the posterior:

$$\boldsymbol{c}^* = \underset{\hat{\boldsymbol{c}}}{\operatorname{argmin}}\, \mathbb{E}[L(\boldsymbol{c},\hat{\boldsymbol{c}}) \mid y_{1:n}] = \underset{\hat{\boldsymbol{c}}}{\operatorname{argmin}} \sum_{\boldsymbol{c}} L(\boldsymbol{c},\hat{\boldsymbol{c}})\mathbb{P}(\boldsymbol{c} \mid y_{1:n}),$$

where $L(\boldsymbol{c},\hat{\boldsymbol{c}})$ is the loss in which we incur using $\hat{\boldsymbol{c}}$ as estimates when the partition takes the value $\boldsymbol{c}$. How to interpret and elicit the loss in practice can change according to the philosophical point of view. Often in parameter estimation the loss is interpreted as the cost of choosing $\hat{\boldsymbol{c}}$ instead of the ideally optimal parameter value $\boldsymbol{c}$ (sometimes interpreted as the *truth*). In a more subjective Bayesian framework, it can be interpreted, together with the model and prior, in terms of the preferences implied on the possible parameter values $\boldsymbol{c}$ via the Bayesian risk. Finally, also in a more frequentist framework the loss can be chosen in terms of the implied properties of the estimator of the unknown parameter $\hat{\boldsymbol{c}}$.

Despite the different philosophical justifications, rarely in applied Bayesian clustering analysis a 0-1 loss function and the resulting MAP estimator are employed

Algorithm 1 Entropy-regularized estimates

Input: MCMC chain of partitions $\{\boldsymbol{c}_m, m = 1,\dots,M\}$, λ
Output: point estimate $\boldsymbol{c}^*$

1: Compute $S(\boldsymbol{c}_m)$ for $m = 1,\dots,M$
2: Compute $w_m = \exp\{\lambda S(\boldsymbol{c}_m)\}$ for $m = 1,\dots,M$
3: $\bar{w}_m \leftarrow w_m / \sum_m w_m$ for $m = 1,\dots,M$
4: Generate $\{\tilde{\boldsymbol{c}}_m, m = 1,\dots,M\}$, sampling with replacement from $\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_M\}$ with prob. $\{\bar{w}_m, m = 1,\dots,M\}$
5: $\boldsymbol{c}^* \leftarrow \text{argmin} \sum_{m=1}^{M} \sum_{\hat{\boldsymbol{c}}} L(\tilde{\boldsymbol{c}}_m, \hat{\boldsymbol{c}})$

due to the large support of the posterior and the fact that the 0-1 loss function does not reflect different levels of distance between two non-coinciding partitions. Widely used alternatives in applications are Binder loss [3] or variation of information loss [see, 19, 30, 5].

We have already stressed how a large presence of noisy clusters is typically undesirable in practice and we claim that this aspect should be reflected in the loss function used for point estimation, so that the loss of each partition is proportional to its entropy. To do so, consider any possible loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ one would like to use to derive the estimate, we can define a new loss function, that we named entropy-regularized, as

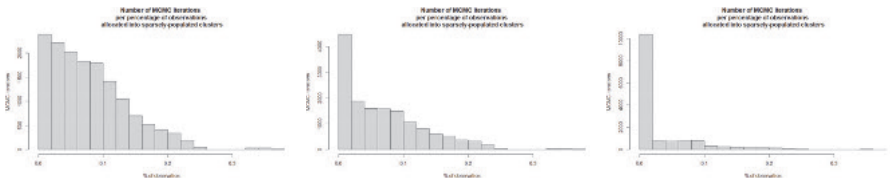
$$\bar{L}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \exp\{\lambda S(\boldsymbol{c})\} L(\boldsymbol{c}, \hat{\boldsymbol{c}}),$$

where, with a little abuse of notation wrt the previous section, $S(\boldsymbol{c})$ is the entropy of the partition identified by $\boldsymbol{c}$ and $\lambda \in \mathbb{R}$. Recall that the base of the logarithm involved in the computation of $S(\boldsymbol{c})$ changes with the argument $\boldsymbol{c}$ and it is equal to the number of unique values in $\boldsymbol{c}$, so that $S(\boldsymbol{c}) = 1$ can be obtained for any number of non-empty clusters $K_n \geq 2$ (provided that $n/K_n \in \mathbb{N}$). Clearly, when λ is positive, for any candidate estimate $\hat{c}$, the loss function is inflated in correspondence of partitions $\boldsymbol{c}$ with high entropy, as desired.

Minimizing the expected entropy-regularized loss function $\bar{L}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ with respect to the posterior is equivalent to minimizing the original loss function $L(\boldsymbol{c}, \hat{\boldsymbol{c}})$ with respect to an entropy-regularized version $\bar{\mathbb{P}}[\boldsymbol{c} \mid y_{1:n}]$ of the posterior distribution, i.e.

$$\bar{\mathbb{P}}[\boldsymbol{c} \mid y_{1:n}] \propto \exp\{\lambda\, S(\boldsymbol{c})\} \mathbb{P}[\boldsymbol{c} \mid y_{1:n}].$$

This result, while immediate to prove, is highly desirable, because it allows implementation of the entropy-correction in a very straightforward and computationally feasible way which is described in Algorithm 1.



(a) Without entropy regularization. (b) With entropy regularization for $\lambda = 10$. (c) With entropy regularization for $\lambda = 20$.

Fig. 2: Percentage of observations in sparsely-populated clusters before and after entropy-regularization.

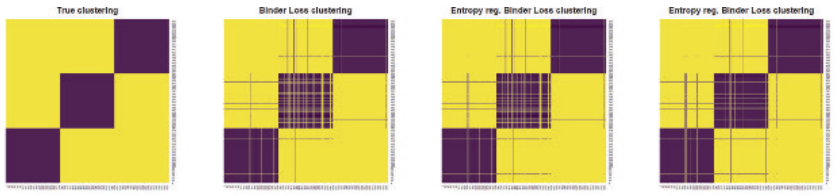
4 Simulation study

We provide here a simulation study, where $n = 1000$ observations are sampled from 3 different univariate Gaussian distributions. Here we refer to "true" clustering as the one implied by the memberships indicators of the Gaussian kernels under the data generating truth. We employ a normal-normal DPM and we compare the posterior estimates obtained minimizing the Binder loss function and the entropy-regularized Binder loss function. We set the concentration parameter $\alpha = 1$, perform 20 000 MCMC simulations, and use the first 5000 as burnin. Defining as sparsely populated clusters those clusters containing 10% or less of observations, we found that in almost a third (4755 out 15 000) of the MCMC iterations, 10% or more of the observations are allocated into sparsely populated clusters, while in almost two thirds (9306 out of 15 000) of MCMC iterations, 5% or more of the observations are allocated into sparsely populated clusters, see Figure 2a. The same counts after entropy-regularization of the posterior (as described in the previous section) are, with $\lambda = 10$, 3981 and 7825 out 15 000, see Figure 2b, and, with $\lambda = 20$, 1393 and 3366 out 15 000, see Figure 2c. However, notice that coherently with the interpretation of the regularization in terms of the loss function, the regularized posterior should be intended only as a computational tool to provide a summary of the posterior distribution and not as an alternative posterior. So that, for instance, uncertainty quantification should be computed using the original posterior.

Finally, Figure 3 shows the true and the estimated clusters with and without entropy regularization and they highlight how the regularization acts allocating observations from noisy clusters into dominating ones. Finally, Figure 4 shows the cluster frequencies for the three point estimates.

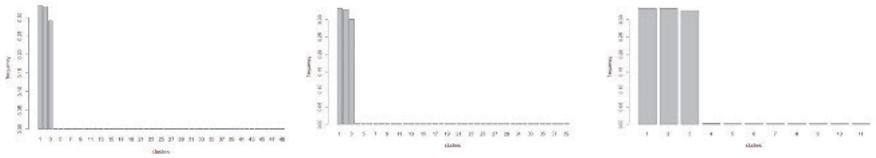
5 Results for the wine dataset

We test the performance of our method also on the wine dataset available on R, where data are the results of a chemical analysis of wines grown in the same region



(a) (b) (c) (d)

Fig. 3: Estimated clustering for the simulation study darker squares denote couples of observations clustered together. Panel (a) shows the true clustering. Panel (b) shows the clustering minimizing the Binder loss. Panel (c) shows the clustering minimizing the entropy-regularized Binder loss for $\lambda = 10$ and panel (d) for $\lambda = 20$.



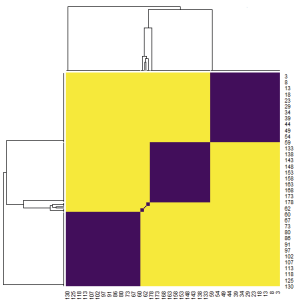
(a) Without entropy regularization.

(b) With entropy regularization for $\lambda = 10$.

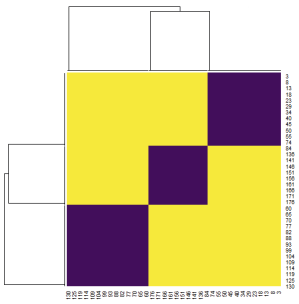
(c) With entropy regularization for $\lambda = 20$.

Fig. 4: Estimated clusters' frequencies.

in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Here we refer to the clustering identified by the three types of wines as "ground truth". We use the 13 constituents to estimate a Dirichlet process mixture model with multivariate Gaussian kernels, and we try to recover the three groups of types of wine through the estimated clustering. After running the MCMC for 10000 iterations and using the first 2000 as burnin, the Binder loss function identifies a partition of seven clusters, while our estimator for $\lambda = 20$ identifies three clusters. See Figure 5 and Figure 6. Lastly, Figure 7 compares the clustering based on three groups of types of wine with the two estimates.

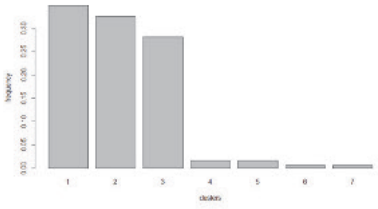


(a) Estimated partition without entropy-regularization.

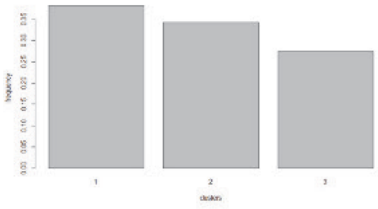


(b) Estimated partition after entropy-regularization.

Fig. 5: Estimated partitions for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based on co-clustering.

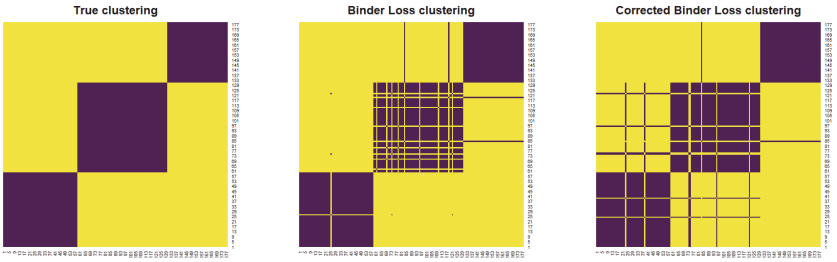


(a) Without entropy-regularization.



(b) With entropy-regularization.

Fig. 6: Estimated clusters' frequencies for the wine dataset.



(a) (b) (c)

Fig. 7: Estimated clustering for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based three groups of types of wine. Panel (a) shows the clustering ground truth. Panel (b) shows the clustering minimizing the Binder loss. Panel (c) shows the clustering minimizing the entropy-regularized Binder loss for $\lambda = 20$.

References

[1] Argiento, R. and M. De Iorio (2019). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Preprint arXiv: 1904.09733*.

[2] Beraha, M., R. Argiento, J. Møller, and A. Guglielmi (2021). MCMC computations for Bayesian mixture models using repulsive point processes. *J. Comput. Graph. Stat.*, in press.

[3] Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika 65*, 31–38.

[4] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

[5] Dahl, D. B., D. J. Johnson, and P. Müller (2021). Search algorithms and loss functions for Bayesian clustering. *Preprint arXiv: 2105.04451*.

[6] De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell. 37*, 212–229.

[7] Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pp. 287–302. Elsevier.

[8] Franzolini, B., A. Lijoi, and I. Prünster (2022). Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes. *Ann. Appl. Stat.*, in press.

[9] Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

[10] Gil-Leyva, M. F., R. H. Mena, and T. Nicoleris (2020). Beta-Binomial stick-breaking non-parametric prior. *Electron. J. Stat. 14*, 1479–1507.

[11] Green, P. J. and S. Richardson (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat. 28*, 355–375.

[12] Greve, J., B. Grün, G. Malsiner-Walli, and S. Frühwirth-Schnatter (2022). Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Aust. N. Z. J. Stat.*, in press.

[13] Hennig, C. (2015). What are the true clusters? *Pattern Recognit. Lett. 64*, 53–62.

[14] Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Series B Stat. Methodol. 69*, 715–740.

[15] Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker (Eds.), *Bayesian nonparametrics*. Cambridge University Press.

[16] Lindsay, B. G. (1995). Mixture models: theory, geometry, and applications. In *NSF-CBMS Regional Conf. Series in Prob. and Stat.*, Volume 5.

[17] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat. 12*, 351–357.

[18] McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annu. Rev. Stat. Appl. 6*, 355–378.

[19] Meilă, M. (2007). Comparing clusterings—an information based distance. *J. Multivar. Anal. 98*, 873–895.

[20] Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *J. Am. Stat. Assoc. 113*, 340–356.

[21] Ngan, H. Y., N. H. Yung, and A. G. Yeh (2015). Outlier detection in traffic data based on the Dirichlet process mixture model. *IET Intell. Transp. Syst. 9*, 773–781.

[22] Petralia, F., V. Rao, and D. Dunson (2012). Repulsive mixtures. *Adv. Neural Inf. Process. Syst. 25*, 1889–1897.

[23] Petrone, S., M. Guindani, and A. E. Gelfand (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Stat. Soc. Series B Stat. Methodol. 71*, 755–782.

[24] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lect. notes-monogr. ser. 30*, 245–267.

[25] Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab. 25*, 855–900.

[26] Polyanskiy, Y. and Y. Wu (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *Preprint arXiv: 2008.08244*.

[27] Saha, S. and A. Guntuboyina (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann. Stat. 48*, 738–762.

[28] Shotwell, M. S. and E. H. Slate (2011). Bayesian outlier detection with Dirichlet process mixtures. *Bayesian Anal. 6*, 665–690.

[29] Stephens, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Series B Stat. Methodol. 62*, 795–809.

[30] Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal. 13*, 559–626.

[31] Xu, Y., P. Müller, and D. Telesca (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics 72*, 955–964.

[32] Zhang, C., Y. Qin, X. Zhu, J. Zhang, and S. Zhang (2006). Clustering-based missing value imputation for data preprocessing. In *4th IEEE Int. Conf. Industr. Inform.*, pp. 1081–1086.

Exact confidence sets from credible sets with finite amounts of data

Regioni di confidenza esatte a partire da regioni di credibilità con dati finiti

B. J. K. Kleijn

Abstract Any model/prior pair that induces a sharp posterior concentration inequality for (metric) neighbourhoods of the truth, permits the interpretation of (metric) enlargements of credible sets of high-enough credible level, as confidence sets of a chosen confidence level. This construction of exact, finite-sample confidence sets is applied for community detection in the sparse two-community stochastic block model, and we argue that the methodology can also be applied in other statistical questions, like trace reconstruction in population genetics. We briefly speculate on computational aspects and the potential extent of applicability of the resulting methods more generally.

Abstract *Ogni coppia modello/distribuzione-a-priori che induca una disuguaglianza di concentrazione fine per intorni (metrici) della verità permette di interpretare gli allargamenti (metrici) delle regioni di credibilità con un livello di credibilità sufficientemente alto come regioni di confidenza con un livello di confidenza a scelta. Applichiamo questa costruzione di regioni di confidenza esatte con un campione finito al rilevamento di comunità tramite il modello sparso a blocchi stocastici per due comunità. Argomentiamo che la stessa metodologia può essere applicata ad altre domande statistiche, come la ricostruzione della traccia genetica nelle popolazioni. Ci soffermiamo brevemente sugli aspetti computazionali e sulla potenziale applicabilità dei metodi che ne derivano in un'ottica più generale.*

Key words: posterior concentration, uncertainty quantification, exact finite-sample confidence set, community detection, sparse random graph

B. J. K. Kleijn
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands, e-mail: kleijn@hushmail.com

1 Statistical uncertainty quantification

The practical value of statistical conclusions is often greatly enhanced if one has an idea of the accuracy of the methods used. Based on the statistical model for the observation, observational randomness results in randomness of methodological outcomes and the analysis of their variability broadly defines *statistical uncertainty quantification*. In frequentist statistics uncertainty quantification is commonly based on the distributions of estimators over the model, leading to so-called confidence sets: we use the data to calculate a whole set of possible estimates, a *confidence set*, that captures the true value with a prescribed probability called the *confidence level*. In Bayesian statistics, estimation uncertainty is quantified with *credible sets*, which contain a prescribed fraction of the posterior probability called the *credible level*. Because for many purposes the interpretation of confidence sets is more natural than that of credible sets, we assume the frequentist perspective and investigate which role Bayesian credible sets can play in the construction of confidence sets.

Questions regarding uncertainty quantification are especially prominent in the more modern forms of data science. In spite of great developments over the past decades, it is not always straightforward to combine machine-learning methods with mathematical statistics: a statistical model is not always defined, often the data does not adhere to standard statistical assumptions, and computability (not of primary importance in mathematical statistics) is a first requirement. Consequently machine learning provides very interesting and completely new algorithmic forms of statistical estimation, that do not lend themselves well to the analysis of statistical uncertainty: it is unfortunately quite common in algorithmic estimation that there is no known construction (or even approximation) for confidence sets.

Bayesian statisticians calculate (or approximate) posterior distributions, making it relatively easy to obtain (approximate) credible sets, simply as the smallest sets (point sets, intervals, ellipsoids, *etcetera*) of the required credible level. By contrast frequentist confidence sets are usually more difficult to obtain: finding confidence sets requires detailed knowledge of the sampling distribution of an estimator, which is available only in the simplest of statistical models. Of course frequentists have found ways to approximate confidence sets: *e.g.* in smooth parametric models for *i.i.d.* samples, so-called *Wald sets* capture the true, underlying value of the parameter with chosen confidence level in the large-sample asymptotic approximation. Alternatively Efron's *bootstrap methods* can be used to find approximate confidence sets, or Bayesian credible sets can serve as approximate confidence sets: the celebrated Bernstein-von Mises theorem implies that Wald sets and credible sets are indistinguishable in the large-sample limit (Le Cam and Yang, 2000). In non-parametric statistics some first examples have been studied: aside from early negative examples Freedman (1999), a semi-parametric Bernstein-von Mises theorem with Gaussian process priors was analysed in (Castillo, 2012a,b) and a semi-parametric Bernstein-von Mises theorem for general priors was obtained in (Bickel and Kleijn, 2012). A Bernstein-von Mises theorem for certain functionals in a Gaussian white noise model was given in (Castillo and Nickl, 2013); adaptive confidence sets in smoothness classes with conjugate Gaussian priors were studied in (Szabó et al., 2015);

a general conversion of *sequences* of credible sets to *sequences* of confidence sets was proposed in (Kleijn, 2021). However all of these proposals involve large-sample limits and due to ever-present computational limitations, asymptotic approximations are not always convenient (or even applicable) in practically feasible circumstances.

The challenge is to find new methods for the construction of (exact) confidence sets that apply to the types of data and models that modern statistics and machine learning have introduced, taking into account also limitations to computational capacities with regard to sample sizes. Below we demonstrate the conversion of (relatively easily obtained) Bayesian credible sets into exact frequentist confidence sets. More precisely, it is shown which minimal credible level is required for a credible set (or its metric enlargement) to be a confidence set of a certain desired confidence level, given sample size and other parameters. To be practical, this conversion must be *exact with finite amounts of data*, because computational limitations often preclude the practical applicability of approximations based on the large-sample limits of preceding references.

We develop the methodology theoretically in section 2. In section 3 we consider a network-science application (community detection in the stochastic block model), and demonstrate that credible sets and their enlargements can serve as exact confidence sets for finite graph sizes. Methods discussed are generalizable to many other applications, like in the population-genetics application of section 4 (trace reconstruction, in which the observed data consists of binary sequences with noise in the form of random deletions, mutations or insertions).

2 Posterior concentration and confidence levels of credible sets

The identification of enlarged credible sets as confidence sets is based on concentration of posterior probability. Below, we develop a perspective that is fully general, in the sense that we do not make many assumptions on the form of the data X (which takes values in a sample space we denote by $\mathscr{X}$), nor on the model used to analyse it: we assume that $X \sim P_\theta$, where θ is an identifiable parameter from a parameter space Θ, which is endowed with a prior distribution Π and associated posterior distribution $\Pi(\cdot|X)$. For the sake of simplicity and to emphasize that the construction holds with finite amounts of data, there is no index n in the discussion of this section. In subsequent sections, we reintroduce graph-/sample-size in our notation.

By posterior concentration, we mean that for every $\theta \in \Theta$, there exists a $U(\theta)$ and $\beta > 0$ such that,

$$E_\theta \Pi\big(U(\theta) \,\big|\, X\big) \geq 1 - \beta. \tag{1}$$

Of course we think of $U(\theta)$ as a small set around the point θ, and β as a small amount of posterior mass that remains outside $U(\theta)$. In most cases the $U(\theta)$ are topological neighbourhoods of the points θ and often the statement is formulated sequentially (with everything depending on some index n that denotes 'sample size' or 'graph size'). Most familiar is the setting where the $U(\theta)$ are (closed) metric balls

with respect to some metric d on Θ,

$$U(\theta) = B(\theta, r) = \{\, \theta' \in \Theta : d(\theta', \theta) \leq r \,\}, \tag{2}$$

for some radius $r \geq 0$ (with $B(\theta, 0) = \{\theta\}$). Posterior concentration forms the centrepiece of the theory of posterior asymptotic convergence: a sequence of posteriors is *consistent* if, for all points in the parameter space(s) and all their neighbourhoods, the sequential version of (1) holds with a sequence of β's that goes to zero. In metric setting, a sequence of posteriors converges at a rate, if, for all points in the parameter space(s) and a sequence of radii that goes to zero, the sequential version of (1), with shrinking balls (2) of said radii holds with a sequence of β's that goes to zero. For that reason, assertions of the type (1) have been one of the main focal points of the theory of non-parametric and asymptotic Bayesian statistics and many statistical models have known bounds of the type (1) (see, *e.g.* (Ghosal and van der Vaart, 2017)).

Note that the posterior probabilities β are not of material significance in the asymptotic perspective: as long as they go to zero, consistency and convergence at a rate are valid. However, posterior concentration as in (1) has another use, for which the value of β plays a central role. Before proving the corresponding lemma, let us sketch the argument that will lead from credible to confidence sets in the metric setting: according to (1), a fraction $1-\beta$ of the posterior mass is concentrated in a ball of radius r around the unknown true value θ of the parameter (in expectation). Hence, any credible set of high enough credible level *must have non-empty intersection with* $B(\theta, r)$ (with high probability), since the total posterior mass does not exceed one. That implies that θ lies at a distance smaller than or equal to r from any credible set of high enough level (with high probability). From this, one deduces that a credible set $D(X)$ enlarged by the radius r can serve as a confidence set, as formalized in the following central lemma.

Lemma 1. *Let Θ be the parameter space for a model $\{P_\theta : \theta \in \Theta\}$ for data X, with prior Π. For all $\theta \in \Theta$, let $U(\theta)$ be subsets of Θ. Assume that for $\theta_0 \in \Theta$ the expected posterior probability of $U(\theta_0)$ is lower-bounded,*

$$E_{\theta_0} \Pi\big(U(\theta_0) \,\big|\, X\big) \geq 1 - \beta, \tag{3}$$

for some $0 < \beta < 1$. For any $0 < \gamma < 1$ and any credible set $D(X) \subset \Theta$ of credible level $1-\gamma$,

$$P_{\theta_0}\big(U(\theta_0) \cap D(X) \neq \varnothing\big) \geq 1 - \frac{\beta}{1-\gamma}.$$

Proof. We first prove that for every $0 < s < 1$,

$$P_{\theta_0}\big(\Pi(U(\theta_0)|X) \geq s\big) \geq 1 - \frac{\beta}{1-s},$$

by contradiction: let $\delta > 0$ be given and define the event,

$$E = \big\{\, x \in \mathcal{X} : \Pi(U(\theta_0)|X = x) \geq s \,\big\}.$$

Suppose that $P_{\theta_0}(E) \leq 1-\beta/(1-s)-\delta$. Then,

$$E_{\theta_0}\Pi(U(\theta_0)|X) \leq P_{\theta_0}(E)+s(1-P_{\theta_0}(E)) \leq 1-\beta-\delta(1-s) < 1-\beta, \quad (4)$$

which contradicts the assumption that $E_{\theta_0}\Pi(U(\theta_0)|X) \geq 1-\beta$. Since this holds for every $\delta > 0$, we have $P_{\theta_0}(E) \geq 1-\beta/(1-s)$. Choose $s > \gamma$. As $D(X)$ has posterior mass at least $1-\gamma$, $U(\theta_0)$ and $D(x)$ cannot be disjoint for $x \in E$. So,

$$P_{\theta_0}\big(U(\theta_0) \cap D(X) \neq \varnothing\big) \geq P_{\theta_0}(E) \geq 1-\frac{\beta}{1-\gamma},$$

which proves the assertion.

This lemma has the following implication for the U-enlargement of credible sets of high enough credible level (see Figure 1).

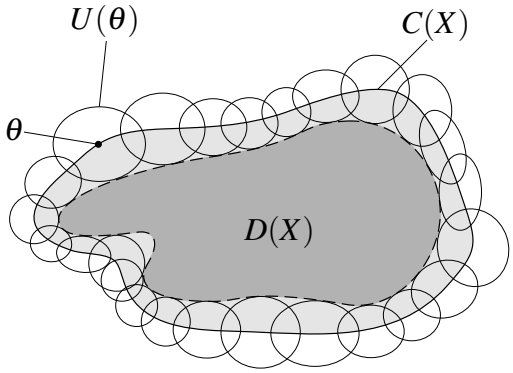


Fig. 1 The relation between a credible set $D(X)$ and its U-enlargement $C(X)$ in Venn diagrams: the extra points θ in $C(X)$ not included in the credible set $D(X)$ are characterized by non-empty intersection $U(\theta) \cap D(X) \neq \varnothing$. [From (Kleijn, 2021)]

Corollary 1. *Assume that (3) holds for some $\beta > 0$. If $U(\theta)$ contains θ for all $\theta \in \Theta$, then, for any level-$(1-\gamma)$ credible set $D(X)$, the set,*

$$C(X) = \{\theta \in \Theta : U(\theta) \cap D(X) \neq \varnothing\}, \quad (5)$$

is a confidence set of confidence level $1-\beta/(1-\gamma)$, i.e. for all $\theta_0 \in \Theta$,

$$P_{\theta_0}\big(\theta_0 \in C(X)\big) \geq 1-\frac{\beta}{1-\gamma}.$$

Proof. The assertion of the above lemma says that:

$$P_{\theta_0}\big(U(\theta_0)\cap D(X)\neq\varnothing\big)\geq 1-\frac{\beta}{1-\gamma}, \tag{6}$$

for any credible sets $D(X)\subset\Theta$ of levels $1-\gamma$. Hence, if $U(\theta_0)$ contains θ_0, then the $C(X)$ satisfies,

$$P_{\theta_{0,n}}\big(\theta_0\in C(X)\big)\geq 1-\frac{\beta}{1-\gamma}.$$

The construction of confidence sets from credible sets then proceeds as follows: sample-/graph-size $n\geq 1$ is fixed and we assume that we have data X^n. Moreover, we have a desired confidence level $1-\alpha$ for the confidence set we are going to construct. Assume that we can show posterior concentration of the type (3) for certain $0<\beta\leq\alpha$. Choose a (minimal) credible level $0<\gamma<1$ such that[1] $\beta/(1-\gamma)\leq\alpha$. Then the sets $C(X^n)$ corresponding to credible sets $D(X^n)$ of levels $1-\gamma$ are exact confidence sets of confidence level $1-\alpha$. In the case (2) that the $U(\theta_n)$ are (closed) d_n-metric balls of (θ_n-independent) radii r in Θ_n, we find that the metric radius-r-enlargements,

$$C(X^n)=\{\theta_n\in\Theta_n : d_n(\theta_n,D(X^n))\leq r\},$$

of credible sets $D(X^n)$ are exact confidence sets of levels $1-\beta/(1-\gamma)\geq 1-\alpha$. (Below, we shall see examples of both radii that are, and radii that are not θ_n-independent.)

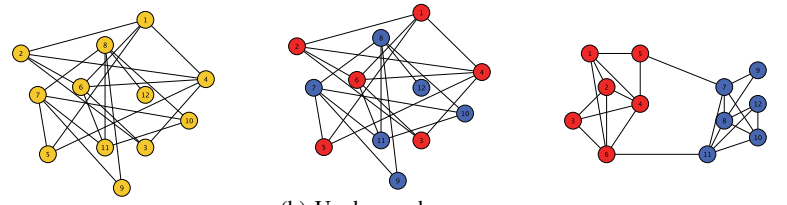
Coming back to the role of the posterior probability β, sharpness of the lower bound (3) is crucial for the construction of confidence sets from credible sets: if the bound is sharp, the required credible level is lowest, enlargement radii are lowest and the resulting confidence sets are smallest; versions of the bound that are too weak lead to required credible levels that are too high, enlargement radii that are too large and result in confidence sets that are too conservative, in that they cover the true value of the parameter with probability strictly above $1-\alpha$. To conclude, we note that the methodology sketched works for *any model/prior pair* for which the posterior displays concentration of the form (1). As noted, inequalities of this type have received a great deal of attention (Ghosal and van der Vaart, 2017) and the main challenge is to assure that known versions of the lower bounds for posterior concentration are *sharp*.

3 Application in sparse stochastic block models

The stochastic block model (Holland et al., 1983) is an inhomogeneous version of the Erdős-Rényi random graph (Erdős and Rényi, 1959) that is employed as one of the canonical models for the study of community structure in network science. The model and its generalizations have applications in physics, biology, sociology, image processing, genetics, medicine, logistics, *etcetera* (Fortunato, 2010; Abbe,

[1] If $\beta\geq\alpha$, there exist no solutions $0<\gamma<1$ that solve $\beta/\alpha\leq(1-\gamma)$. In that case no confidence set of confidence level α can be derived from credible sets.

2018). The observation is a graph X^n with n vertices that each belong to one of several communities, and edges that occur independently with probabilities that depend on the communities of the vertices they connect. For example, if the probability p of finding an edge between two vertices from the same community is much larger than the probability q of finding an edge between two vertices from different communities, then one expects high connectivity within and low connectivity between communities. Thinking of the random graph X^n as data and the community assignments of the vertices as unobserved, the statistical challenge is estimation of the vertices' true community assignments $\theta_{0,n} = (\theta_{0,n,1}, \ldots, \theta_{0,n,n})$ in the parameter space[2] $\Theta_n = \{0,1\}^n$, a task commonly referred to as *community detection* (Girvan and Newman, 2002) (see Figure 2). Parameter spaces are endowed with their so-called *Hamming distances* (all denoted k), which simply count the number of differing components between θ and θ'.



(a) Observation X^n (b) Unobserved communities (c) Community detection

Fig. 2 A realisation of the stochastic block graph X^n (Fig. 1(a)) with $n = 12$ vertices from two unobserved communities: vertices 1 through 6 belong to the red community and vertices 6 through 12 to the blue (Fig. 1(b)). Community detection (Fig. 1(c)) estimates the communities of Fig. 1(b), based on the presence or absence of edges in Fig. 1(a).

Aside from a large volume of interest in general estimation of parameters in the stochastic block model (for a recent overview, see (Abbe, 2018)), over the last decade there has also been great interest in asymptotic lower bounds for edge sparsity that leave consistent community detection (only just) possible as the graph size n grows. In (Decelle et al., 2011; Abbe et al., 2016; Massoulié, 2014; Mossel et al., 2016) and many other publications, asymptotic limitations on the estimation problem are studied in the context of the so-called *planted bi-section model*, which is a stochastic block model with two equally-sized communities of n vertices each and edge probabilities p_n (within communities) and q_n (between communities). Sharp conditions on edge probabilities for recovery of the communities have been established: for example, it is possible to recover the community assignments with a small, fixed fraction of mis-assignments asymptotically, if and only if (Decelle et al., 2011; Mossel et al., 2016),

[2] Strictly speaking, the equivalence relation $\sim$ that exchanges 0's for 1's and vice versa must be modded out, so $\Theta_n = \{0,1\}^n/\sim$ and k counts differences modulo $\sim$ as well. (See (Kleijn and van Waaij, 2021).)

$$\frac{n(p_n-q_n)^2}{2(p_n+q_n)} > 1, \tag{7}$$

for large values of n. A stronger, more precise form of consistency (so-called *exact recovery*) is possible, if and only if (Abbe et al., 2016; Abbe, 2018; Mossel et al., 2016),

$$\Big(\big(\sqrt{a_n}-\sqrt{b_n}\big)^2-2\Big)\log(n)+\log(\log(n))\to\infty, \tag{8}$$

(where $np_n = a_n\log(n)$ and $nq_n = b_n\log(n)$.)

Conditions for (1) in the planted bi-section model have been analysed in (Kleijn and van Waaij, 2018): it is shown that the posterior recovers community assignments consistently under sufficient conditions that are sharp *c.f.* (7) and (8). Here, we generalize to the full two-class stochastic block model, *i.e.* a graph X^n with n vertices from two communities without the restriction that the communities have equal sizes. In (Kleijn and van Waaij, 2021) it is shown that if we equip the spaces Θ_n of community assignments with uniform priors, then we have (inequality (1) with $k_n = 0$):

$$E_{\theta_{0,n}}\Pi\big(\{\theta_{0,n}\}\,\big|\,X^n\big) \geq 1-\tfrac{1}{2}n\rho(p_n,q_n)^{n/2}\,e^{n\rho(p_n,q_n)^{n/2}}, \tag{9}$$

where $\rho(p,q) = \sqrt{p}\sqrt{q}-\sqrt{1-p}\sqrt{1-q}$ denotes the Hellinger affinity between two Bernoulli experiments with probabilities $0\leq p,q\leq 1$. If,

$$n\rho(p_n,q_n)^{n/2}\to 0,$$

the posterior recovers the true community assignment exactly in the large-graph limit. This condition generalizes (8) to the case of unequally-sized communities and (like (8)) applies in cases where edge probabilities p_n and q_n are sparse enough to ensure that the expected degree of any vertex in the graph grows no faster than $\log(n)$ with growing graph size n. It is also shown in (Kleijn and van Waaij, 2021) that for any fractions $0<\lambda_n<1/2$,

$$\begin{aligned} E_{\theta_{0,n}}\Pi\big(k(\theta,\theta_{0,n})\leq\lambda_n n\,\big|\,X^n\big) \\ \geq 1-\frac{1}{2}\Big(e\,\lambda_n^{-1}\rho(p_n,q_n)^{n/2}\Big)^{\lambda_n n}\Big(1-e\,\lambda_n^{-1}\rho(p_n,q_n)^{n/2}\Big)^{-1}, \end{aligned} \tag{10}$$

which means that the posterior achieves almost-exact recovery at Hamming rate $k_n = \lambda_n n$, whenever,

$$\lambda_n n\Big(\log(\lambda_n)+\frac{n}{4}\big(\sqrt{p_n}-\sqrt{q_n}\big)^2-1\Big)\to\infty.$$

This last condition applies when edges are so sparse that the expected vertex degree remains finite in the limit. Analyses in (Kleijn and van Waaij, 2018, 2021) show that these conditions generalize the consistency conditions (7) and (8), applicable in the planted bi-section case, suggesting that the bounds (9) and (10) are also sharp.

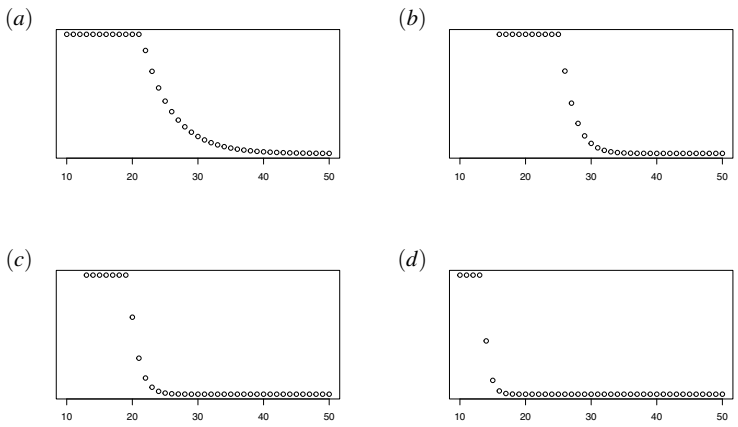


Fig. 3 Credible levels $1-\gamma$ required for a confidence set of confidence level 0.95 as a function of graph size $10 \leq n \leq 50$, with fixed edge probabilities $p = 0.9$ and $q = 0.1$ and various enlargement radii: in (a) we do not enlarge the credible set, in (b) – (d), we enlarge by Hamming radii $0.05n$, $0.10n$ and $0.25n$ respectively. (For the low graph sizes n where γ is missing or $\gamma = 0$, no confidence set of confidence level 0.95 can be derived from credible sets.) [From (Kleijn and van Waaij, 2021)]

To address our main goal, it is noted that corollary 1 can be used with the lower bounds (9) and (10) for the exact conversion of credible sets into confidence sets for any graph size: the method is visualized in Figure 3, which displays the credible levels $1-\gamma$ required for a credible set (or its enlargements by Hamming-radii $0.05n$, $0.10n$ and $0.25n$ respectively) to be interpretable as confidence sets of confidence level $1-\alpha = 0.95$, as a function of graph size n. In all graphs of Figure 3, there is a sharp drop in required credible level around a certain *critical graph size* $n(p,q;\alpha,\lambda)$ that depends on edge sparsity, desired confidence level and enlargement radius: when n passes the critical graph size, the frequentist starts to have confidence in community assignments of high posterior probability, not just in subsets of almost full posterior probability. Figure 3 offers a picture of the relationship between credible and confidence levels that is unexpected from the large-sample-approximate point of view (which would suggest equality or forms of proportionality).

These results on community detection in sparse graphs invite numerical verification and exploration: to begin with, the graphs shown in Figure 3 concern graph sizes in which Markov chain Monte Carlo simulations are feasible (McDaid et al., 2013; Geng et al., 2019; Jiang and Tokdar, 2021), so *cross-validation of the confidence levels* of (enlarged) credible sets with simulated data is not just possible, but highly desirable as a concrete verification of the main claim. Of course, the natural extension would be a demonstration of the new methods for uncertainty quantification in an application with real data, for example, clustering in protein-protein interaction networks.

A second computational direction is somewhat more speculative, but could be highly rewarding. Markov chain Monte Carlo methods have their limitations and scalability is one of them: when the dimension or cardinal of the parameter space is very high, samplers simply cannot be run long enough to generate a representative sample of the entire posterior. In network science, cardinals tend to grow very fast with increasing graph size and MCMC sampling in the two-community stochastic block model does not stay feasible for graph sizes much beyond those of Figure 3. There are, however, reasons to suspect that even undersampled posteriors are useful for our purposes: samples that that are too small tend to under-represent primarily the tails of a posterior and not so much its bulk. For the construction of credible sets we need specifically those parameter values with high posterior probability, that is, parameter values from the bulk. So it appears worthwhile to explore the possibility that some form of *early stopping* of MCMC samplers may still produce credible sets that are useful for frequentist uncertainty quantification. To analyse questions regarding the validity and limitations of this argument, one could start by cross-validating the confidence levels of enlarged credible sets with simulated data and undersampled posteriors. The possible upshot is construction of confidence sets from credible sets for *graphs that are (much?) larger* than common limitations of MCMC sampling suggest.

4 Application in population genetics: trace reconstruction

Trace reconstruction (Thornton, 2004) originates from population genetics: when similar-but-not-equal strands of genetic code are found in a variety of present-day species or sub-species and it is assumed that they all originate from a single, original version of the strand in some common ancestor, what did that original strand of genetic code look like? The underlying statistical modelling assumption is, that all observed present-day versions of the strand are the result of a (simplified version of the) random process by which genetic code deteriorates over time, *e.g.* by random deletions, insertions and mutations of bits of code.

In its simplest form the question is reduced to describe an original strand $\theta_{0,n} \in \Theta_n = \{0,1\}^n$ with n bits of *binary* information from which N so-called *traces* are drawn independently, through a *deletion channel*: for every trace, bits from $\theta_{0,n}$ are deleted independently with some probability $0 < p < 1$. Deletions take place *without leaving empty spots*, so every random deletion shortens the trace's length by one and traces have lengths that are distributed $\mathrm{Bin}(n,p)$. The trace reconstruction problem is inhomogeneous, in that accuracy of estimation varies strongly with the true $\theta_{0,n} \in \Theta_n$: intuitively it is clear that an original strand of the form $\{0,\ldots,0\}$ is easily identified from only a few traces, while strand of more diverse nature require many more trace data to reconstruct.

Preliminary explorations of the problem indicate the following: let $F_n(X_n,\theta)$ denote the square root of the number of possible deletions d in D_{n,n_j} that match the observed X_n (of length n_j) to a trace of θ. With a uniform prior for Θ_n, the expected

posterior probability for the Hamming ball $V_{n,m}$ of radius $m \geq 0$ is upper bounded by:

$$\mathrm{E}_{\theta_{0,n}} \Pi\big(V_{n,m} \,\big|\, X^n\big) \geq 1 - \sum_{\theta \notin V_{n,m}} \alpha_n(\theta_0, \theta)^N, \text{ with } \alpha_n(\theta_0, \theta) = \mathrm{E}_{n,\theta_0}\left(\frac{F_n(X_n, \theta)}{F_n(X_n, \theta_0)}\right), \tag{11}$$

Lower bounds for the right-hand side of (11) can be surmised from the work of probabilists, who have analysed the minimal number N of traces needed to reconstruct an original strand of (large) length n correctly with high probability: the main result is (Nazarov and Peres, 2017) which requires N of order $\exp(O(n^{1/3}))$, and (Peres and Zhai, 2017; Holden and Lyons, 2018) look at particular cases and average N needed with unknown original binary sequence.

Inequality (11) or subsequent lower bounds for the right-hand side can be combined with Lemma 1 in the same way that the bounds (9) and (10) were, to construct confidence sets as enlarged credible sets. What is different in this analysis, is the inhomogeneous dependence of the sets $U(\theta)$ on the parameter θ: while community detection is equally difficult for any community assignment, trace reconstruction is highly dependent on the true $\theta_{0,n}$ and this will be reflected in the resulting confidence sets (or in conditions on $\theta_{0,n}$).

Acknowledgements The author thanks J. van Waaij for a collaboration on frequentist uncertainty quantification in the stochastic block model,

References

L. Le Cam and G. Yang, "Asymptotics in statistics: Some basic concepts", Springer New York, 2000.

D. Freedman, "On the Bernstein-von Mises theorem with infinite-dimensional parameters", *The Annals of Statistics* **27** (1999), no. 4, 1119–1140.

I. Castillo, "Semiparametric Bernstein-von Mises theorem and bias, illustrated with Gaussian process priors", *Sankhya* **74** (2012)a, no. 2, 194–221.

I. Castillo, "A semiparametric Bernstein-von Mises theorem for Gaussian process priors", *Probab. Theory Relat. Fields*, 2012b, no. 152, 53—-99.

P. Bickel and B. Kleijn, "The semi-parametric Bernstein-von Mises theorem", *Annals of Statistics* **40** (2012), no. 1, 206–237.

I. Castillo and R. Nickl, "Nonparametric Bernstein-von Mises theorems in Gaussian white noise", *Annals of Statistics* **41** (2013), no. 4, 1999–2028.

B. Szabó, A. van der Vaart, and J. van Zanten, "Frequentist coverage of adaptive nonparametric Bayesian credible sets", *Annals of Statistics* **43** (2015), no. 4, 1391–1428.

B. Kleijn, "Frequentist validity of Bayesian limits", *Annals of Statistics* **49** (2021), no. 1, 182–202.

S. Ghosal and A. van der Vaart, "Fundamentals of nonparametric Bayesian inference", Cambridge University Press, 2017.
P. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps", *Social Networks* **5** (1983), no. 2, 109–137.
P. Erdős and A. Rényi, "On random graphs i", *Publicationes Mathematicae*, 1959.
S. Fortunato, "Community detection in graphs", *Physics Reports* **486** (2010), no. 3, 75–174.
E. Abbe, "Community detection and stochastic block models: Recent developments", *Journal of Machine Learning Research* **18** (2018), no. 177, 1–86.
B. Kleijn and J. van Waaij, "Confidence sets in a sparse stochastic block model with two communities of unknown sizes", arXiv:2108.07078.
M. Girvan and M. Newman, "Community structure in social and biological networks", *Proc. Nat. Acad. Sc.* **99** (2002), no. 12, 7821–7826.
A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications", *Phys. Rev. E* **84** (2011) 066106.
E. Abbe, A. Bandeira, and G. Hall, "Exact recovery in the stochastic block model", *IEEE: Transactions on Information Theory* **62** (2016), no. 1,.
L. Massoulié, "Community detection thresholds and the weak ramanujan property", in "Proc. 46th Symposium on the Theory of Computing", pp. 1–10. 2014.
E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model", *Electron. J. Probab.* **21** (2016).
B. Kleijn and J. van Waaij, "Recovery, detection and confidence sets of communities in a sparse stochastic block model", arXiv:1810.09533.
A. McDaid, T. Murphy, N. Friel, and N. Hurley, "Improved Bayesian inference for the stochastic block model with application to large networks", *CSDA* **60** (2013), no. C, 12–31.
J. Geng, A. Bhattacharya, and D. Pati, "Probabilistic community detection with unknown number of communities", *JASA* **114** (2019), no. 526, 893–905.
S. Jiang and S. Tokdar, "Consistent Bayesian community detection", arXiv:2101.06531.
J. Thornton, "Resurrecting ancient genes: experimental analysis of extinct molecules.", *Nat. Rev. Genet.* **5** (2004) 366—-375.
F. Nazarov and Y. Peres, "Trace reconstruction with $exp(o(n^{1/3}))$ samples", in "Proc. 49th ACM SIGACT Symposium on Theory of Computing", pp. 1042—-1046. 2017.
Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice", *IEEE 58th Symposium on Foundations of Computer Science*, 2017.
N. Holden and R. Lyons, "Lower bounds for trace reconstruction", arXiv:1808.02336.

Empirical Bayesian analysis of componentwise maxima in multivariate samples

Analisi empirico-bayesiana dei massimi a componenti in campioni multivariati

Simone A. Padoan and Stefano Rizzelli

Abstract Statistical theory and methods for the analysis of maxima, computed componentwise in a multivariate sample, has been an active research area in the last decade. Under mild assumptions, extreme-value theory justifies modelling random vectors of linearly normalized sample maxima by multivariate max-stable distributions. Various proposals for Bayesian inferential procedures have been formulated in recent years, though they typically disregard the asymptotic bias inherent in the use of max-stable models, incorporating no information on norming sequences in prior specifications for scale and location parameters. The semiparametric empirical Bayesian approach in [6] suitably addresses this point via data-dependent priors. In this contribution we review its consistency properties.

Abstract *Lo sviluppo di metodi e strumenti teorici per l'analisi statistica dei massimi, calcolati per ciascuna delle componenti di un campione multivariato, ha rappresentato un importante tema di ricerca negli ultimi dieci anni. La teoria dei valori estremi giustifica la modellizzazione (di una trasformazione lineare) dei massimi campionari con distribuzioni max-stabili multivariate. Le procedure bayesiane proposte negli ultimi anni non tengono conto della distorsione insita nell'uso di modelli max-stabili, non incorporando alcuna informazione sulle costanti di normalizzazione nelle distribuzioni a priori. L'approccio semi-parametrico empirico-bayesiano in [6] affronta adeguatamente questo punto, attraverso pseudo distribuzioni a priori i cui parametri dipendendono dai dati. In questo contributo riassiumiamo le sue proprietà di consistenza.*

Key words: Extreme-value copula, Multivariate max-stable distribution, Semiparametric estimation, Angular measure, Posterior consistency.

Simone A. Padoan
Department of Decision Sciences, Bocconi University, e-mail: simone.padoan@unibocconi.it

Stefano Rizzelli
Department of Statistical Sciences, Catholic University, e-mail: stefano.rizzelli@unicatt.it

1 Introduction

In the last two decades, two main approaches to the statistical analysis of multivariate extremes have emerged. In the first one, tail features of the data generating distribution are inferred by analysing maxima, computed componentwise over data subsets (blocks). In the second one, the tail behaviour of a set of variables of interest is studied by analysing multivariate peaks over a threshold. See, e.g., §1 of [4] for a discussion. In this contribution, we focus on the former approach.

Whenever data can be sensibly considered as realisations of independent and identically distributed (i.i.d.) random vectors (r.v.s), a typical inferential routine is:

- Step 1: split the data sample into blocks of equal length;
- Step 2: over each block, compute the maximum for each variable;
- Step 3: fit a multivariate max-stable model to the so-obtained sample of maxima.

In the above methodological scheme, Step 3 is grounded on multivariate extreme-value theory, according to which limiting distributions of linearly normalised maxima (when they exist) are max-stable. In practice, however, max-stable distributions constitute a misspecified model class for maxima. According to statistical folkore in extreme-value analyis, the estimation bias due to model misspecification is modereate whenever maxima are computed over sufficiently large blocks of observations. Whereas this common belief has multiple theoretical validations for frequentist inference (e.g., [3]), rigorous mathematical results for Bayesian inference have been established only very recently by the authors of this note, in [6]. In this contribution, we review the semiparametric framework proposed in our earlier work.

The theoretical results established in [6] are complex, requiring a supplemental material document of significant length for a proper and formal argumentation. Therefore, in the reminder of this note, we do not aim at expounding our asymptotic theory, but rather at offering a concise and simplified account. As a preliminary step, in §2 and §3, we provide the necessary background on limiting behaviour of maxima and on the max-stable distributions used for statistical modelling of the latter. In §4 we describe the data-dependent prior formulation for the finite and infinite dimensional components of the considered max-stable models, providing new concrete examples of prior specifications. This empirical Bayes approach plays a crucial role in compensating for the lack of knowledge *a priori* about the tail features of observables, from a practical viewpoint, and in guaranteeing consistency of posterior inferences, from a theoretical viewpoint. The main findings in this respect are summarised in §6, containing also a short discussion of their important consequences for statistical prediction of future extremes.

2 Limiting distributions for maxima

Let $\mathbb{N}_+ = \{1,2,\ldots\}$ be the set of positive integers and, for a given $d \in \mathbb{N}_+ \setminus \{1\}$ and any $\lambda \in \mathbb{R}$, set $[d] = \{1,\ldots,d\}$ and $\lambda_d := (\lambda,\ldots,\lambda) \in \mathbb{R}^d$. In what follows,

the computation of maxima as well as sums, products, etc. involving vectors in $\mathbb{R}^d$ operate component by component. Let $X_i, i = 1,2,\ldots$ be i.i.d. d-dimensional r.v.s with distribution F. If there exist norming sequences $a_m \in (0,\infty)^d$ and $b_m \in \mathbb{R}^d$, $m = 1,2,\ldots$ and a distribution G with nondegenerate margins such that

$$\lim_{m\to\infty} F^m(a_m x + b_m) = G(x), \tag{1}$$

for all continuity points x of G, we say that G is a multivariate extreme-value distribution and that F is in its max-domain of attraction, in symbols $F \in \mathscr{D}(G)$. By Fisher-Tippett-Gnedenko theorem, (e.g., Proposition 0.3 in [8]), for $j \in [d]$ the marginal distribution G_j is of one of the following types

$$G_j(x) = \begin{cases} \exp(-x^{-\rho_j}), & x > 0, \rho_j > 0, \\ \exp(-\exp(-x)), & x \in \mathbb{R}, \\ \exp(-(-x)^{\omega_j}), & x < 0, \omega_j > 0, \end{cases}$$

known as Fréchet, Gumbel and (reverse) Weibull distributions, up to location and scale parameters. The distribution G_j is of the first, second or third type depending on the weight of tha tail of F_j ([8], Ch. 1.1–1.3).

As for the dependence structure of G, its copula function C_η belongs to the class of extreme-value copulas (e.g., [1], p. 272) and is fully characterised by a probability measure η on the unit simplex $\mathscr{S} := \{w \in [0,1]^d : \|w\|_1 = 1\}$ via the relation

$$C_\eta(u) = \exp\left(-d\int_{\mathscr{S}} \max\{(-\ln u_d)w_1, (-\ln u_1)w_2, \ldots, (-\ln u_{d-1})w_d\}\mathrm{d}\eta(w)\right),$$

for all $u \in (0,1]^d$. The measure η is commonly called angular (or spectral) probability measure and satisfies the mean constraints $\int_{\mathscr{S}} w_j \eta(\mathrm{d}w) = 1/d$, for $j \in [d]$. It governs dependence strenght in the tails of F, as can be intuitively deduced, e.g., from the first order approximations to the probability of simultaneous exceedances for all the components of the i-th r.v. X_i

$$\mathbb{P}\left(\cap_{j=1}^d \{X_{i,j} > F_j^{\leftarrow}(e^{-1/t})\}\right) \sim dt^{-1}\int_{\mathscr{S}} \min(w_1,\ldots,w_d)\mathrm{d}\eta(w),$$

where $t \to \infty$ and $F_j^{\leftarrow}(e^{-1/t})$ is an increasingly large quantile of the j-th margin of F, for $j \in [d]$. In particular, the more η is concentrated in proximity of the center of the simplex $(1/d)_d$, the higher is the probability of simultaneous peaks above large levels in all the components of X_i.

3 Max-stable models

The class of location/scale multivariate extreme-value distributions arising from (1) is precisely the class of max-stable distributions with nondegenerate margins (e.g.,

Proposition 5.9 in [8]). Hence, the asymptotic distribution of the r.v. of linearly normalised componentwise maxima must satisfy the max-stability property $G^t(x) = G(\alpha(t)x+\beta(t))$, for some functions $\alpha : (0,\infty) \mapsto (0,\infty)^d$ and $\beta : (0,\infty) \mapsto \mathbb{R}^d$, and all $t > 0$. In turn, the class of extreme-value copulas coincides with that of max-stable copulas (e.g., [1], p. 273), therefore we must have

$$C_\eta(u) = C_\eta(u_1^{1/k},\ldots,u_d^{1/k})^k,$$

for all $u \in [0,1]^d$ and $k \in \mathbb{N}_+$. Extreme-value copulas cannot be fully characterised using parametric families (e.g., [1], Ch. 9.2), as they are indexed to the angular probability measure η, ranging over an infinite-dimensional space. This underpins the semiparametric nature of the multivariate max-stable distributions class.

Angular probability measures can be fairly complex objects, as they can place mass on all the the subspaces of the simplex $\{w \in \mathscr{S} : w_j > 0, \forall j \in I\}$, with $I \subset [d]$, including singletons which only contain the j-th vertex e_j, $j \in [d]$. To simplify statistical inference, a common approach is to consider subfamilies of measures which only place mass on specific subspaces (e.g., [9]). Hereafter, we focus on the class $\mathscr{E}$ of angular probability mesures having null mass outside the subset

$$\{\mathscr{S} \cap (0,1)^d\} \cup \{e_1\} \cup \cdots \cup \{e_d\}$$

and absolutely continuous restriction to $\mathscr{S} \cap (0,1)^d$. Such angular probability measures can be characterised in the following terms. Define the polytope $\mathscr{R} := \{v \in [0,1]^{d-1} : \|v\|_1 \leq 1\}$ and the projection map $P : \mathscr{S} \mapsto \mathscr{R} : (w_1,\ldots,w_{d-1},w_d) \mapsto (w_1,\ldots,w_{d-1})$. Then, for any $\eta \in \mathscr{E}$, there exist masses $p_j \in [0,1/d]$, $j = 1,\ldots,d$, and an integrable nonnegative function $h : \mathring{\mathscr{R}} \mapsto [0,\infty)$ allowing the representation

$$\eta(B) = \sum_{j=1}^{d} p_j \delta_{e_j}(B) + \int_{P(B\cap(0,1)^d)} h(v)\mathrm{d}v,$$

for all Borel sets $B \subset \mathscr{S}$; see also [5], pp. 3313–3314, for an equivalent representation in the bivariate case $d = 2$. The function h is referred to as angular density ([6], Def. 2.2). Overall, the class $\mathscr{E}$ results into a flexible nonparametric family of extreme-value copulas.

To study asymptotic properties of statistical inference on multivariate max-stable models, for technical reasons, we further restrict our attention to classes of scale or location-scale multivariate max-stable distributions with homogeneous margins, firstly introduced in Definition 2.1 of [6], conveniently constraining shape parameters to be larger than 1 for Weibull margins (see [6], §4.2).

Definition 1. We refer to the family of multivariate max-stable distributions $G_{\theta,\eta}(x) = C_\eta(G_{\theta_1}(x_1),\ldots,G_{\theta_d}(x_d))$ as:

- *multivariate ρ-Fréchet*, when $G_{\theta_j}(x_j) = \exp(-(x_j/\sigma_j)^{-\rho_j})$, with $\theta_j = (\rho_j,\sigma_j) \in (0,\infty)^2$, $x_j > 0$, for all $j \in [d]$;

- *multivariate Gumbel*, when $G_{\theta_j}(x_j) = \exp(-\exp(-(x-\mu_j)/\sigma_j))$, with $\theta_j = (\sigma_j, \mu_j) \in (0,\infty) \times \mathbb{R}$, $x_j \in \mathbb{R}$, for all $j \in [d]$;
- *multivariate ω-Weibull*, when $G_{\theta_j}(x_j) = \exp(-(-(x-\mu_j)/\sigma_j)^{\omega_j})$, with $\theta_j = (\omega_j, \sigma_j, \mu_j) \in (1,\infty) \times (0,\infty) \times \mathbb{R}$, $x_j < \mu_j$, for all $j \in [d]$.

In the three cases, the marginal parameters and their spaces are $\theta = (\rho, \sigma) \in \Theta = (0,\infty)^{2d}$, $\theta = (\sigma, \mu) \in \Theta = (0,\infty)^d \times \mathbb{R}^d$ and $\theta = (\omega, \sigma, \mu) \in \Theta = (1,\infty)^d \times (0,\infty)^d \times \mathbb{R}^d$, respectively. We denote the probability density of $G_{\theta,\eta}$ by $g_{\theta,\eta}$.

In the reminder of this note, the model $\mathscr{G} := \{G_{\theta,\eta} : \theta \in \Theta, \eta \in \mathscr{E}\}$, resulting from one of the three multivariate distribution classes of Definition 1, serves as an approximate observational model for the empirical Bayesian analysis of maxima. We assume that n maxima $M_{m,i} = \max(X_{(i-1)m+1}, \ldots, X_{im})$, $i = 1, \ldots, n$, are computed componentwise from disjoint blocks of m r.v.s, extracted from a simple random sample $X_1, \ldots, X_{nm}$, with unkown distribution F_0. The distribution of componetwise maxima F_0^m is allowed to lie outside the max-stable model class $\mathscr{G}$, which is thus misspecified for any finite block-size m. On the other hand, to the purpose of asymptotics, we assume that the block-size m gets larger as n increases, i.e. $m \equiv m(n) \to \infty$ as $n \to \infty$. Moreover, we postulate that Condition 4.1 and 4.4 in [6] are satisfied, compactly denoting the whole set of assumptions on the data-generating mechanism by the acronym "DGM". Thus, as the block-size m increases, we have that the density of unnormalised maxima $f_m(x) := (\partial/\partial x)F_0^m(x)$ merges in Hellinger distance with the sequence of max-stable densities $g_{\eta_0,\theta_{0,m}}$, where the parameter $\theta_{0,m}$ equals (ρ_0, a_m) or (a_m, b_m) or (ω_0, a_m, b_m) in the Fréchet, Gumbel and Weibull limiting cases, respectively. In what follows, this fact guarantees that, when using the appropriate distribution class in Definition 1, the density estimation bias due to model misspecification is asymptotically negligible.

4 Data-dependent prior

In this section, we present the general lines of our prior specification. We assume that the prior on the angular measure η and the marginal distributions parameter θ are specified independently. We postulate a Bernstein polynomial prior Π on η, constructed as in Condition 3.8 of [6]. A concrete example is as follows.

Example 1. For any integer $k > d$, consider the representation

$$\eta_k(B) = \sum_{j=1}^{d} \gamma_{ke_j} \delta_{e_j}(B) + \int_{P(B \cap (0,1)^d)} \sum_{q \in \mathscr{Q}_k} \gamma_q \{(k-1)!/(k-d)!\} b_q(v) \mathrm{d}v,$$

for all Borel set $B \subset \mathscr{S}$, where $\mathscr{Q}_k$ is the set of multi-indices

$$\mathscr{Q}_k := \{q \in [k-d+1]^d : \|q\|_1 = k\},$$

$\{\gamma_{ke_1},\dots,\gamma_{ke_d},\gamma_q, q\in\mathscr{Q}_k\}$ are nonnegative weights summing up to 1 and, for all $q\in\mathscr{Q}_k$, b_q is the Bernstein polynomial basis function

$$b_q(v)=(k-d)!\prod_{j=1}^{d-1}\frac{v_j^{q_j-1}}{(q_j-1)!}\frac{(1-\sum_{j=1}^{d-1}v_j)^{q_d-1}}{(q_d-1)!},\quad v\in\mathring{\mathscr{R}}.$$

Then, a prior for the angular probability measure with full support on the space $\mathscr{E}$ can be specified in the following hierarchical way, using the above representation. For any integer $k>d$, set

$$\begin{aligned}
&w_k:=\{w_q, q\in\mathscr{Q}_k\}\sim\text{Dirichlet}(\alpha_k)\\
&E_k|w_k\sim\text{Uniform}\left(0,\min\left\{1,\left[\max_{1\le j\le d}\sum_{l=1}^{k-1}dk^{-1}\sum_{q\in\mathscr{Q}_k:q_j=l}w_q\right]^{-1}\right\}\right)\\
&\{\gamma_q, q\in\mathscr{Q}_k\}|E_k,w_k\sim\delta_{E_kw_k}\\
&\gamma_{ke_j}|E_k,w_k\sim\delta_{d^{-1}-\sum_{l=1}^{k-1}\sum_{q\in\mathscr{Q}_k:q_j=l}E_kw_k},\quad\forall j\in[d]
\end{aligned}$$

for an arbitrary $\alpha_k\in(0,\infty)^{\text{card}(\mathscr{Q}_k)}$; finally, let k be a discrete Weibull random variable with unit scale and shape parameter $d-1$, truncated outside $\mathbb{N}_+\setminus[d]$.

In priciple, prior specification for marginal scale an location parameters should be based on knowledge about the norming sequences a_m and b_m, available before observing the data. In practice, typical lack of prior knowledge can be bypassed resorting to a data-dependent prior distribution. Hence, we consider a data dependent prior density sequence ψ_n for θ of the following general form:

$$\psi_n(\theta)=\begin{cases}
\pi_{\text{sh}}(\rho)\times\prod_{j=1}^d\pi_{\text{sc}}\left(\frac{\sigma_j}{\widehat{a}_{m,j}}\right)\frac{\mathrm{d}\sigma_j}{\widehat{a}_{m,j}},\\
\pi_{\text{sh}}(\omega)\times\prod_{j=1}^d\pi_{\text{sc}}\left(\frac{\sigma_{n,j}}{\widehat{a}_{m,j}}\right)\frac{\mathrm{d}\sigma_j}{\widehat{a}_{m,j}}\times\prod_{j=1}^d\pi_{\text{loc}}\left(\frac{\mu_j-\widehat{b}_{m,j}}{\widehat{a}_{m,j}}\right)\frac{\mathrm{d}\mu_j}{\widehat{a}_{m,j}},\\
\prod_{j=1}^d\pi_{\text{sc}}\left(\frac{\sigma_j}{\widehat{a}_{m,j}}\right)\frac{\mathrm{d}\sigma_j}{\widehat{a}_{m,j}}\times\prod_{j=1}^d\pi_{\text{loc}}\left(\frac{\mu_j-\widehat{b}_{m,j}}{\widehat{a}_{m,j}}\right)\frac{\mathrm{d}\mu_j}{\widehat{a}_{m,j}},
\end{cases}$$

for the three limit classes ρ-Fréchet, ω-Weibull, Gumbel respectively, where $\widehat{a}_m=(\widehat{a}_{m,1},\dots,\widehat{a}_{m,d})$ and $\widehat{b}_m=(\widehat{b}_{m,1},\dots,\widehat{b}_{m,d})$ are estimators of a_m and b_m with good asymptotic behaviour, while π_{sh}, π_{sc} and π_{loc} are smooth Lebesgue probability densities. A precise enumeration of the properties they are assumed to comply with can be found in Condition 4.6, 4.14 and 4.18 of [6] for the Fréchet, Gumbel and Weibull modelling setup, respectively.

Example 2. In the ρ-Fréchet modelling setup, for positive κ,λ,τ, examples of valid choices of π_{sh} and π_{sc} are the product of d half-Gaussian densities

$$\pi_{\text{sh}}(\rho)=\prod_{j=1}^d\frac{\sqrt{2}}{\tau\sqrt{\pi}}e^{-\rho_j^2/2\tau}$$

and the (positive) Weibull probability density $\pi_{\mathrm{sc}_j}(\sigma) = (\kappa/\lambda)(\sigma/\lambda)^{\kappa-1}e^{-(\sigma/\lambda)^{\kappa}}$. In the Gumbel modelling setup, the latter choice of π_{sc} can be paired with a Gaussian selection $\pi_{\mathrm{loc}}(\mu) = (2\pi\varsigma)^{-1/2}e^{-(\mu-\zeta)^2/2\varsigma}$, $\zeta \in \mathbb{R}, \varsigma > 0$. Finally, in the ω-Weibull modelling setup, valid choices of $\pi_{\mathrm{sh}}, \pi_{\mathrm{sc}}, \pi_{\mathrm{loc}}$ can be derived by truncating the previous ones outside suitably large compact sets.

The overall set of assumptions on a data-dependent prior specification for (θ, η) is hereafter compactly denoted by the acronym "DDP".

5 Consistency

Once one of the three distribution families in Definition 1 is chosen for statistical inference, the data dependent prior $\Psi_n \times \Pi$ and the likelihood

$$\mathscr{L}_n(\theta, \eta) = \prod_{i=1}^{n} g_{\theta,\eta}(M_{m,i}), \quad (\theta, \eta) \in \Theta \times \mathscr{E},$$

give rise to the empirical Bayes posterior defined via

$$\Pi_n(B) := \frac{\int_B \mathscr{L}_n(\theta, \eta)\mathrm{d}(\Psi_n \times \Pi)(\theta, \eta)}{\int_{\Theta\times\Theta} \mathscr{L}_n(\theta, \eta)\mathrm{d}(\Psi_n \times \Pi)(\theta, \eta)},$$

for all $\Psi_n \times \Pi$-measurable sets B; see aso [7] for an account on emperical Bayes methods. In turn, Π_n induces a posterior distribution $\tilde{\Pi}_n$ over the class of max-stable densities $\{g_{\theta,\eta}, \theta \in \Theta, \eta \in \mathscr{E}\}$. The corresponding posterior predictive density is $\hat{g}_n(x) = \int_{\mathscr{G}} g(x)\mathrm{d}\tilde{\Pi}_n(g)$. In what follows we provide a posterior consistency theorem which unifies the results in Theorems 4.7, 4.15, 4.19 and Corollary 5.2 in [6]. To study posterior concentration for the finite dimensional parameter, we make use of the following notion of neighbourhoods of $\theta_{0,m}$:

$$B_n = \begin{cases} \{(\rho, \sigma) : \|\rho - \rho_0\|_1 + \|\sigma/a_m - \mathbf{1}\|_1 < \varepsilon\} \\ \{(\sigma, \mu) : \|\sigma/a_m - \mathbf{1}\|_1 + \|(\mu - b_m)/a_m\|_1 < \varepsilon\} \\ \{(\omega, \sigma, \mu) : \|\omega - \omega_0\|_1 + \|\sigma/a_m - \mathbf{1}\|_1 + \|(\mu - b_m)/a_m\|_1 < \varepsilon\} \end{cases}$$

in the ρ-Fréchet, Gumbel and ω-Weibull cases, respectively, with $\varepsilon > 0$.

Theorem 1. *Let $M_{m,1}, \ldots, M_{m,n}$ be i.i.d. according to F_0^m. Let F_0 and $G_{\theta_{0,m},\eta_0}$ comply with DGM. Then, under DDP, as $n \to \infty$:*

i. *for every ball in Lévy-Prohorov metric L with center η_0 and radius $\varepsilon > 0$ and every sequence of neighbourhood B_n of $\theta_{0,m}$*

$$\Pi_n(\{B_n \times L\}^c) = o_p(1);$$

ii. *for every sequence of Hellinger balls H_n with radius $\varepsilon > 0$ and center $g_{\theta_{0,m},\eta_0}$*

$$\tilde{\Pi}_n(H_n^c) = o_p(1);$$

iii. *the Hellinger distance between* $\hat{g}_n$ *and* f_m *converges to* 0 *in probability.*

In simple terms, the results at point i.-ii. of Theorem 2 guarantee that the posterior distributions $\tilde{\Pi}_n$ and Π_n concentrate around the max-stable density $g_{\theta_{0,m},\eta_0}$, which approximates the true density of maxima, and the corresponding parameter $\theta_{0,m}$. We point out that stronger asymptotic results, valid in almost sure terms, can be obtained in the Fréchet and Gumbel modelling framework (see Theorems 4.7 and 4.15 in [6]). The result at point iii. of Theorem 2 has important upshots for statistical prediction. As claimed by George and Xu in [2]: "Of the many possible forms a prediction can take, the richest is a predictive density, a probability distribution over all possible outcomes. Such a comprehensive description of future uncertainty opens the door to sharper risk assessment and better decision making." This is especially true in the case of prediction of future extreme quantities, such as the next maximum levels of an observable process. In the (empirical) Bayesian approach, a natural estimator of the true, unkown predictive density of future observations is the posterior predictive density. Theorem 2.iii establishes consistency of the latter as an estimator of f_m under a sufficiently strong metric to guarantee that, as $n \to \infty$,

$$\sup_B |\hat{G}_n(B) - F_0^m(B)| = o_p(1), \tag{2}$$

where the supremum ranges over Borel subsets B of $\mathbb{R}$ and $\hat{G}_n(B) = \int_B \hat{g}_n(x)\mathrm{d}x$ denotes the probability measure associated to the posterior predictive distribution. Intuitively, (2) warrants that draws from $\hat{G}_n$ are representative of the behaviour of a future r.v. of maxima over a block of size m, provided that m and n are large enough. This property is particularly attractive from the point of view of practical implementation of statistical prediction, using MCMC methods to sample from $\hat{G}_n$.

References

1. Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons Ltd., Chichester.
2. George, E. I. and Xu, X. (2010). Bayesian predictive density estimation. *Frontiers of Statistical Decision Making and Bayesian Analysis. In honor of James O. Berger*, 83-.95. Springer.
3. Bücher, A. and Segers, J. (2014). Extreme value copula estimation based on block maxima of a multivariate stationary time series *Extremes*, **17**, 495– 528.
4. Bücher, A., Volgushev, S. and Zou, N. (2019). On second order conditions in the multivariate block maxima and peak over threshold method. *J. Multivar. Anal.*, **173**, 604–619.
5. Marcon, G., Padoan, S. A. and Antoniano-Villalobos, I. (2016). Bayesian inference for the extremal dependence. *Electron. J. Stat.*, **10**, 3310–3337.
6. Padoan, S. A. and Rizzelli, S. (2021). Consistency of Bayesian inference for multivariate max-stable distributions. *Ann. Statist.* (in print).
7. Petrone, S., Rizzelli, S., Rousseau, J. and Scricciolo, C. (2014). Empirical Bayes methods in classical and Bayesian inference. *METRON*, **72**, 201–215.
8. Resnick, S. I. (2008). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

9. Sabourin, A. and Naveau, P. (2014). Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Comput. Statist. Data Anal.*, **71**, 542–567.

Processing of textual data in large corpora

Predictive performance comparisons of different feature extraction methods in a financial column corpus

Confronto della capacità predittiva di diversi metodi di estrazione delle variabili dal corpus di una rubrica finanziaria

Andrea Sciandra and Riccardo Ferretti

Abstract This work concerns the processing of a corpus made up of a financial weekly column. Specifically, we focused on document-level index extraction and textual feature extraction. Moreover, some feature extraction methods had been compared to evaluate their predictive capacity. Results confirm the hypothesis that vectors derived from word embedding do not improve the predictive power compared to other feature extraction methods but remain a fundamental resource for capturing semantics in texts.

Abstract *Questo contributo riguarda il trattamento di un corpus costituito da una rubrica finanziaria settimanale. In particolare, ci siamo concentrati sull'estrazione di indici a livello di documento e sull'estrazione di variabili testuali. Inoltre, abbiamo confrontato alcuni metodi di estrazione delle variabili per valutare la loro capacità predittiva. I risultati confermano l'ipotesi che i vettori derivati dal word embedding non migliorano la capacità predittiva rispetto ad altri metodi di estrazione delle variabili, ma restano una risorsa fondamentale per cogliere la semantica nei testi.*

Key words: behavioural finance, sentiment analysis, lexical complexity, feature extraction, word embedding, principal component regression

---

[1] Andrea Sciandra, Department of Communication and Economics, University of Modena and Reggio Emilia; email: andrea.sciandra@unimore.it

Riccardo Ferretti, Department of Communication and Economics, University of Modena and Reggio Emilia; email: riccardo.ferretti@unimore.it

This research has been partially supported by the University fund for Research (FAR 2020) of the University of Modena and Reggio Emilia.

1 Introduction

This work focuses on a financial column, named 'Letter to investor' (*Lettera all'investitore*), which has been published on the Sunday edition of the leading Italian financial newspaper (Il Sole 24 Ore). The column analyses every week an Italian stock through second-hand news, reporting balance sheet and income statement data, managers' outlook, stock's past performance, and, in some cases, analysts' recommendations. In previous research, we showed how the mechanism of attention grabbing (AGH) is at work. According to AGH, stale information published in print media can lead retail investors to buy stocks that grab their attention [1] to the extent that past analysts' recommendations may induce abnormal movements in stock prices and returns. Cervellati et al. [2] and Ferretti and Sciandra [3] showed that the publication of articles concerning single listed companies' profiles and financial analysts' recommendations are followed by an asymmetric reaction of stock prices. More precisely, they find a statistically significant stock price increase when the recommendation is positive (overweight or buy) and a substantial stationarity when the recommendation is not positive (hold or underweight or sell). In a more recent work, Ferretti and Sciandra [4] pointed out how the absence of explicit recommendations (approximately from 2015) in the same column, calls into question the role of the article sentiment, as they showed how investors transform articles content into implicit recommendations that, when highly positive, can direct their buying decisions. This result explained the importance of the textual analysis of this corpus, which will be deepened in this work in terms of text processing, textual feature extraction, and summary indexes especially related to the polarity and to the lexical complexity of the articles. Moreover, some feature extraction methods will be compared to evaluate their predictive capacity. The underlying hypothesis, based on previous research from other fields (especially social media [12]), is that vectors derived from word embedding do not improve the predictive ability compared to other feature extraction methods. The prediction targets are the abnormal returns calculated on the first day after the publication of the column.

2 Data processing

We collected all the 'Letter to investor' columns published, from January 2005 to December 2020, mentioning single Italian companies listed on the domestic Stock Exchange. In the time span 2005-2014 most of the columns contain explicit trading advice, that disappear since 2015 (overall: 350 stocks with explicit recommendation, and 366 without). Therefore, the 'Letter to investor' corpus consists of 716 articles, with an average length of about 1500 words, totalling 1104925 tokens and 482735 types. The type-token ratio is therefore very high (0.437), primarily due to the presence of: proper names (managers, companies, banks, rating agencies, countries), numbers and shares, dates, acronyms, anglicisms, etc.

The first task performed on the corpus was the lemmatisation using the R `udpipe` library [13]. The treebank on which this procedure was based, the Italian Stanford Dependency Treebank (ISDT), seems quite suitable for the purpose, as it was created using newspaper articles and Wikipedia pages. Following, we chose to select only nouns, adjectives, verbs, and adverbs for the next phase of feature extraction.

We then calculated, using a bag-of-words approach, some stylistic features pertaining to the lexical complexity and some lexica-based features pertaining the polarity of the texts. With regard to sentiment analysis, in previous works we exploited the NRC general lexicon, pointing out the need for resources in Italian comparable to the financial lexicon of Loughran and McDonald [6]. Loughran and McDonald lexicon contains lists of positive and negative terms, and other potentially interesting lists of words, e.g., related to uncertainty. Therefore, we decided to automatically translate the lexicon via 'eTranslation', an online machine translation service provided by the European Commission[2], qualitatively reviewing the result.

We also computed some lexical complexity measures, regarding readability and lexical diversity. The purpose of this task was to provide further dimensions that could potentially affect the reader and consequently the abnormal returns. In particular, these indices aim to discriminate the articles complexity and the authors' style of writing, as some journalists have taken turns as editors of the column over the years. For the predictive models, among several metrics we selected two indices of readability (mean sentence length, mean word syllables) and two indices of lexical diversity (Dugast's Uber Index U, Simpson's D) [14]. Since readability indices often contain specific weights for a given language, in this work we chose two unweighted indices[3]. Instead, lexical diversity is generally measured with respect to the type-token ratio. Considering the high level of correlation found between the several available indices, we chose U and D because we already tested them [5] and, even though they are dependent on the text length, weekly column's structure and layout did not vary in the observed period [4].

2.1 Feature extraction

The main goal of the feature extraction phase is to obtain a limited set of variables from the texts of the column, which will be used as predictors of abnormal returns. We chose to compare three different strategies: using the frequencies of a set of words determined by the value of the RAKE (Rapid Automatic Keyword Extraction) index, selecting the most important words based on the TF-IDF index, and creating a set of vectors derived from word embeddings.

RAKE [10] index derived from a keyword extraction algorithm based on the ratio of the degree to the frequency of each word. The algorithm creates a word

[2] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

[3] In future studies it would be useful to exploit Italian based indices, such as READ.IT and GulpEase.

degree matrix with each row displaying the number of times a given word co-occurs with another word in the sentences that make up a document. The degree of a word is calculated as the sum of the number of co-occurrences, then it is divided by the occurrence frequency. In this way, a ranking of the most relevant words in a text can be performed.

TF-IDF [11] is a widely used index that evaluates how relevant a word is to a document in a corpus and it is based on the ratio between the term frequency and the inverse document frequency of the given term. TF-IDF brings out the words that occur many times in a few documents and those words would be relevant to distinguish documents. We decide to compute the numerator as the normalized term frequency, i.e., the relative term frequency within a document. In order to obtain a selection of the most relevant terms (with a minimum frequency of 5), we then summed the TF-IDF normalised values within each document, thus obtaining a ranking for the terms. In this case, the selected predictors will be weighted precisely according to the TF-IDF index with respect to each document.

Word embedding is a popular text representation where words that have the same meaning will have a similar vector representation [7]. We chose to train our word embedding with the column corpus using the Global Vector (GloVe) model [9], usually able to identify synonyms or to suggest a word to complete a sentence. The GloVe model use an unsupervised learning algorithm to map the words into a N-dimensional space, where the semantic similarity among words is explored through the distance among the words. GloVe model builds a words co-occurrence matrix and then uses the matrix factorization technique for word embedding. Since word embedding techniques use context to create the word representations, after the corpus lemmatisation and the vocabulary creation (with a minimum frequency of 5), we defined a context window (a string of words before and after a focal word) of 3 words that was used to train our word embedding model. After obtaining 100 vectors for each word included in the corpus vocabulary, the word embedding features for each column were computed as the averages of the word vectors for all the vocabulary words appearing in the column [8]. A few examples of the semantic power of the word embedding trained in the corpus are provided in Table 1.

Table 1: Examples of similar words (cosine similarity) extracted from trained word embedding.

Input word	***Word 1 (similarity)***	***Word 2 (similarity)***	***Word 3 (similarity)***
strategy	development (0.751)	company (0.733)	growth (0.731)
plan	foresees (0.761)	industrial (0.723)	investment (0.671)
business	activity (0.905)	group (0.778)	industry (0.741)

3 Experiments and results

To compare the predictive power of 100 features obtained using RAKE, TF-IDF and word embedding respectively, we tested different statistical learning models (Linear Regression, Partial Least Squares Regression (PLS), Principal Component Regression (PCR), Random Forest, and Support Vector Machines with Radial Basis Function Kernel (SVM)) estimating the value of abnormal returns. Abnormal returns (ARs) were computed following the Market Adjusted Model:

$$AR_{jt} = R_{jt} - R_{mt}$$

where R_{jt} is the stock return of company j (mentioned in the column) on day t, R_{mt} is the stock market return (MILAN COMIT GLOBAL + R - PRICE INDEX) on day t, and AR_{jt} is the abnormal return of company j on day t (AR_{jt} are averaged across companies to get the mean Abnormal Return on day t, AR_t).

The experiments setting was the same for comparison purposes. The values of the ARs were then estimated through each of the five statistical models using 100 features selected through RAKE, TF-IDF and word embedding. To the 100 features of each model, we added five econometric control variables collected from DataStream and Borsa Italiana databases (company's size, price-to-book value (PBV), past performance, company's beta, and presence of concurrent news), four sentiment scores (NRC sentiment; Loughran-McDonald sentiment, uncertainty, and modal words scores) and four lexical complexity indices. Hence, a total of 114 predictors are included in each model. We performed a 5-fold cross-validation with 100 repetitions on the training set. The training set was made up of stocks with recommendations, while the test set was made up of stocks without recommendations. We obtained the best results through Principal Component Regression in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)[4]. Table 2 shows the results of the models on the test set, while Figure 1 shows the most important features in the PCR-RAKE model based on weighted sums of the absolute regression coefficients[5]. It is important to stress that we found among the most important features: words extracted from RAKE, indices of sentiment and uncertainty, econometrics, and lexical complexity (Fig. 1). In contrast, in models with TF-IDF and word embedding few non-textual variables appeared among the most important ones. Furthermore, it should be mentioned that using terms in the models also allows for greater interpretability, which is simply not possible using word embeddings.

[4] Given a large set of variables, PCR probably overcomes the multicollinearity issue better than other techniques.

[5] Partial Least Squares regression also showed good results, especially through features extracted by word embedding and TF-IDF.

Table 2: Models results (MAE and RMSE) for each set of features – test set.

Model	*Features type*	*MAE*	*RMSE*
Linear Regression	RAKE	0.89281880	1,13017700
	TF-IDF	0.82722900	1.05565800
	Word Embedding	0.88001680	1.10798300
PLS	RAKE	0.11087380	0.12742180
	TF-IDF	0.03618955	0.04379134
	Word Embedding	0.03621648	0.04382179
PCR	RAKE	0.02606201	0.03253241
	TF-IDF	0.03017980	0.03764856
	Word Embedding	0.03017943	0.03764809
Random Forest	RAKE	0,07227743	0,09321358
	TF-IDF	0,04359055	0,05658084
	Word Embedding	0,06354684	0,07858456
SVM	RAKE	0,07826125	0,09849686
	TF-IDF	0,10417250	0,12097520
	Word Embedding	0,07064870	0,09008151

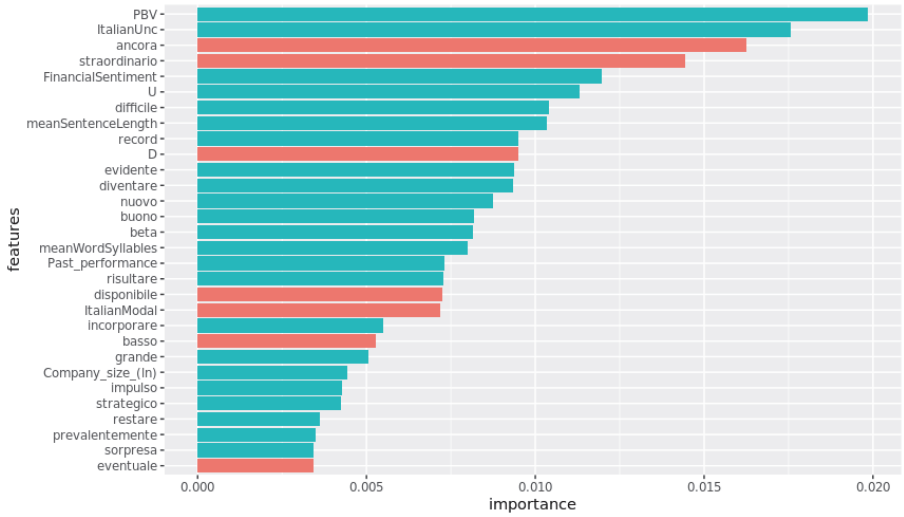


Figure 1: Most important features - RAKE PCR model (blue bars indicate a positive effect, red bars a negative effect).

4 Conclusions

Results confirmed our hypothesis, as RAKE features performed better in terms of both MAE and RMSE in the PCR model. The PCR model result for the word embedding features is similar to that achieved with TF-IDF. The main reason in our

opinion is that word embedding defines the multidimensional coordinates of each word, but to extract features for each text, we have to average each coordinate among the document words, resulting in fuzzy measures. We believe that the main usefulness of word embedding is in the recovery of semantics, while its use as features should be reviewed, for example through universal dependencies [5]. A further possibility to explore for exploiting word embeddings could be the use of measures like RAKE and TF-IDF to weight differently the numerical vectors. Future developments of this research should also consider n-grams and improve the translation of the financial lexicon for sentiment.

References

1. Barber, B.M., Odean, T.: All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. The Rev. of Financial Stud., 21,785–818 (2008).
2. Cervellati, E.M., Ferretti, R., Pattitoni, P.: Market reaction to second-hand news: Inside the attention-grabbing hypothesis. Appl. Econ, 46(10), 1108-1121 (2014).
3. Ferretti, R., Sciandra, A.: Does the attention-grabbing mechanism work on Sundays? Influence of social and religious factors on investors' attention. Rev. of Behav. Fin. (2021)
4. Ferretti, R., Sciandra, A.: Media and Investors' Attention. Estimating analysts' ratings and sentiment of a financial column to predict abnormal returns. In: SIS 2021 Book of Short Papers, Pearson, 1543-1548 (2021).
5. Lai, M., Cignarella, A.T., Finos, L., Sciandra, A.: WordUp! at VaxxStance 2021: Combining Contextual Information with Textual and Dependency-Based Syntactic Features for Stance Detection. In: XXXVII Int. Conf. of the Spanish Society for NLP*,* 2943, 210-232. CEUR (2021).
6. Loughran, T. and McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The J. of Finance, 66(1), 35-65 (2011).
7. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. Proc. of International Conf. on Learning Representations (2013).
8. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT), 17(3), 1-23. (2017).
9. Pennington J., Socher R., Manning C.D.: GloVe: Global Vectors for Word Representation, in Empirical Methods in Natural Language Processing (EMNLP), 1532-1543 (2014).
10. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text mining: applications and theory, 1, 1-20 (2010).
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management*,* 24(5), 513–523 (1988).
12. Sciandra, A.: COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings, 2020 IEEE Symposium on Computers and Communications (ISCC), 1-6 (2020).
13. Straka M., Straková J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. ACL, Vancouver, Canada, 88-99 (2017).
14. Tweedie, F.J., Baayen, R.H.: How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32(5), 323-352 (1998).

Topics and trends in the end-of-year addresses of the Presidents of the Italian Republic (1949-2021)

Temi e tendenze nei discorsi di fine anno dei Presidenti della Repubblica Italiana (1949-2021)

Matilde Trevisani and Arjuna Tuzzi

Abstract The aim of this study is to analyse the corpus of end-of-year speeches of the Presidents of the Italian Republic from a diachronic perspective in order to identify groups of words that share the same pattern and depict main topic trends. The procedure adopted for the recognition of the dynamics of word frequencies moves from a statistical learning perspective and envisages decisions that concern the normalization of occurrences, the smoothing of the trajectories, and the curve clustering. In resulting clusters, emerging topics as well as those that have disappeared over years are clearly visible but, above all, the individual trait of the President stands out as the most relevant element that determines the contents of his discourses.

Riassunto *Questo studio ha come obiettivo l'analisi del corpus dei discorsi di fine anno dei Presidenti della Repubblica Italiana da una prospettiva diacronica allo scopo di identificare gruppi di parole che condividono lo stesso andamento e riconoscere le tendenze degli argomenti principali. La procedura adottata per il riconoscimento della dinamica delle frequenze delle parole parte da una prospettiva di apprendimento statistico e richiede decisioni che riguardano la normalizzazione delle occorrenze, il lisciamento delle traiettorie e la classificazione delle curve. Nei gruppi ottenuti, sono chiaramente visibili gli argomenti emergenti e quelli scomparsi nel corso degli anni ma soprattutto è il tratto individuale del Presidente che spicca come l'elemento più importante che determina i contenuti dei suoi discorsi.*

Key words: presidential addresses, functional data analysis, chronological textual data, curve clustering, topic trends

[1] Matilde Trevisani, Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" – Università degli studi di Trieste, email: matilde.trevisani@deams.units.it

Arjuna Tuzzi, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia applicata – Università degli studi di Padova, email: arjuna.tuzzi@unipd.it

1 Introduction

The President of the Republic represents the highest office in the Italian system and also one of the most popular political representatives in civil society. The presidential end-of-year speech (or similar Christmas speech) is an established tradition in many countries all over the world; in Italy it has been held every year since 31st December 1949 (second year of Einaudi's office). President Luigi Einaudi gave his short messages on the radio and so did Giovanni Gronchi during the first year of its office. In 1956 Gronchi gave his message on TV and, since then, every New Year Eve the presidential address is simulcast by main Italian TV channels and it has become a popular media event. Previous interdisciplinary studies have already explained how the duration, style, habits, contents and media coverage of the messages have changed over time (Cortelazzo and Tuzzi, 2008) and further in-depth analyses observed that the presidential end-of-year addresses, unlike other speeches by other relevant political representatives, is strongly influenced by the individual traits and choices of the President (Cortelazzo, 2018).

The main aim of this study is to analyse the corpus of Italian end-of-year addresses from a diachronic perspective in order to draw the temporal evolution of content words, identify groups of words that share the same temporal pattern, and highlight main topics trends. The study is meant to update and innovate a previous study (Trevisani and Tuzzi, 2012, 2013) not only through the integration of the corpus with the speeches delivered until 2021 (the previous work stopped at Napolitano's 2011 speech) but also by using different strategies for: normalizing the raw occurrences of words (relative frequencies versus chi-square transforms), smoothing trajectories (previously we used wavelets whereas now splines), and detecting curve clusters (model-based versus distance-based methods).

2 Corpus and data

The corpus includes 73 end-of-year addresses (1949-2021) by 10 Presidents: Luigi Einaudi, Giovanni Gronchi, Antonio Segni, Giuseppe Saragat, Giovanni Leone, Sandro Pertini, Francesco Cossiga, Oscar Luigi Scalfaro, Carlo Azeglio Ciampi, Giorgio Napolitano, Sergio Mattarella. The office of the President usually lasts seven years. The exceptions are Segni, who resigned from his position after two years and Napolitano, who got a second office and resigned two years later (Mattarella also obtained a re-election in 2022 but, at the moment, delivered seven discourses).

The corpus has been constantly updated through the collation of the digitalised version with the audio-visual version made available on the Quirinale website (www.quirinale.it) in order to have a text that corresponds with the version actually spoken. For text pre-processing and lemmatization, the tools included in the *Taltac* software (Bolasco et al., 2019) were exploited as they are suitable for the type of corpus under scrutiny. Texts have been furtherly revised manually. The corpus is available both in the original version and in the lemmatized version, in the latter words

(graphic forms) are replaced by the lemma and grammatical category (lemma_CAT) pair that responds to the part-of-speech perspective.

With a total of 123,792 word tokens and 6,953 lemma-types, the corpus is large and shows a good degree of redundancy (lemma-token ratio 5,6%, hapax legomena 38,1%, mean frequency 18). Speech length in terms of total occurrences (Figure 1) is irregular over time, starting with Einaudi's short radio messages up to Scalfaro's broad addresses. The messages have been expanding over time with some interesting trends: early presidents tended to deliver speeches that were shorter at the start of their office and broader at the end, whilst a substantial length stabilization is perceptible for the last three presidents (Ciampi, Napolitano, Mattarella).

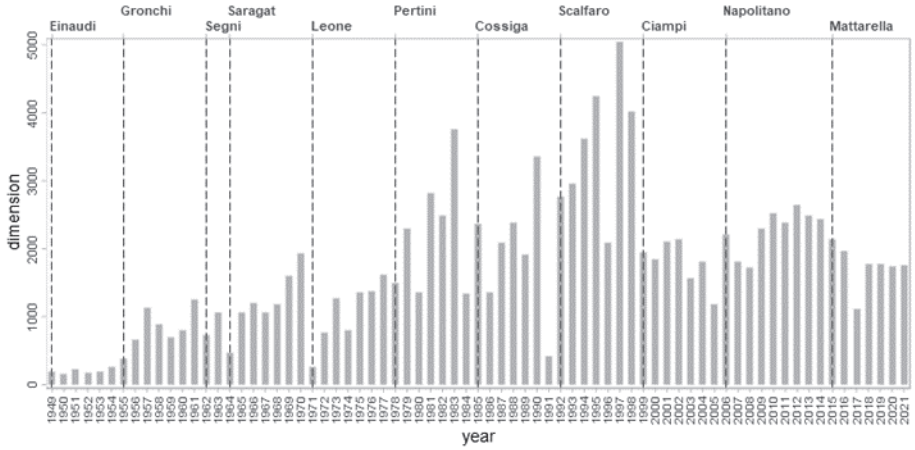


Figure 1: Subcorpora dimension: total number of word-tokens in end-of-year speeches

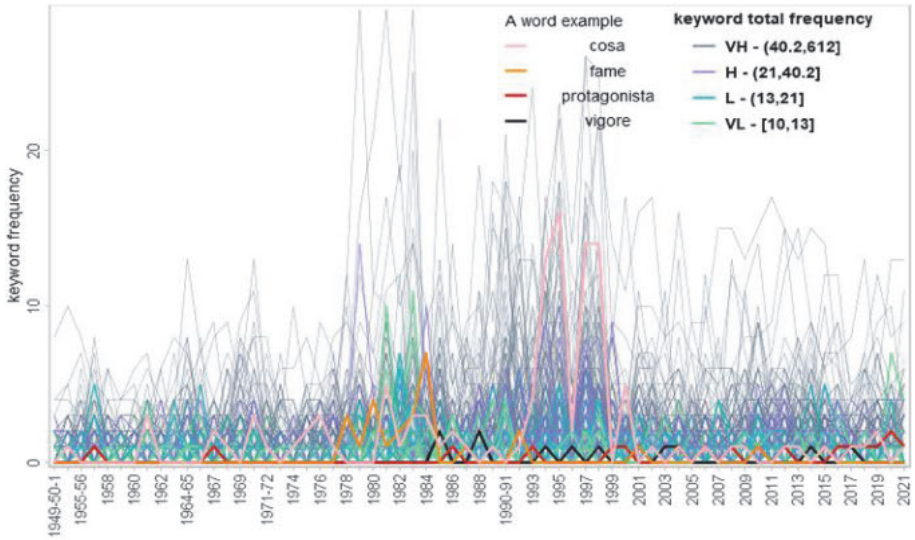


Figure 2: Keyword trajectories (original data): word frequency classes (VH=very high, H=high, L=low, VL=very low) are color-highlighted and a word from each frequency class is drawn

In order to study the contents of the speeches, in this analysis we decided to select nouns with a frequency higher than 10 yielding a total of 612 entries of the lemmary. The starting matrix is a (lexical) contingency table that reports the occurrences of each noun in each discourse (term-document matrix) and, since the corpus is diachronic, these occurrences are observed with reference to the discourse year. Thus, each noun profile consists in an ordered sequence of frequencies, which we can represent as a trajectory (Figure 2). Finally, we aggregated the shortest discourses with the contiguous ones (yielding 65 time-points): 1949-1951 and 1952-1954 (Einaudi); 1955-1956 (Gronchi); 1964-1965 (Saragat); 1971-1972 (Leone); 1990-1991 (Cossiga).

3 Curve clustering

The pursuit of patterns within the trajectories of word occurrences starts from the assumption that the occurrence of a word at a time-point is able to reflect the word's vitality and, if so, it represents the observable and discrete realization of this underlying latent and continuous function. Starting from such a time series, we want to reconstruct the general dynamics, that is, the shape of the underlying function, which generated the observed trajectory. In this context, it seems appropriate to adopt a functional data analysis (FDA) approach (Ramsay and Silverman, 2005).

The procedure adopted for the recognition of dynamics in the trajectories traced by lemmas works from a statistical learning perspective in three steps (Trevisani, 2018, Trevisani and Tuzzi, 2018): (1) normalization of occurrences, (2) smoothing of trajectories, (3) curve clustering. The calculations were performed with the support of *R* (R core team, 2022) libraries `fda` (Ramsay et al., 2021), `kml` (Genolini et al., 2005), `clusterCrit` (Desgraupes, 2018), and `clusterSim` (Walesiak and Dudek, 2020), supplemented by ad hoc R code.

1) Normalization must be chosen on the basis of both the specific behaviour of the trajectories and the objectives pursued by the classification. We must take into account the variable size of texts across years (number of word-tokens of each message) as well as understand to what extent limit the effect of high unequal popularity of words (Fig. 2). We have chosen the "chi-square" normalization as in previous studies it proved able to capture life cycles of words that are born and die within the period considered. These are typically words with moderate or low frequency in the entire corpus for being sparse over time, thus the chi-square transformation: $y_{ij} = n_{ij} / (n_{i.} \sqrt{n_{.j}} / n)$ (where n_{ij} is the raw frequency of word i at time point j, $n_{i.}$ is the i-row sum, $n.j.$ is the j-column sum, and n the matrix total of the corpus table) causes them to emerge and hence drive subsequent clustering.

2) The normalized frequencies are smoothed in order to eliminate roughness and extreme irregularity of trajectories. Smoothing is performed through B-splines and optimized by the roughness penalty approach to estimation (Ramsay and Silverman, 2005, Trevisani and Tuzzi, 2018). In a previous study (Trevisani and Tuzzi, 2015) we applied a wavelet-based decomposition which proved successful in recognizing the typical bumpy trend of word trajectories. Yet, this time our objective is recognizing

continuous, hence, more easily interpretable shapes, which lead us to opt for the spline functions and, in particular, for B-splines as they consist in a very flexible basis system for non-periodic functional data.

3) The smoothed trajectories are clustered into groups that share similar temporal patterns. The curve clustering resorts to a k-means distance-based algorithm that exploits the Euclidean distance as a measure of (dis)similarity. Incidentally, we chose a distance-based approach to functional clustering since our objective is setting up a procedure eminently exploratory (the procedure is asked to look for "interesting patterns", without prescribing any specific interpretation, to be submitted to subject matter experts who potentially formulate new questions and hypotheses and drive to research insights) and mostly automated (the procedure is asked to be fast and relatively easy to use and understand even by non-statisticians of interdisciplinary research groups). The alternative model-based approach is typically chosen for confirmatory analyses and is generally more demanding in terms of computing and inferential expertise (see an overview of functional data clustering in Jacques and Preda, 2014).

The algorithm has been repeated 20 times (by different initial solutions) for each potential number of clusters (from a minimum of 2 to a maximum of 26 groups), thus generating 20 possible partitions for each number. To determine the optimal cluster number, about 50 different quality criteria are queried and a list of prioritized solutions is formed on the ground of the top choices found (Figure 3a). Once the optimal cluster numbers have been established, the partitions mostly indicated by the quality criteria are selected (among the 20 available) and, among them, the one that maximizes the degree of overlap with the others (generalized Rand index) is the winner as it should assure stability and coherence of groups (Wagner and Wagner, 2007). In our example, if we select a clustering into 4 groups – which is the best 'parsimonious' solution, five partitions resulted as the mostly indicated by the validation criteria and partition 14 resulted as the one maximizing the average Rand index calculated on the four couples at comparison (Figure 3b).

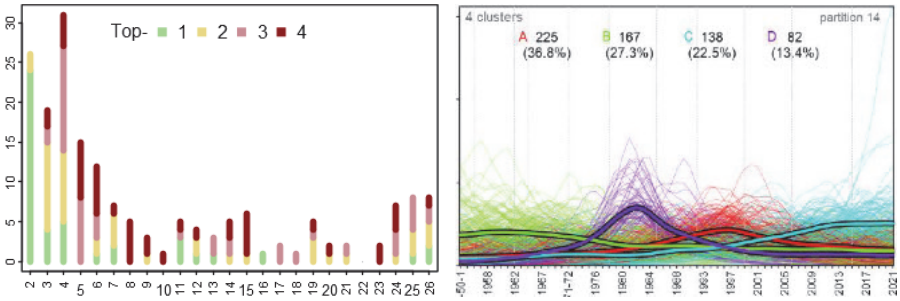


Figure 3: (a) Top-1 to top-4 cluster numbers (on the left) and (b) Best clustering solution: 4 clusters as in the best partition (14) according to the mean Rand index (on the right)

4 Results

The indications obtained from both quality criteria and clustering visualization led us, at the end of this contest, to the solution envisaging 11 clusters (Figure 4) since it ensures a finer-grained partition and was often selected as the first best option.

Some clusters gather words that occur less and less in presidential speeches (cluster I: *auspicio*-hope, *fortuna*-fate, *ricostruzione*-reconstruction, *progresso*-advancement, *miseria*-misery, *benessere*-well-being) and that characterized past periods (cluster G: *lavoratore*-worker, *sindacato*-labor union, *imprenditore*-entrepreneur, *ceto*-class, *classe*-class, *categoria*-class, *organizzazione*-organization, *reddito*-income, *produzione*-production, *investimento*-investment, *assistenza*-assistance) but, with particular relevance, the office of a specific President (cluster G: Saragat).

In their historical periods, some Presidents take on the role of drivers of specific clusters, such as the clear-cut ones of Pertini in the 1980s (cluster H: *terrorismo*-terrorism, *terrorista*-terrorist, *ordigno*-bomb, *strage*-massacre, *angoscia*-anxiety, *dolore*-sorrow, *fame*-hunger, *disoccupazione*-unemployment, *preoccupazione*-concern), Scalfaro in the 1990s (cluster C: *partito*-party, *autorità*-authority, *magistrato*-magistrate, *magistratura*-judiciary, *giudice*-judge, *giudizio*-judgment, *garanzia*-guarantee, *criminalità*-crime), Ciampi at the turn of the new millennium (cluster F: *Europa*-Europe, *Euro*, *Unione Europea*-European Union, *dialogo*-dialogue, *stabilità*-stability).

There are then clusters which exhibit a more widespread (transversal to several Presidents), and in some cases fluctuating, temporal evolution: on one side, cluster B with a decreasing trend from early ’90 and, on the opposite side, clusters A, D and E, the first two with an increasing trend from around mid ’80, the third showing an upward trend from end ’90 reaching a plateau over 2010-2021. They seem to mark a cultural change that took place around the 1990s but remain difficult to interpret. Moreover, cluster E shows a substantial continuity of thematic discourses in the terms of Napolitano and Mattarella and reveals the challenges, paradigm shifts and emerging topics of the new era, starting around 2011 (*futuro*-future, *coesione*-cohesion, *memoria*-memory, *rischio*-risk, *sfida*-challenge, *cambiamento*-change, *opportunità*-opportunity, *emergenza*-emergency, *malattia*-desease, *ricerca*-research, *innovazione*-innovation, *università*-university, *conoscenza*-knowledge, *scienza*-science, *donna*-woman, *territorio*-territory, *impresa*-enterprise, *immigrato*-migrant).

Finally, it is worth noting the existence of clusters/singletons that include specific words (cluster J: *Francesco*, *concittadini*-fellow citizens; cluster K: *pandemia* pandemic) and are soaring in the last years.

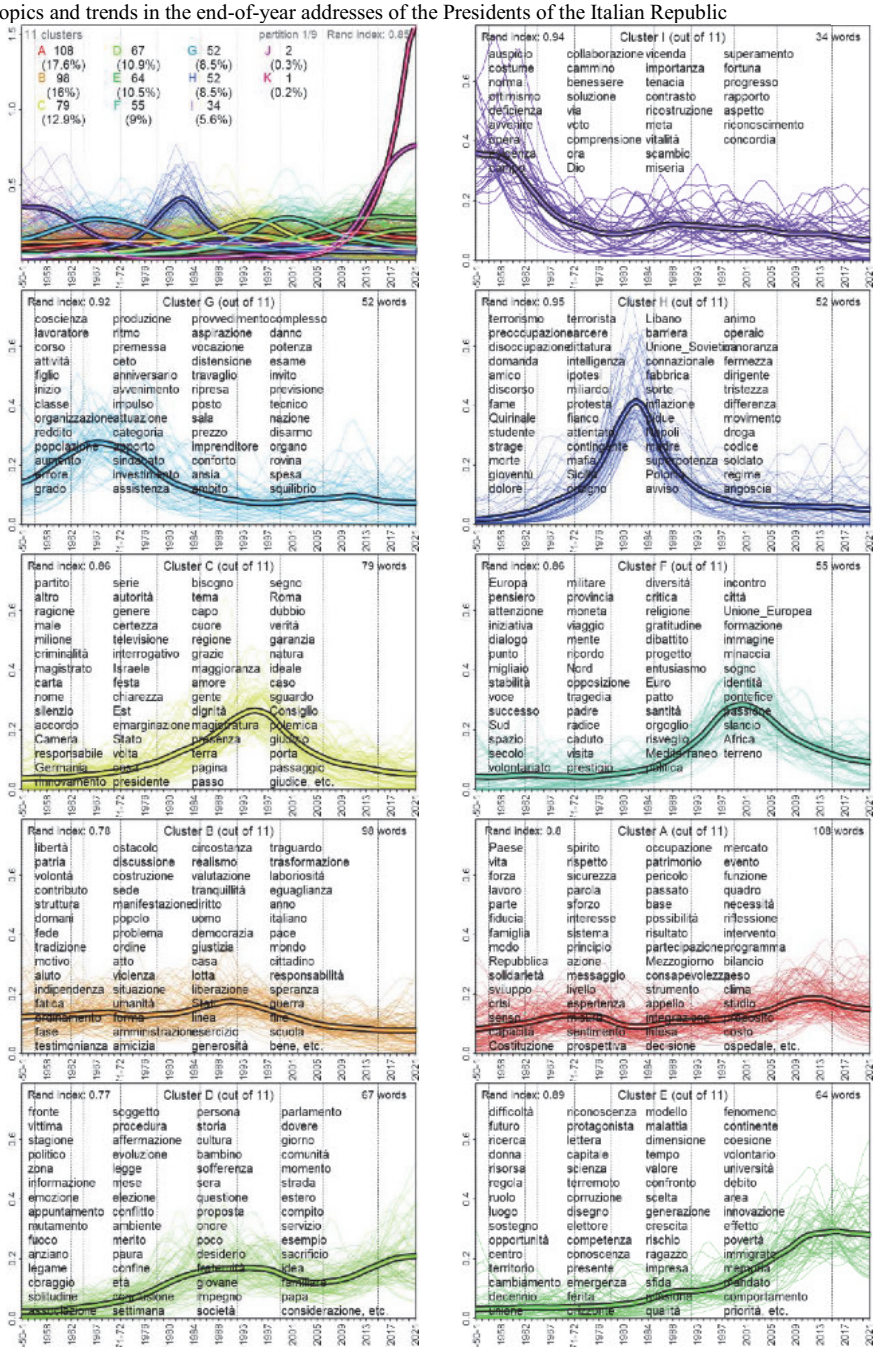


Figure 4: Final selected clustering: the overall 11 clusters and 9 clusters

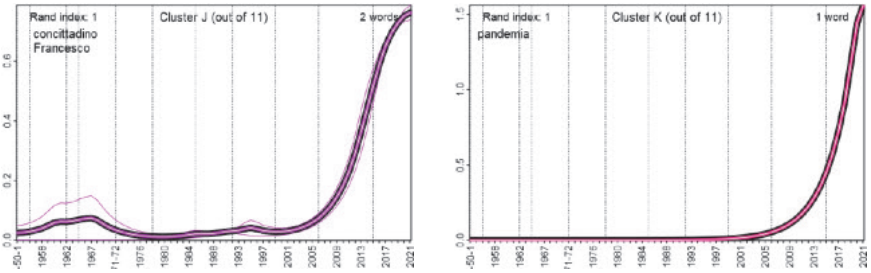


Figure 5: Clusters of sporadic and exceptional ('black swan' events) words

5 Discussion and conclusions

The presence of the last two clusters and the difficulty in some cases of clearly interpreting temporal evolutions offer us the opportunity to point out once again that in the three-step procedure (normalization, smoothing, clustering) the most relevant decision concerns data normalization, as it heavily affects results in terms of both the subsequent smoothing and clustering. In fact, a chronological corpus is typically characterized by extremely irregular (peak-and-valley) trajectories of word frequencies over time (if we consider the term-document matrix by row) and by a marked disparity of frequency classes between the most popular words and the rest of the others (if we look at data by column). Then, unless we let most popular words drive the clustering process (thereby, limiting the normalization to a standardization by column for the uneven dimension over time), we need to adequately treat frequency asymmetry in order to effectively gather the synchrony of word histories at comparison. The "chi-square" double normalization here adopted leads to valuable results and solves the problem of high-frequency words but, as a well-known effect, emphasizes the role of rare words and, in some conditions, also shows undesirable effects on either smoothing or clustering.

As far as normalization is concerned, the authors are reviewing several types of normalization, and their possible effects on clustering results, within two broad classes: the sole normalization per column, leaving the word popularity unaltered (different curve amplitude influences the cluster formation; an example in Trevisani and Tuzzi, 2015) and the double (by both column and row) normalization (comparison is focussed on curve phase or synchrony). In the latter case, it is important to choose whether to treat words of different popularity homogeneously (e.g. the minmax normalization) or differently according to their characteristics (e.g. chi-square normalization which emphasizes rare words, as in the present paper, or non-linear normalization which remedies for each word asymmetry as in Sciandra et al., 2021).

In conclusion, the choice of a data normalization must take into account the specific features of textual data above mentioned and the objectives pursued: on the one hand, it is necessary to understand how to reduce the effect of text size, on the

other, it is necessary to understand what role to assign to the popularity of words and the shape of the curves in the clustering solution.

From this perspective, at present, it is still difficult to imagine standard procedures that are not tailored on a case-by-case basis.

References

1. Bolasco, S., Baiocchi, F., Canzonetti, A.: Taltac2, release 2.11.2 (2019)
2. Cortelazzo, M.A.: Il linguaggio dei presidenti. In Cassese, S., Galasso, G., Melloni, A. (eds) I presidenti della Repubblica. Il capo dello stato e il Quirinale nella storia della democrazia in Italia, pp. 901-929, Bologna, Il Mulino (2018)
3. Cortelazzo, M.A., Tuzzi, A. (eds.): Messaggi dal colle. I discorsi di fine anno dei presidenti della Repubblica. Marsilio, Venezia (2008)
4. Desgraupes, B.: clusterCrit: Clustering Indices, R package version 1.2.8 (2018)
5. Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C.: kml and kml3d: R Packages to Cluster Longitudinal Data, *J. Stat. Softw.* 65(4), 1-34 (2015)
6. Jacques, J., Preda, C.: Functional data clustering: A survey. Advances in Data Analysis and Classification, 8(3), 231-255 (2014).
7. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022)
8. Ramsay, J.O., Graves, S., Hooker, G.: fda: Functional Data Analysis, R package version 5.5.1. (2021)
9. Ramsay, J.O., Silverman, B.W.: Functional data analysis. Springer series in Statistics, Springer, New York (2005)
10. Trevisani, M., Tuzzi, A.: Chronological analysis of textual data and curve clustering: preliminary results based on wavelets. In: Società Italiana di Statistica, Proceedings of the XLVI Scientific Meeting. CLEUP, Padova (2012)
11. Trevisani, M., Tuzzi, A.: Shaping the history of words. In: Obradović, I., Kelih, E., Köhler, R. (eds) Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO), pp. 84-95. Akademska Misao, Belgrade (2013)
12. Trevisani, M., Tuzzi, A.: A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal, Q&Q 49, 1287-1304 (2015)
13. Trevisani, M., Tuzzi, A.: Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories, Knowl.-Based Syst. 146, 129-141 (2018)
14. Trevisani, M.: Functional Data Analysis and Knowledge-Based Systems. In: Tuzzi, A. (ed) Tracing the Life Cycle of Ideas in the Humanities and Social Sciences, pp. 167-187, Cham, Springer (2018)
15. Sciandra, A., Trevisani, M., Tuzzi, A.: Sulle tracce dell'espressione dell'interiorità: analisi diacronica di un corpus di narrativa italiana del XIX-XX secolo, Rivista internazionale di tecnica della traduzione / International Journal of Translation 23, 219-233 (2021)
16. Wagner, S., Wagner, D.: Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik Karlsruhe (2007)
17. Walesiak, M., Dudek, A.: The Choice of Variable Normalization Method in Cluster Analysis. In Soliman, K.S. (ed) Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges. Proceedings of the 35th International Business Information Management Association Conference (IBIMA), 1-2 April 2020, pp. 325-340, Seville, Spain, International Business Information Management Association (2020).

Thematic analysis on online education issues during COVID-19

Analisi tematica sulla didattica a distanza durante il COVID-19

Valerio Basile, Michelangelo Misuraca and Maria Spano

Abstract The emergency induced by COVID-19 pandemic posed the greatest challenge that any national education system has ever faced, and this issue was widely discussed on all media sources as well as on social media platforms. In this paper, we aim at analysing the discourse on online teaching developed by Italian tweeters during the past school year, creating a digital storytelling. Employing thematic analysis, an approach used in bibliometrics to highlight the conceptual structure of a research domain, different time slices have been described, bringing out the most discussed topics. The mapping of these topics allowed obtaining an easily readable discourse representation, paving the way for a novel use of thematic analysis in social sciences.

Abstract *L'emergenza causata dalla pandemia da COVID-19 ha posto la più grande sfida che qualsiasi sistema educativo nazionale abbia mai affrontato, e tale tema è stato ampiamente discusso da tutti i mezzi d'informazione oltre che sui social media. In questo lavoro ci proponiamo di analizzare il discorso sulla didattica online dei tweeter italiani durante il passato anno scolastico, creando uno storytelling digitale. Utilizzando l'analisi tematica, approccio usato in bibliometria per evidenziare la struttura concettuale di un dominio di ricerca, sono stati descritti diversi intervalli temporali, evidenziando i temi più discussi. La mappatura di questi temi ha permesso di ottenere una rappresentazione del discorso facilmente leggibile, aprendo la strada a un nuovo uso dell'analisi tematica nelle scienze sociali.*

Key words: topic detection, thematic analysis, text analytics

---

Valerio Basile
DI - University of Turin, e-mail: valerio.basile@unito.it

Michelangelo Misuraca
DiScAG - University of Calabria, e-mail: michelangelo.misuraca@unical.it

Maria Spano
DiSES - University of Naples Federico II, e-mail: maria.spano@unina.it

1 Introduction

From the first case of the *severe acute respiratory syndrome coronavirus 2*, commonly known as COVID-19, reported in China on December 2019, the contagion spread rapidly worldwide, becoming in a few weeks a global pandemic [17]. Italy was directly interested by this new viral infection at the end of February 2020, as one of the first countries in Europe, when a small cluster of cases was detected in Lombardy, in Northern Italy. Due to the rapid spread of COVID-19 and the occurrence of new clusters in different areas of the national territory, the Italian government decided to take drastic actions. Unprecedented social distancing measures were soon put into practice, which significantly triggered a radical and rapid change in everyday life. Some radical measures were taken, such as local, regional, and international travel bans, quarantine practices, and a complete shutdown of all non-essential private and public activities in the whole country on the 9th of March.

All Italian schools suspended face-to-face education and remote teaching became the rule almost overnight. Emergency distance education tools were rapidly put into use to offer some form of continuance for students' education. Hundreds of students and teachers have tried to adapt to this new situation, where both lessons and exams sessions have moved to the online environment [11]. The COVID-19 pandemic has posed perhaps the greatest challenge that any national education system has ever faced [6] and the topic was widely discussed in the last two years. Even after the relaxation of restrictions, the reopening of schools has sparked a series of debates (e.g., how to contain the contagion in classes, how to cope with positive cases detected at school, teaching staff's mandatory vaccination) on all media sources, including social media platforms. Internet and social networks like Facebook and Twitter have become an integral part of daily life. People search information online and share their opinions, allowing a rapid circulation of news and producing a huge amount of textual contents, especially with the development of Web 2.0 tools and the definition of a new virtual environment to communicate and collaborate [16].

In this work, we aim at investigating people's views and highlighting the main topics related to the debate of how the Italian education system have faced with the pandemic, by analysing the massive amount of comments shared by Italian users on Twitter between July 2020 and October 2021. The problem of extracting the topics embodied in a collection of texts has been faced in different ways by scholars [9]. When there is no prior knowledge concerning the analysed texts, several unsupervised approaches can be carried out [10]. To enhance topics' visualisation and interpretation, and to track the topical trends over a given time-span, here we refer to *thematic analysis*, a technique broadly used in bibliometrics to explore the conceptual structure of a research field. This approach allowed us to extract and automatically label the different topics of the collection and highlight the evolution of the discourse about the reaction of education system to COVID-19, offering interesting insights on this current issue.

2 Methods and Data structure

Before applying any kind of statistical methods on natural language texts, it is necessary to transform them in a structured form. The first step is to scan each text and identify all the different words, obtaining a list containing all the words used in the entire collection, commonly known as "vocabulary". Then, a pre-processing stage is necessary to reduce the vocabulary's dimension and to avoid non-informative words, removing the so-called stop-words, i.e. the most common terms used in the language and in the specific analysed domain. According to the *vector space model* [12], each text can be then represented as a vector $\mathbf{d_i}$ in the space spanned by the q words belonging to the vocabulary:

$$\mathbf{d_i} = (w_{i1}, \ldots, w_{ij}, \ldots, w_{iq}), \tag{1}$$

where w_{ij} represents the importance of the j-th word in the text-vector. Different weighting schemes can be applied to reflect the importance of words [13], but here we refer to binary weights and assign 1 to words appearing in the text and 0 to words not appearing in the text. All the text-vectors can be juxtaposed and organized in a matrix $\mathbf{D}$ ($p \times q$). To partially recover the contextual information lost in the *bag of words* coding, from matrix $\mathbf{D}$ it is possible to derive a q-dimensional matrix $\mathbf{A} = \mathbf{D}^{\mathsf{T}}\mathbf{D}$, whose generic element $a_{jj'}$ ($j \neq j'$) represents the number of texts in which two words j and j' co-occur (i.e., they both appear in the texts). The a_{jj} elements on the principal diagonal of $\mathbf{A}$ count the total number of texts containing the single word j. The co-occurrence of two words belonging to a text can be normalized by *association strength* [7]. This measure assumes values in the interval [0,1] and reflects the strength of the association among words. A matrix like $\mathbf{A}$ can be seen as undirected weighted graph, where each word is a node and the association between linked words is expressed as an edge, visualizing both single words and subsets of words frequently co-occurring together. To detect subgroups of strongly linked words, where each subgroup corresponds to a topic of the analyzed collection, we refer to community detection algorithms [8]. In particular, *Louvain algorithm* [4] showed high effectiveness with respect to other competing proposals [18].

The topics obtained through the community detection can be projected in a so-called *strategic diagram*, obtaining a thematic mapping of the surveyed domain, in accordance with Callon's *centrality* and *density* [5]. These measures express the role of a topic in organising the domain's conceptual structure. Callon centrality can be read as the relevance of the topic in the entire research domain, while Callon density can be read as a measure of the topic's development.

The strategic diagram allows highlighting four different kinds of topics, depending on the quadrant in which they are mapped:

- higher values of centrality and density define the *hot topics*, well developed and relevant for structuring the conceptual framework of the domain;
- higher values of centrality and lower values of density define the *basic topics*, significant for the domain and cross-cutting to its different areas;

- lower values of centrality and density define *peripheral topics*, not fully developed or marginally interesting for the domain;
- lower values of centrality and higher values of density define *niche topics*, strongly developed but still marginal for the domain under investigation.

It is possible to express the complexity of each topic by scaling its representation on the diagram in accordance with the number of related words. To facilitate the reading of the map, each topic can be labelled with the associated most occurring keywords. Jointly analysing the conceptual structure of different temporal sub-periods, it is possible to shape the topical evolution of the domain, revealing the trajectories of the different topics across time.

To track the evolution of contents shared by twitter users about the school during the emergency period, we considered the large set of data offered by *40wita* [3], the most extensive repository holding tweets written in Italian about the COVID-19. The tweets were selected from the primary collection *Twita* – a massive archive of tweets written in Italian [2] – by using a list of 43 different keywords that include both terms related to the COVID-19 (e.g., *covid19*, *coronavirus*) and other terms and hashtags popular in Italy during the emergency period (e.g., *#iorestoacasa* $\longrightarrow$ "I stay at home", *#andratuttobene* $\longrightarrow$ "everything will be fine").[1]

Starting from this collection, we further filtered only tweets dealing with the schools and remote teaching, by considering the following terms: *dad, nodad, didattica, scuola, genitore, studente, allievo, scolaro, insegnante, maestro, professore, docente, preside, lezione, esame* (including both writing variants and hashtags). We focused our attention on the tweets published between July 2020 and October 2021, the period in which several actions for the reopening and the emergency management in schools were taken by the national government as well as by the regions. In this way, we retrieved a set of 91,098 tweets (without retweets) accompanied by some meta-data, like the username, the publishing date, the retweet count, the like count.

3 Main findings

To highlight the main topics and studying their evolution over time, we decided to divide our reference period (July 2020 and October 2021) into four-time slices. In Table 1, some descriptive statistics about the collection are reported.

[1] The full list of keywords ow 40wita is: covid, covid19, covid-19, corona virus, coronavirus, quarantena, autoisolamento, auto-isolamento, iorestoacasa, stateacasa, COVID19Italia, redditodicittadinaza, eurobond, coronabond, restiamoacasa, preghiamoinsieme, NoMes, milanononsiferma, bergamononsiferma, l'italianonsiferma, abbraccciauncinese, iononsonounvirus, iononmifermo, aperisera, covidunstria, italiazonarossa, bergamoisrunning, quarantena, chiudetetutto, apritetutto, CuraItalia, ciricordiamotutto, oggisciopero, chiudiamolefabbriche, iononrinuncioalletradizioni, andràtuttobene, INPSdown, percheQuando, cercareDi, ringraziarevoglio, 600euro, CineINPS, COVID19Pandemic.

Table 1 Descriptive statistics on the tweets posted in the four time slices

Time slice	Tweets	Tokens	Types	Avg. tweets per day	CV
July 2020 - October 2020	41,928	2,422,454	21,377	340.88	0.61
November 2020 - February 2021	24,488	1,376,376	15,502	252.45	0.34
March 2021 - June 2021	14,058	814,497	11,578	140.58	0.68
July 2021 - October 2021	10,624	621,405	9,171	52.84	0.48

We observed a great amount of tweets posted between July 2020 and October 2020, covering the period in which the Italian government had to put in practice actions for the reopening of schools in safe. Then, in the subsequent time slices the number of posted tweets decreases, highlighting how people have become accustomed to the measures taken to contain the infections.

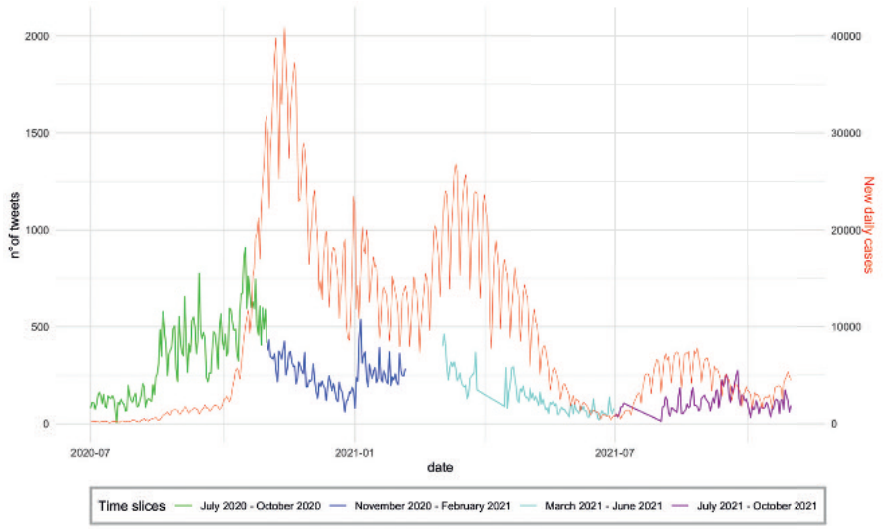


Fig. 1 Daily tweets and COVID-19 new cases in Italy (July 2020 - October 2021)

Figure 1 reports the day-wise distribution of tweets over the four time slices as well as the distribution of the COVID-19 daily new cases in Italy in the same reference period. We observed that in the all analysed periods the number of tweets does not exceed one thousand, with a maximum value of 910 posts on October 16, 2020. Conversely, as highlighted by the different scale on the right of the figure, the number of new daily positive cases is considerably higher, exceeding the 10,000 cases on the same date. The highest daily number of positive cases 40,902 is recorded on November 13, 2020, the period in which the second wave of the pandemic occurs.

On each subset of tweets, we performed the same pre-processing procedure based on the following steps:

- removing URLs, usernames, hashtags and emoticons;
- normalizing each text by stripping special characters and any delimiter different from blank;
- tokenizing texts in bigrams (i.e., sequence of contiguous words were considered as unique entries of the vocabulary). The tokenization is performed with UDpipe [15] with a model pre-trained on Italian Twitter language [14] ;
- filtering out Italian stop-words (e.g., preposition, articles).

At the end of pre-processing, we kept the first 1,500 most occurring bigrams and we derived for each time slice a term co-occurrence matrix as an input for the *thematic analysis*. Figures 2–5 show the maps obtained in the different sub-periods.

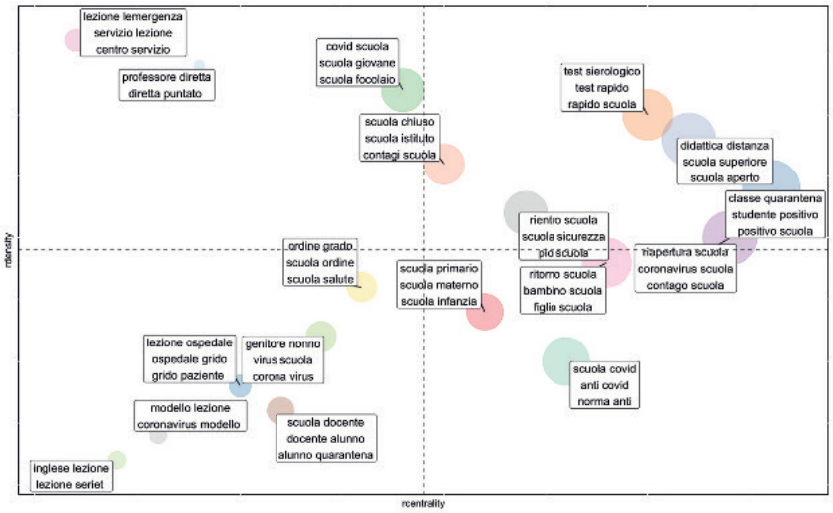


Fig. 2 Thematic map of topics discussed in the period July 2020 - October 2020

In the first sub-period (July 2020 - October 2020) topics are mainly related to the re-opening of schools (e.g., *rientro scuola*, *riapertura scuola*, *ritorno scuola*, *scuola sicurezza*), to the guidelines for containing the contagion (e.g., *scuola covid*, *norma anti*, *anti covid*) and to the introduction of rapid antigen tests for COVID-19. All these topics appeared on the right of the map, reflecting their importance in the discourse.

In the second sub-period (November 2020 - February 2021) the hot topic concerning the online education (*didattica distanza*, *scuola superiore*, *scuola aperta*) mainly applied in the high schools, was also extended to middle schools (*didattica distanza*, *scuola media*, *negozio scuola*), becoming a basic topic. The reason is that during this sub-period a second wave of the pandemic occurs and the main discourses deal with the closure of schools, except for primary schools (*scuola primaria*, *didattica presenza*, *lezione presenza*).

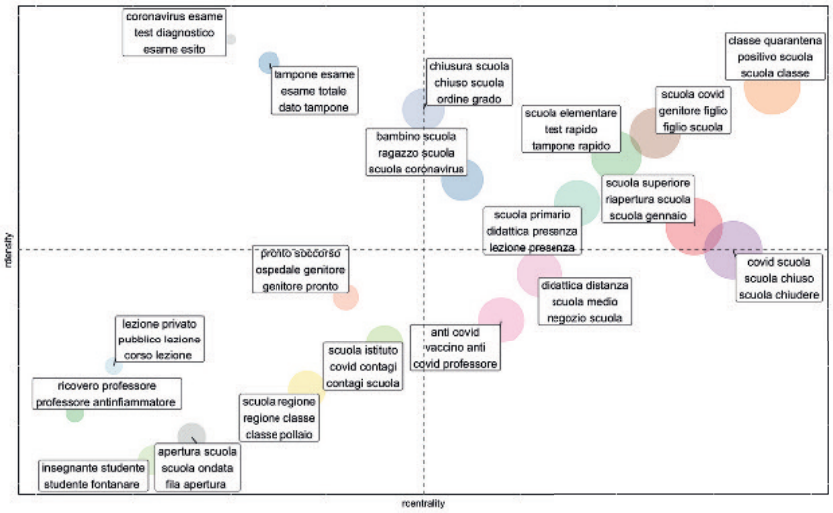


Fig. 3 Thematic map of topics discussed in the period November 2020 - February 2021

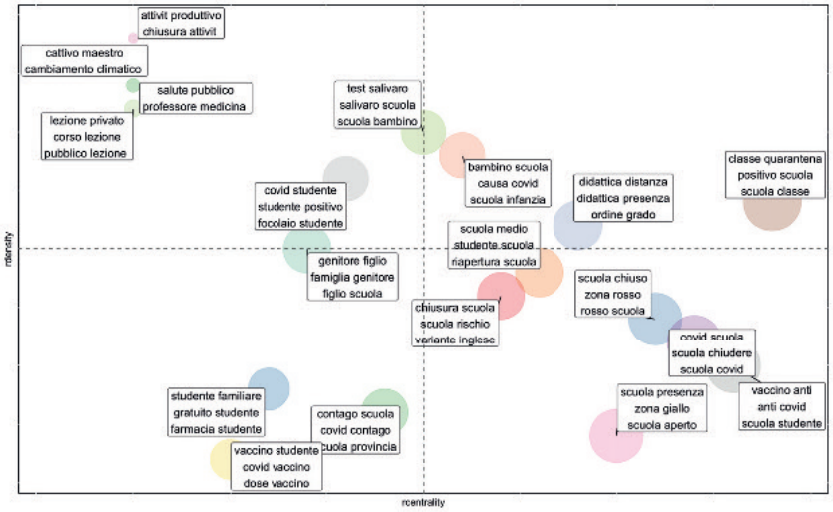


Fig. 4 Thematic map of topics discussed in the period March 2021 - June 2021

The third sub-period (March 2021 - June 2021) the discourse is mainly focused on the different measures applied in the Italian regions to contain the spread of COVID-19, where in low-risk regions (yellow zone) schools remain open, while in regions with a high number of infections there is a return to distance learning. During this period the emergent topic (third quadrant) is related to the vaccination for school staff as well as for students.

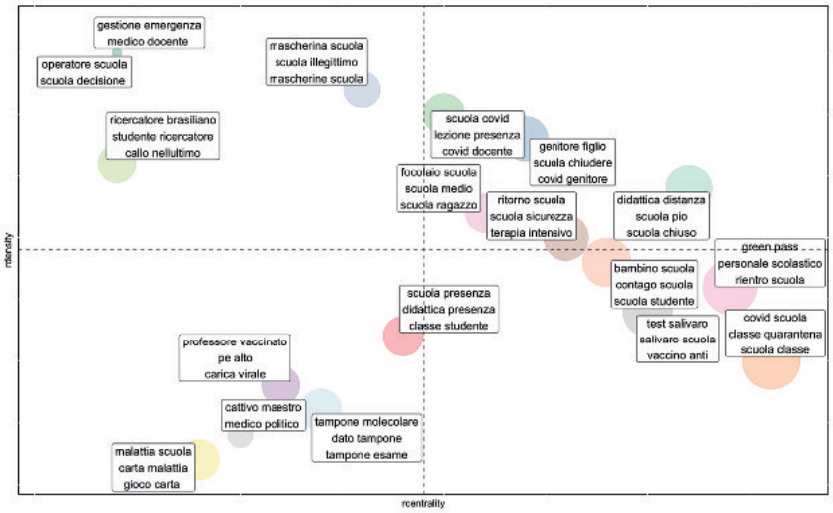


Fig. 5 Thematic map of topics discussed in the period July 2021 - October 2021

In the last sub-period (July 2021 - October 2021) the topic dealing with the mandatory vaccination for school staff (*green pass*, *personale scolastico*, *rientro scuola*) become a basic topic, because in those months the vaccination campaign was practically completed. It is interesting to note how some topics are the same in the different time slices (e.g., *classe quarantena*, *studente positivo*, *classe scuola*) changing only the position in the different quadrants. The practice of putting in quarantine positive students and then the whole class was a hot topic for the first 3 slices, becoming a basic topic in the last period.

4 Conclusion

By analysing the comments posted on Twitter from July 2020 to October 2021, in this paper we highlighted how the discourse on social media about online teaching developed in Italy during the last year of the COVID-19 pandemic. The obtained graphical representations summarize many aspects of the debate about online teaching. Obviously, the presented results are only a small part of what could be observed starting from the thematic maps. Moreover, dividing our reference period (July 2020 and October 2021) into four-time slices allowed us to track the evolution of topics, but the setting of the sub-periods could affect the results, highlighting obviously the main topics of that particular time slices. Performing a thematic analysis allowed to discover the main topics discussed in Italy, distinguishing different topical categories on the basis of their relevance and their development in the discourse. Nevertheless, future developments will be devoted to evaluate different kinds of

pre-processing procedures, for reducing the variability of the vocabulary, and to automatic labelling, in order to make easier the interpretation of the topics detected through the analysis.

References

1. Aria, M., Cuccurullo, C., D'Aniello, L., Misuraca, M., Spano, M.: Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy. Sustainability, **14**, 3643 (2022) doi:https://doi.org/10.3390/su14063643
2. Basile, V., Lai, M., Sanguinetti, M.: Long-term Social Media Data Collection at the University of Turin. In: Proceedings of the Fifth Italian Conference on Computational Linguistics, Turin, Italy. http://ceur-ws.org/Vol-2253/paper48.pdf (2018)
3. Basile, V., Caselli, T.: 40twita 1.0: A collection of Italian tweets during the COVID-19 pandemic. Available online: http://twita.di.unito.it/dataset/40wita (accessed on 10 December 2021)
4. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. P10008 (2008)
5. Callon, M., Courtial, J.P., Laville, F.: Co-word analysis as a tool for describing the network of interactions between basic and technological research - The case of polymer chemistry. Scientometrics **22**, 155–205 (1991)
6. Daniel, S.J.: Education and the COVID-19 pandemic. Prospects **49**, 91–96 (2020)
7. van Eck, N., Waltman, L.: How to normalise co-occurrence data? An analysis of some well-known similarity measures. J. Am. Soc. Inf. Sci. Technol. **60**, 1635–1651 (2009)
8. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
9. Ibrahim, R., Elbagoury, A., Kamel, M.S., Karray, F.: Tools and approaches for topic detection from Twitter streams: survey. Knowl. Inf. Syst. **54**, 511–539 (2018)
10. Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In: Iezzi, D.F., Mayaffre, D., Misuraca, M. (eds.) Text Analytics. Advances and Challenges, pp. 17-28. Springer, Heidelberg, Germany (2020)
11. Pokhrel, S., Chhetri, R.A.: Literature review on impact of COVID-19 pandemic on teaching and learning. High. Educ. Future **8**, 133–141 (2021)
12. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
13. Salton, G., Buckley C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**, 513–523 (1988)
14. Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., Tamburini, F.: PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan (2018)
15. Straka, M., Hajič, J., Straková, J.: UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portorož, Slovenia, 4290-–4297 (2016).
16. Westerman, D., Spence, P.R., Van Der Heide, B.: Social Media as information source: Recency of updates and credibility of information. J. Comput.-Mediat. Commun. **19**, 171–183 (2014)
17. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 (accessed on 15 January 2022).
18. Yang, Z., Algesheimer, R., Tessone, C.J.: A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. **6**, 30750 (2016)

What do we learn by applying multiple methods in topic detection? An empirical analysis on a large online dataset about mobility electrification

Che cosa impariamo applicando diversi metodi per identificare gli argomenti di un corpus? Analisi empirica su un grande insieme di dati online sull'elettrificazione della mobilità

Fabrizio Alboni, Margherita Russo and Pasquale Pavone*

Abstract Identifying the topics covered in a corpus is one of the central issues in automatic text analysis. The objective of our paper is to contribute to the comparative analysis of different methods. In particular, we compare the results obtained through the use of the most common methods for topic identification, applied to the same corpus. The analysis is performed on a large original textual database created from an e-mobility newsletter. To compare the results between the methods, we refer to two criteria. First of all, the semantic consistency of the different models is evaluated by applying the UMass score and Pointwise mutual information. Secondly, the degree of association between the topics identified by the different models is processed using a heat-map and Cramer's V.

Abstract L'identificazione degli argomenti trattati in un corpus è uno dei temi centrali dell'analisi automatica dei testi. Obiettivo del nostro articolo è contribuire all'analisi comparata di diversi metodi. In particolare, confrontiamo i risultati ottenuti attraverso l'uso dei metodi più comuni per l'identificazione di argomenti, applicati allo stesso corpus. L'analisi viene effettuata su un ampio database testuale originale creato a partire da una newsletter sulla mobilità elettrica. Per confrontare i risultati tra i metodi, facciamo riferimento a due criteri. In primo luogo, la coerenza semantica dei vari modelli è valutata applicando il punteggio UMass e il Pointwise mutual information. In secondo luogo, il grado di associazione tra gli argomenti identificati dai diversi modelli viene elaborato con una heat-map e con la V di Cramer.

Key words: topic detection, text mining, Cramer's V, coherence indexes, electric mobility

* Fabrizio Alboni, Università degli studi di Modena e Reggio Emilia, fabrizio.alboni@unimore.it
Margherita Russo, Università degli studi di Modena e Reggio Emilia; margherita.russo@unimore.it:
Pasquale Pavone, Università degli studi di Modena e Reggio Emilia, pasquale.pavone@unimore.it:

Introduction

The continuous proliferation and availability of digitized textual information has led, especially over the last two decades, to an increase in demand - in both academia and industry - for systems and algorithms capable of extracting information of interest from unstructured, semi-structured and fully structured textual data. This availability of data makes it possible, on the one hand, to carry out qualitative analyses of document collections in all research contexts and, on the other hand, to develop Apps with the most diverse objectives in the context of everyday life. Research activities in the field of text analysis have developed rapidly: many of Text Mining's approaches [1, 3, 4, 7, 20] effectively combine linguistic resources, computational methods and statistical techniques for the analysis of texts, representing a highly interdisciplinary field. In general, these processes do not involve only the training of the models, but also require numerous additional procedures, pre-processing of texts, transformation and reduction of the dimensionality of the data being analysed.

Among the many objectives that can be defined within a text analysis, of particular importance are the clustering techniques of documents on the basis of their similarity in terms of content, and more in detail the identification of the topics covered in the collection of documents [2, 11, 13]. In this regard, one of the most used methods in this context is topic modelling which, starting from a first work by Blei et al. [5] was developed by Griffiths et al. [8] introducing the Latent Dirichlet Allocation (LDA) as a generative model for identifying topics within a corpus. This method is sometimes overused within any type of context without the necessary adaptation of the analysis strategy to the characteristics of the corpus. As alternatives to LDA, multiple methodologies have been formulated for the exploration of topics, such as: Latent Semantic Analysis (LSA) [6]; the Reinert method [15, 16]; Non-negative Matrix Factorization (NMF) [9]; Correspondence and cluster analysis [10].

The goal of our paper is contributing at the broad debate on text analysis, as it is summarized by Lebart [11]. He focuses on the comparison of different techniques (NMF, LDA, Correspondence Analysis and Clustering) applied on a given middle size and homogeneous corpus, i.e. Shakespeare's 154 Sonnets.

In our paper we rely on a large original textual database created on a newsletter– issued daily by electrive.com on electric mobility. Russo et al. [18] have already analysed that data set (for the period 21th August 2018-15th September 2021) to identify the emerging topics in a domain of rapid transformation of the automotive industry. Entities disambiguation techniques, topic detection based on Correspondence and cluster analysis have been already commented by Russo et al. [18] who identify eight main classes of topics and 24 subtopics. In this paper, we update the database (with news item until 8th March 2022) and address the exploration of some clustering techniques.

In his analysis, Lebart concludes that the various methods "concur on the same topics […] despite the amazing variety of their theoretical backgrounds", and he underlines that results depend on "on various parameters and options", and that "exploratory or descriptive tools… have been essential to visualize the complexity of the process and to assess the obtained results" [13, p. 11].

In our paper, we refer to the expert classification (directly provided by the newsletter editors) and to three alternative methods for clustering/topic detection, based, respectively, on probabilistic, cluster-based and factorial methods: LDA; Reinert method; Correspondence analysis to select the most relevant factors explaining variability within the corpus, on which a hierarchical cluster analysis is applied. The rational for our choice is argued in the paper together with a discussion of the pros and cons of the various clustering techniques. Ad hoc visual tools for the comparative analysis have been created by using Tableau. Analytical methods to compare the results refer to the hierarchical cluster analysis as a benchmark.

Automatic analysis enables speed, consistency and reproducibility, and produces a systematic analysis of a comparative and contextual type, thus allowing to overcome the limitations of classifications and analyses based on the subjective opinion of whoever reads and classifies the texts one by one. On one hand, the limits deriving from the expert reading of large quantities of texts is generally overlooked, even though they can produce significant distortions with effects in interpretation of the results. On the other hand, the adoption of automatic techniques of topic detection/clustering process of text analysis must be characterized by transparency in the specification of the methods of analysis and in the interpretation of the results, favouring their reproducibility both to qualify their scientific character and to favour their use in a systematic way over time or for corpora with similar characteristics. Along this direction, the paper suggests some key challenges to be made explicit in adopting topic models.

Data and methods

1.1 Data

The data in analysis are composed by a collection of news published in English by electrive.com, a daily newsletter covering a wide range of relevant information on developments in electric transport in Europe, the USA and China. As an exploratory step, we analysed the data source "electrive.com", provided as a service offered online by a private publishing company (Rabbit Publishing GmbH). It covers a wide range of relevant information on the developments in electric transport, and its daily newsletter is not only made available on the website, but is also relayed on the main social media, including Twitter.

Using the Twitter API, tweets from September 12, 2018 to March 8, 2022 were downloaded from the timeline of the electrive.com Twitter page. Within each tweet, we identified the link to the news URL. From the news page, with a web data extraction procedure (web scraping), we extracted the following information of each item of news: title, full text, associated tags, category, date of publication and links to the information sources.

Of the ten categories proposed by electrive.com - Air, Automobile, Battery & Fuel Cell, Energy & Infrastructure, Fleets, Politics, Short Circuit, Two-Wheeler, Utility

Vehicles, Water - the major category, "Automobile", encompasses nearly 38% of the news, followed by "Battery and fuel cells" and "Energy & infrastructure", each with nearly 14% of the news items.

1.2 Methods

The first step to be able to proceed to the analysis of the texts consists in structuring the textual information in a lexical and textual database. This step was carried out using TaLTaC2 software.

The electrive.com corpus is composed of 5,216 news items (title and full text) published in the period 12/09/2018-08/03/2022 and consists of a vocabulary of 54,230 different words (i.e. types) for a total size of 2,175,691 word occurrences (i.e. tokens).

By means of grammatical tagging of the vocabulary words, it was possible to distinguish between the different grammatical types of words (structure words versus content words) and also to lemmatize them, i.e. to relate each word to its canonical form, resulting in a reduction of the forms under analysis. Furthermore, thanks to the use of a lexical-textual model [14], it was possible to recognize the multiword expressions present in the texts. The recognition of these forms yielded lexical analysis units with less semantic ambiguity.

Thanks to the specific characteristics of news writing, it was also possible to distinguish easily between common nouns and proper nouns. In fact, the news was clearly and carefully written; use of uppercase and lowercase allows to identify proper nouns (of people and companies) and acronyms (defined by all capital letters). It was also possible to recognize all the words identified by the electrive.com magazine as TAGs of individual news items. At the same time, all the types (simple and compound) referring to nations (and national adjectives) mentioned in the text were identified.

In order to classify the news items on the basis of their similarity in terms of content, only common nouns (simple words and multiword expressions) and adjectives were selected for each news item. A vector space model representation was then generated, in which each news item is defined as a vector composed of the selected keywords. In the next step the matrix <news × keywords> (5,125 × 8,489) has been analysed through the different methods selected in order to define the topics covered within the corpus.

In addition to expert classification, three methods for clustering/topic detection have been implemented: LDA, Reinert method (ALC), Correspondence analysis and Cluster analysis (CA) to select the most relevant factors explaining variability within the corpus, on which a hierarchical cluster analysis is applied[1].

To compare the results across methods, we refer to two criteria. First of all, we check for semantic coherence in topic models [12, 17, 19], by applying two measures of coherence: the UMass score, based on a log-conditional-probability measure, and

[1] The following libraries have been implemented in R: *topicmodels* (for LDA), *FactoMiner* (for correspondence analysis), *quanteda* and *rainette* (respectively for, preparing the dataset and elaborating the Reinert method).

a variant of the UCI metric, based on the normalized pointwise mutual information. Both are intrinsic measures based on the co-occurrences in the corpus of the 10 most important words defined in each topic[2]. Secondly, we elaborate a cluster heat-map, to compare the results obtained by the different methods and the degree of association between the topics. Cramer's V was also used to measure the strength of association between the classifications produced by the different methods. With both criteria we refer to a given number of topics that is defined with the CA method.

Results: discussion and further developments

The first result refers to the optimal number of topics obtained with each method, in comparison with the 10 categories defined by the expert classification, with the three largest groups – "automobile", "battery and fuel cells", "energy & infrastructure" – encompassing, respectively, 35.5% and 14.1%, 13,8% of all news items.

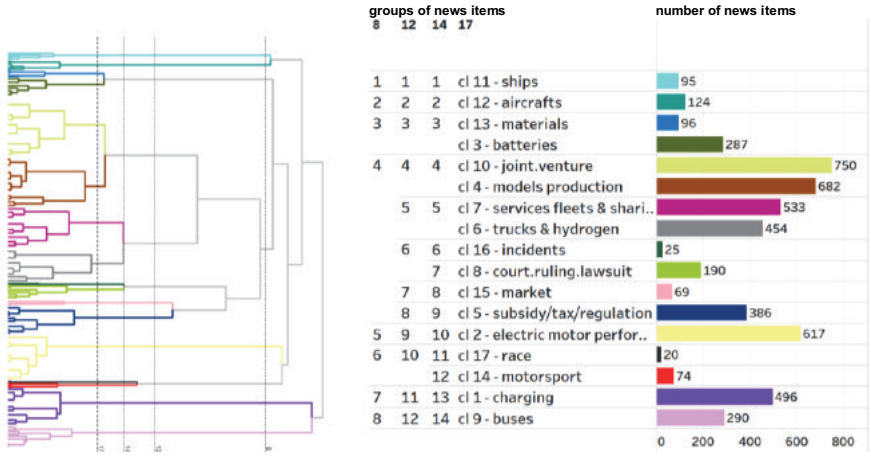
When considering the CA method, Figure 1 shows the dendrogram of the hierarchical clustering on the 10 factors of the correspondence analysis. The several cuts shown in the figure highlight the results from optimal number of clusters according to several methods (detailed results upon request). Our interpretation of results from an economic point of view indicates a cut at 17 clusters, which allows for a better disaggregation of the vast category "automobile" (split in the two groups of production differentiated with regard to features of economic organisation of production and a specific group describing electric motor performance), of the "battery &fuel cells" category (split in its components, respectively, of material and production), and of the" energy & infrastructure" (split in charging infrastructures vs. services). This number of 17 topics becomes the benchmark for all the other methods (details on optimal numbers are available upon request).

From the interpretative point of view of the topics encompassed in each cluster, the two methods that offer the hierarchical structure of texts of news items (CA and ALC) seem to be advantageous, since they allow us to define a greater or lesser number of topics, thus passing from a greater to a lesser detail, preserving the hierarchy of topics.

When moving on a more analytical comparison across methods of topic detection and documents classification, two key challenges must be addressed.

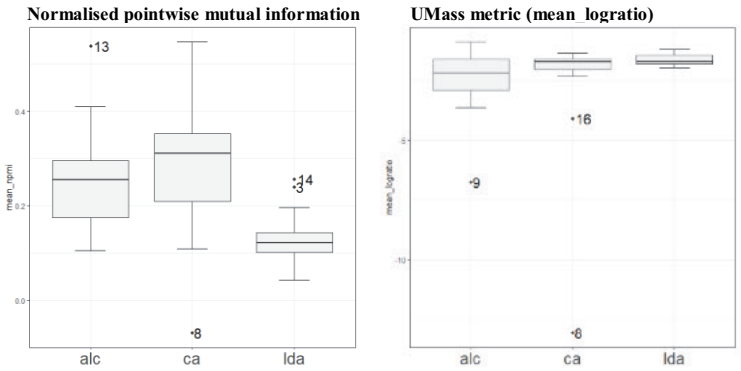
[2] The analysis in R uses the text2vec library. The selection of the terms identifying the topics is specific to the different methods used, respectively: test-value for CA; a chi-square test for ALC and the terms with the highest probability in the LDA.

Figure 1: Dendrogram: results of the CA method, with 8, 12, 14 and 17 groups



The first challenge concerns the ability of each algorithm to express semantically coherent topics. In this perspective, we implement two coherence indexes, both refer to a given number of topics that is defined with the CA method. Box plots in Figure 2 show the results of measures based, respectively, on normalized pointwise mutual information, NMPI, (left pane) and UMass score (right pane). According to NMPI, ranking of relative coherence shows the highest median value for CA method, with a high dispersion and cluster 8 as outlier that indeed has a miscellaneous of issues ("court.ruling.lawsuit"); with UMass score, LDA performs better, both in terms of median and overall topics, while in the case of CA method it highlights not only the case of cluster 8, but also of cluster 16, close by in the cluster.

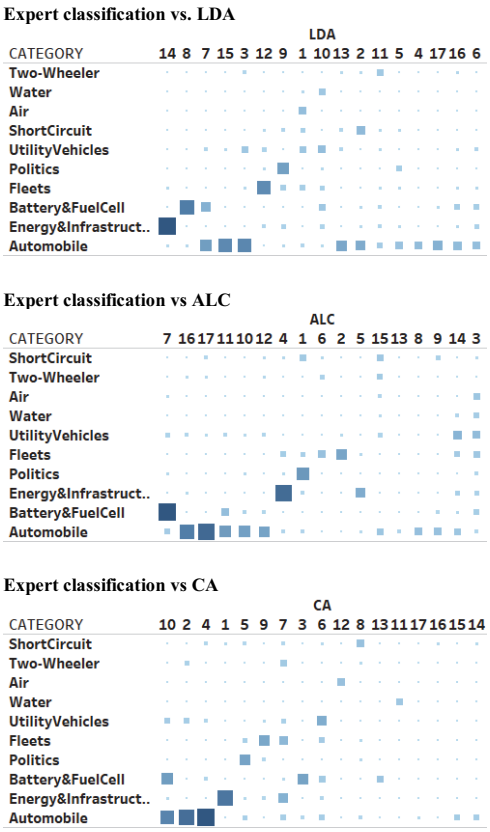
Figure 2 – Box plots of coherence indexes



The other challenge refers to the metrics that can be used in comparing the results obtained with topic models under analysis, in terms of document classification. In this paper we use heat-maps (Figure 3) to visualise the results obtained by pairs of methods in groupings news items (in the 10 expert categories and 17 groups). We can observe

that for some methods there are more areas of classification that overlap while in others overlapping is less significant.

Figure 2 – Expert classification (10 categories) and the three topic models under analysis (17 groups): heat-maps of relative correspondences



By comparing pairs of methods (results in Table 1) we observe a significant association between them (all p.values of the chi-squared test are <0.001). A summary measure of the strength of the association is provided by Cramer's V. It shows that CA is the methods that most closely approximate expert classification (67-68%), and is a confirmation of the superiority of the CA method in terms of readability of topics.

Table 1: Cramer's V indices for the three topic models

model.1	model.2	chi-squared	df	p.value	Cramer.V
CA	CATEGORY	21486.7	144	<0.001	0.6767
LDA	CATEGORY	14096.4	144	<0.001	0.5471
CA	ALC	24834.6	256	<0.001	0.5450
CA	LDA	24750.3	256	<0.001	0.5441
ALC	CATEGORY	12859.6	144	<0.001	0.5223
ALC	LDA	22410.6	256	<0.001	0.5174

Following Lebart [11] we intend to compare the results with other models (such as LSA and NMF), and to explore other models and methods for visualizing the comparative perspective on topic models and, in particular, Additive Trees, Self-Organizing Maps and Correspondence Analysis on the results of topic detection and clustering methods will be implemented. A third aspect to be explored is the analysis of the semantic similarity of the topics produced by the various algorithms. A fourt aspect concerns a general issue to be discussed, i.e. the specificity of cross-method results with respect to the characteristics of the corpus. In our database each document essentially deals with one topic and it would be important to discuss the comparison of topic models in cases of corpora with different structural features, in particular with regard to the variety of topics they might be include in each document.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012).
2. Allan, J.: Topic detection and tracking: event-based information organization. Springer Science & Business Media (2012).
3. Berry, M.W.: Survey of text mining. Computing Reviews. 45, 9, 548 (2004).
4. Berry, M.W., Kogan, J.: Text mining: applications and theory. John Wiley & Sons (2010).
5. Blei, D.M. et al.: Latent dirichlet allocation. Journal of machine Learning research. 3, Jan, 993–1022 (2003).
6. Deerwester, S. et al.: Indexing by latent semantic analysis. Journal of the American society for information science. 41, 6, 391–407 (1990).
7. Feldman, R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge ; New York (2007).
8. Griffiths, T.L. et al.: Topics in semantic representation. Psychological Review. 114, 2, 211–244 (2007). https://doi.org/10.1037/0033-295X.114.2.211.
9. Hassani, A. et al.: Text mining using nonnegative matrix factorization and latent semantic analysis. Neural Computing and Applications. 1–22 (2021).
10. Lebart, L. et al.: Exploring textual data. Springer, Dordrecht; London (1998).
11. Lebart, L.: Looking for topics: a brief review. In: Text Analytics, Advances and Challenges. pp. 215–223 Springer (2020).
12. Mimno, D. et al.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 conference on empirical methods in natural language processing. pp. 262–272 (2011).
13. Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In: Text Analytics. pp. 17–28 Springer (2020).
14. Pavone, P.: Automatic Multiword Identification in a Specialist Corpus. In: Tuzzi, A. (ed.) Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. pp. 151–166 Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-97064-6_8.
15. Ratinaud, P., Marchand, P.: Application de la méthode ALCESTE à de “gros” corpus et stabilité des “mondes lexicaux”: analyse du “CableGate” avec IRaMuTeQ. Actes des 11eme Journées internationales d’Analyse statistique des Données Textuelles. 835–844 (2012).
16. Reinert, M.: “Alceste” - une méthodologie d’analyse des données textuelles et une application: “Aurelia” de Gerard De Nerval. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique. 26, 1, 24–54 (1990). https://doi.org/10.1177/075910639002600103.
17. Röder, M. et al.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015).
18. Russo, M. et al.: Agents and artefacts in the emerging electric vehicle space. Int. J. Automotive Technology and Management. (2021).
19. Stevens, K. et al.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 952–961 (2012).
20. Sullivan, D.: Document warehousing and text mining. Wiley, New York (2001).

Businesses in industry: new challenges in sustainability, innovation, performance and competitiveness

Multidimensional assessment of Eco-Innovation and its link with Marketing Innovations.

Misurazione multidimensionale dell'ecoinnovazione e relazione con le innovazioni di Marketing.

Ida D'Attoma and Marco Ieva

Abstract Academics have studied the role of marketing innovations in leading to positive consequences for the environment. But, little is known on the enlargement of environmental benefits portfolio that can be achieved by marketing innovation. In this respect, we aim to study the environmental contribution driven by marketing innovation through an empirical analysis of the Community Innovation Survey in Germany related to 2012-2014. We construct an eco-innovation indicator using a PCA-based strategy. Then, we run a fractional regression where the eco-innovation indicator is function of the marketing innovations. Results show that innovation in placement yields an enlargement of the environmental benefits portfolio.

Abstract Alcuni studi esaminano il ruolo delle innovazioni di marketing nel portare benefici per l'ambiente. Tuttavia, è poco esplorato il legame tra l'innovazione di marketing e l'ampliamento di benefici per l'ambiente. L'obiettivo del lavoro è studiare il contributo in termini di benefici per l'ambiente veicolato dall'innovazione di marketing tramite un'analisi empirica basata sull'indagine comunitaria sull'innovazione per il triennio 2012-2014. E' proposto un indicatore di eco-innovazione ed è analizzata la sua relazione con le innovazioni di marketing tramite l'uso di un 'fractional probit'. E' emerso che l'innovazione in posizionamento porta ad un ampliamento dei benefici per l'ambiente.

Key words: Eco-innovation indicator, marketing innovation, fractional probit

[1] Ida D'Attoma, Department of Statistical Sciences – University of Bologna; email: Ida.dattoma2@unibo.it

Marco Ieva, Department of Economics and Management- University of Parma; email: marco.ieva@unipr.it

1 Introduction

In recent years a growing attention has been paid by governments, organizations and companies on environmental issues, such as the decrease of natural resources, the growth of the number of natural disasters and of pollution all over the world. The above factors have pushed consumers and companies to strive for a challenging balance between consumption requirements and sustainability [4]. In this context mobilizing industry for a clean and circular economy represents an important and necessary move to achieve the reduction of greenhouse gasses with the aim of fighting the increase in global warming. In this respect, marketing innovation could play a key role [2]. To the author knowledge, no studies have focused on understanding to which extent marketing innovations could really contribute to an enlargement of environmental benefits. Therefore, the aim of this paper is to evaluate the role of marketing innovation strategies on eco-innovation of manufacturing firms. Eco-innovation is defined as innovation that results in a reduction of negative environmental impact, no matter whether or not that effect is intended [11]. Specifically, a large body of literature used to assess it through a binary measure (e.g., [8]), but its use might cause the loss of valuable information regarding the eco-innovative intensity (e.g., [15]). Only recently, some scholars used a count measure to capture to some extent the simultaneous adoption of more types of eco-innovations (e.g. [5,3]). However, count measures can suffer from the double counting problem [10]. Hence, the present study addresses the abovementioned gap and measures the firms' eco-innovation by adopting a PCA-based analytic strategy to group together environmental benefit indicators. Marketing innovations are also considered both at the overall and individual level, adopting the widely employed 4Ps classification: innovation in product design and packaging, innovation in price, innovation in placement and innovation in promotion. To this aim, we use data for Germany collected via the Community Innovation Survey (CIS) carried out in 2014, as a special section on 'Innovation with Environmental Benefits' was present. The results of our study provide support for the role of marketing innovation overall considered, and when disentangled in 4Ps, results provide support for the role of innovation in placement in leading to an enlargement of benefits for the environment.

2 Data and variables

The dataset is based on the 2014 CIS survey for Germany. We focus on manufacturing sectors as they have the "potential to become a driving force for realising a sustainable society by introducing efficient production practices and developing products and services that help reduce negative impacts" [11 p.2]. The analysis has concentrated on

those innovative manufacturing firms which have obtained at least one environmental benefit from their technological and non-technological innovation activity (EI firms) in the period 2012-2014. The final sample is thus composed of 1137 firms. Literature on environmental innovation has developed a taxonomy of the different environmental benefits (EBs). EBs of innovation may origin from the production of a good or service, they may be related to after-sale consumption or use of a good or service by the end consumer [8].
The data structure follows the abovementioned taxonomy to identify the major EBs. EBs within the firm are identified as: reduced use of material (ECOMAT), reduced energy consumption for production (ECOENO), reduction in water soil pollution (ECOPOL), replacement of polluting materials with materials that are better for the environment (ECOSUB), replacement of fossil energy with renewable energy (ECOREP) and recycling of waste and material that is related to the production (ECOREC). External EBs that are obtained, due to innovations, during the consumption or use of goods and services from the end user are identified as follows: reduction of energy consumption from the end user (ECOENU), reduction of water, soil or air pollution (ECOPOS), recycling of product after usage (ECOREA) and increase of product life (ECOEXT).
The ten EBs are then used to build up the dependent variable of our model, while the independent variables of interest are a set of four dummies indicating whether the firm has introduced one of four marketing innovation activities that are classified, according to the OSLO Manual, in consistency with the 4Ps of the marketing: place, product, promotion and price. In addition to that other variables were included in the model as controls such as: regulatory push-pull factors, demand, technology conditions and firm-specific factors.

3 Methodology

The empirical analysis consists of two main stages. The first one is dedicated to the construction of a composite indicator of eco-innovation, while the second one investigates its link with marketing innovations controlling for other strategical, behavioral and structural variables.

3.1 First-stage: the construction of the eco-innovation indicator

We exploit the previously described environmental dimensions (EBs) to construct an indicator of Eco-innovation on each EI firm. It is based on individual indicators (the ten EBs) that focus on efforts and activities rather than on actual results. Each single EB indicator fails to provide an overall picture of the eco-innovation activity of a firm. By contrast, the micro-level composite Eco-innovation indicator here proposed addresses the sustainability at firm level in terms of all eco-innovation activities undertaken.
We built the Eco-innovation indicator following a widely adopted methodology [11] that involves three main steps: i) normalization, ii) multivariate analysis, iii) weighting

and aggregation. To consider the net contribution of each correlated indicator [1]to the composite one a principal component analysis (PCA) based strategy was conducted.
Let consider the sample of N elementary units (the EI firms). On each firm i, K primary indicators[2] EB_k are measured, K>1 (in our case K=10). Let E_{ik} be the value of the primary indicator EB_k for the firm i. The problem we try to solve is to define a unique numerical indicator for each firm - CEI_i- as a composite of the K primary indicators that keep track of the involvement in eco-innovation activities.
The variables exploited for the PCA are first transformed using the transformation proposed by [1], since this allows to take into consideration if a given EI firm is more or less focused on a single EB. Therefore, the rule is separately applied to the two groups of EBs: internal (EB_{int}) or external (EB_{ext}). In particular, following [1] each of the six (four) EBs experienced within the firm (by the end user) is divided by the total number of EBs reached by each EI firm within the firm (by the end user) as in eq. 1:

$$EB^*_{int,k,i} = \frac{EB_{int,k,i}}{\sum_k EB_{int,k,i}}$$
$$EB^*_{ext,k,i} = \frac{EB_{ext,k,i}}{\sum_k EB_{ext,k,i}} \quad (1)$$

By means of such a transformation the EI firms are represented on the basis of the predominance ascribed to internal or external EBs of their innovation activity. Higher values will emerge for more focused EI firms, while lower values will be observed for EI firms less focused but with more EBs.
PCA was then conducted on the set of these ten transformed variables[3] and the components with corresponding eigenvalue larger than 1 were retained, accounting for 72% of the original variability.
When using PCA to construct weights, the standard procedure is to use the eigenvector associated to the first component to serve as weights for the primary indicators ([7], [6]). However, it might not explain alone an adequate portion of the variance of the indicators, thus requiring more components to retain. Several scholars consider factor loadings of all the retained factors (e.g., [13], [14]) in order to preserve a larger proportion of the variation in the original data and here we follow such strand of literature. In particular, once the principal components have been retained, variables' weights were attributed by multiplying the contribution of each k-th EB indicator to the *m* most important components retained j – say L_{kj} —with their proportion of explained variance (λ_j) as in (2):

$$W_k = \sum_{k=1}^{K} \sum_{j=1}^{m} |L_{kj}| \cdot \lambda_j \quad (2)$$

[1] The tetrachoric correlation among EB indicator is medium-high. The matrix is available upon request.
[2] These indicators, before transformation, are all qualitative binary in our data.
[3] The transformed variables were standardized before PCA

with W_k as the weight of the k-th EB indicator, L_{kj} as the loading value of the k-th EB indicator on the principal component j and λ_j as the proportion of the explained variance of the j-th PC. Final weights were rescaled to sum up to one.
After variables' weighting, they were aggregated through a linear additive method: the composite indicator for each EI firm (CEI) resulting from the summation of the k weighted EB indicator as in (3):

$$CEI_i = \sum_{k=1}^{K} EB_{\text{ik}}\ W_k \tag{3}$$

where CEI_i is the composite Eco-innovation indicator of each EI firm, with $\sum_{k=1}^{K} w_k = 1$, $0 \leq w_k \leq 1$ for all k=1,…,K and i=1,…,N.
Finally, the CEI values, which can be either positive or negative, were normalized using the min-max normalization procedure. By doing so, the value of the indicator can range from 0 to 1, facilitating the interpretation. In particular, higher values of the indicator will correspond to a larger portfolio of EBs which can be interpreted as a high involvement in eco-innovation activities in terms of effort undertaken by firms in the direction of benefits for the environment.

3.2 Second-stage: the role of marketing innovation on CEI

The derived CEI indicator was used as outcome in a fractional response model ([12]) in order to analyse the four marketing innovation activities (place, product, promotion and price) in shaping firms' involvement in eco-innovation. In particular, the CEI indicator is a continuous variable bounded in [0,1] with the possibility of observing values at the boundaries. The fractional response regression model here adopted accomplished the dependent variable bounded in [0,1] and ensured that $E(Y|X)$ is also in [0,1]. Indeed, we model the mean of the dependent variable Y conditional on covariate X as follows (4):

$$E(y_i|x_i) = G(\boldsymbol{x_i\beta}) \tag{4}$$

with $[(x_i, y_i): i = 1,2, \ldots N]$ as the set of independent sequence of observations, $0 \leq y_i \leq 1$, N as the sample size and $G(\cdot)$ as a known function satisfying $0 < G(z) < 1, \forall\, z \in R$. This ensures that the predicted values of y lie in the interval [0,1]. Typically, $G(\cdot)$ is chosen to be a cumulative distribution function (cdf), with the two most popular examples being the logistic function and the standard normal cdf. We opted for the probit functional form for $G(\cdot) = \Phi(\boldsymbol{x_i\beta})$. The estimation procedure used is a quasi-likelihood method [12] where the log-likelihood function is defined as in (5):

$$lnL = \sum_{i=1}^{N} y_i ln\ G(\boldsymbol{x_i\beta}) + (1 - y_i)[1 - G(\boldsymbol{x_i\beta})] \tag{5}$$

4 Results

Our results, collected in Table 1, show that introducing a marketing innovation has an overall significant positive relationship with the eco-innovation indicator. However, when disentangled in the 4Ps marketing practices, they reveal that not all marketing innovation strategies are equally important for the environment. In particular only marketing innovation in placement leads to higher values of eco-innovation. Placement innovation involves the delivery of products to shops or to end consumers: delivery can be optimized to reduce pollution, and to generate EBs. Decentralized distribution of items and enabling local stores to manage recycled items are additional possible way of obtaining EBs. Findings are in line with previous studies suggesting the potential positive role of placement innovation towards the environment (e.g., [9]).

Table 1. The role of marketing innovation strategies on eco-innovation in Germany

MARKETING INNOVATIONS	Marginal effect (std. err.)	Z	p>|Z|
Overall	0.0193 (0.0085)	**2.25**	**0.024**
Packaging	-0.0067 (0.0055)	-1.20	0.231
Promotion	-0.0017 (0.0051)	-0.35	0.728
Price	0.0034 (0.0031)	1.09	0.274
Placement	0.0127 (0.0049)	**2.59**	**0.010**

Notes. Marginal effects are the derivatives of the conditional mean functions. They are averaged over firms. Coefficients of controls were not displayed due to space limitations. Source: own elaboration of CIS 2014 data.

5 Conclusions

Our work explored the link between the introduction of marketing innovations and the eco-innovation undertaken by firms measured in terms of the involvement in activities that lead to the achievement of environmental benefits. The eco-Innovation indicator here proposed cannot be considered an overall eco-innovation measure as, due to data availability, it does not consider the plurality of goals involved and the whole production process. It does not consider inputs (investments aiming at triggering sustainable activities), outputs (the immediate results of activities), socio-economic and resource efficiency outcomes. However, thanks to the CIS data, it is able to capture relevant and extensive information about eco-innovative activity. This information would be lost if binary or count type variables are used, but it is enhanced by our methodology. Results confirm the role of marketing innovation overall considered and extend the work from [2] by providing an additional overlook on the contribution of the four marketing innovation activities (place, product, promotion and price) to an enlargement of the innovation activities portfolio that leads to benefits for the environment.

Disclaimer

The anonymous data of the Community Innovation Survey 2014 used in the analysis of this paper was provided by EUROSTAT. All results and conclusions are given by the authors and represent their opinion and not those of EUROSTAT, the European Commission or any of the national authorities whose data have been used. The responsibility for all conclusions drawn from the data lies entirely with the authors.

References

1. Caravella, S., Crespi, F. (2020). Unfolding heterogeneity: The different policy drivers of different eco-innovation modes. Enviromental Science and Policy, 114: 182-193.
2. D'Attoma, I. and Pacei, S. (2020). The determinants of eco-innovation strategies. An empirical investigation of two European countries. In: Electronic Conference Proceedings of Sinergie-Sima Management Conference Grand Challenges: Companies and Universities Working for a Better Society. Pisa, pp. 247-254.
3. D'Attoma, I, and Ieva, M. (2022). The role of marketing strategies in achieving the environmental benefits of innovation. Journal of Cleaner Production, 342, 130957
4. Garcia-Granero, E.M., Piedra-Muñoz, L. and Galdeano-Gómez, E. (2018). Eco-innovation measurement: A review of firm performance indicators. Journal of Cleaner Production, 191, 304-317
5. Ghisetti, C., Marzucchi, A. and Montresor, S. (2015). The open eco-innovation mode. An empirical investigation of eleven European countries. Research Policy, 44/5, 1080-1093.
6. Greco, S., Ishizaka, A., Tasiou, M. and Torrisi, G. (2019). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Social Indicator Research, 141: 61-94.
7. Greyling, T., and Tregenna, F. (2016). Construction and Analysis of Composite Quality of Life Index for a region of South Africa. Social Indicator Research, 131(3), 887-930.
8. Horbach, J. (2016). Empirical determinants of eco-innovation in European Countries Using the Community Innovation Survey. Environmental Innovation and Societal Transitions, 19, 1-14.
9. Medrano, N. Cornejo-Cañamares, M. and Olarte-Pascual, C. (2020). The impact of marketing innovation on companies' environmental orientation. J.Bus.Ind.Market. 35(1), 1-12.
10. Nardo, M. et al. (2005). Handbook on constructing composite indicators. Paris: OECD Publishing.
11. OECD, 2010. Eco-innovation in industry. Enabling Green Growth, OECD Publishing, Paris.
12. Papke, L.E. and Wooldridge, J. M. (1996). "Econometric Methods for Fractional Response Variables with an Application to 401 (K) Plan Participation rates", Journal of Applied Econometrics, Vol. 11, 619-632.
13. Salvati, L. and Carlucci, M. (2014). A composite index of sustainable development at the local scale: Italy as a case study. Ecological Indicators, 43, 162-171.
14. Tapia, C., Abajo, B., Feliu, E., Mendizabal, M., Martinez, J.A. et al. (2017). Profiling urban vulnerabilities to climate change: An indicator-based vulnerability assessment for European cities. Ecological Indicators, 78, 142-155.
15. Triguero, A., Moreno-Mondéjar, L. and Davia, M.A. (2013). Drivers of different types of eco-innovation in European SMEs. Ecological Economics, 92, 25-33.

Circular Economy practices in the European SMEs: company-level and country-level drivers

Pratiche di Economia Circolare nelle PMI Europee: fattori determinanti a livello di impresa e di Paese

Francesca Bassi, Josè G. Dias, Nunzio Tritto

Abstract This paper studies the willingness of small and medium-sized companies (SMEs) in the European Union (EU) to undertake Circular Economy (CE) practices. The dataset comes from a survey involving more than 10,000 SMEs in the EU. This hierarchical dataset – companies within countries – was analyzed using a multilevel factor model that takes the heterogeneity between countries into account. Company-level variables and country-level covariates are inserted in the models. Both at company and country levels, there are factors that explain the attitude towards CE. Factor scores at both levels suggest a division between Western and Eastern countries (with some exceptions) regarding willingness to undertake CE activities by SMEs, which identify regional consequences of the EU policies towards CE.

Abstract *Questo lavoro studia l'intenzione delle Piccole e Medie Imprese (PMI) Europee di adottare pratiche di Economia Circolare (EC). I dati provengono da un'indagine su circa 10.000 PMI dislocate negli Stati membri dell'UE. Data la struttura gerarchica dei dati si sono utilizzati modelli di analisi multilivello, all'interno dei quali sono state inserite sia varabili di impresa che covariate di Paese. I risultati evidenziano la significatività di fattori ad entrambi i livelli nell'adozione delle diverse pratiche di EC. Essi suggeriscono inoltre una divisione (con alcune eccezioni) tra Paesi dell'Ovest e dell'Est dell'Europa per quel che riguarda l'attitudine alle pratiche di EC, che identificano la necessità di politiche eterogenee per favorire la transizione verso la sostenibilità.*

Key words: Circular Economy, sustainability, EU, SMEs, multilevel models

[1] Francesca Bassi, University of Padova; email: francesca.bassi@unipd.it
Josè G. Dias, ISCTE Lisbon; email: jose.dias@iscte.pt
Nunzio Tritto, University of Padova.

1. Introduction

In the last century, the quality of human life has exponentially improved thanks to science and technological innovation. However, this improvement has also led to several outcomes threatening the ecosystem equilibrium. In particular, the huge population growth caused resource insufficiency and the use of petroleum as source of energy has been associated to greenhouse effect. Scientific research documents that communities and individuals deal with climate change and many official institutions encourage a sustainable way of life. Many papers treat green behavior by consumers, others focus on big industries and their commitment to preserving our planet. In this context, minor importance has been given to small and medium-sized enterprises (SMEs), even if they are the engine of the world economy.

The introduction of the concept of Circular Economy (CE) can be traced at the end of the 20th century when seminal papers were published and it attracted the attention of many scholars (Lieder and Rashid, 2016). The terminology of CE was officially adopted in China in 2002, when the government approved the first law for CE promotion, which came into force in January 2009 (The Standing Committee of the National People's Congress of China, 2008). The main target was to reduce pollution and protect the planet. From this point on, worldwide institutions (including EU) unavoidably face these issues that might also bring competitive advantages to companies. The concept of CE has evolved in the world of business in an attempt to find a compromise between economic growth and environmental protection. This concept is in contrast with the most used idea of linear economy, i.e. take-make-use-dispose. Different definitions of CE exist, according to the field in which they are focused on (Lieder and Rashid, 2016). Thinking about eco-industrial development, CE may be defined as a creation of a closed-loop material flow in the entire economic system (Geng and Doberstein, 2008). According to the 3R principles (reduction, reuse, and recycling), the aim of CE is a circular (closed) flow of materials, use of raw materials, and energy through multiple phases (Yuan et al., 2006). In general, we can define CE "as a regenerative system in which resource input and waste, emission, and energy leakage are minimized by slowing, closing, and narrowing material and energy loops. This can be achieved through long-lasting design, maintenance, repair, reuse, remanufacturing, refurbishing, and recycling" (Geissdoerfer et al., 2017, p. 759).

The term sustainability is strictly related with the concept of CE. Indeed, sustainability is so much a broad topic that Johnston et al. (2007) found around 300 definitions. The study of Geissdoerfer et al. (2017) revealed that the concept of CE is seen as a condition of sustainability; consequently, there are several differences in terms of motivations, goals, and beneficiaries. However, the research found some common points, especially in the business world and in the effort to protect the environment. In the last years, the concept of sustainable development has been advanced, adding to the term sustainability a deeper conception of progress. In 2015, the UN established Sustainable Development Goals (SDGs), 17 targets to reach within 2030 in the perspective for a better future (United Nations, 2015). The European Union has committed to implement the SDGs both in internal and external policies.

This paper analyzes the willingness of European SMEs to undertake specific activities related to Circular Economy (CE) and to identify the potential drivers of this behavior. Data are collected from a sample of SMEs operating in the 28 EU Member States. Country-level characteristics are also included and their impact on the overall willingness to undertake CE activities is evaluated. The collected data is hierarchical: SMEs are nested into countries; this required the specification and the estimation of multilevel models.

SMEs are defined by the European Commission as companies with less than 250 employees and with an annual turnover that does not exceed 50 million euros, or a total annual balance that does not exceed 43 million euros (European Commission, 2003).

In 2015, more than 99% of EU enterprises could be classified as SMEs; they covered around two thirds (66.3%) of total employment (and the percentage is in continuous growth), and 55.8% of the total turnover (Papadopoulos et al., 2018). These numbers show the importance of SMEs in the European Union economy. At the same time, SMEs have a strong impact on the environment. In fact, it is estimated that about 60-70% of the total pollution is caused by them (Hoogendoorn et al. 2015).

The paper is organized as follows: Section 2 reviews the reference literature and steps forward the hypotheses that are tested. Section 3 describes the data. Section 4 introduces statistical methods. Section 5 presents the results and Section 6 concludes.

2. Literature review and hypotheses

Much research has been conducted on the identification of factors that may trigger and sustain the willingness of companies to promote CE. These factors can be classified into two categories: specific characteristics of the company that may play a role in undertaking CE activities and country-level factors, macro initiatives that either enhance or create barriers to the development of sustainable companies.

The size of the company affects the choice to undertake CE activities (Bianchi and Noci, 1998). SMEs are classified into: micro enterprises, with less than 10 employees and with an annual turnover or a total balance sheet lower than 2 million euros; small enterprises, with less than 50 employees and with an annual turnover or a total annual balance sheet lower than 10 million euros; and medium enterprises, with a number of employees between 50 and 250 and with an annual turnover between 10 and 50 million euros, or a total annual balance sheet between 10 and 43 million euros. Bigger companies have access to more resources, while smaller companies struggle with the absence of financial resources to invest even in simple sustainable activities, like building and handling recycling schemes (Hollins, 2011). Micro and small SMEs tend to be more active in terms of waste, recycling, and innovation. However, they can face difficulties, such as looking for funding, that can affect the implementation of CE activities, especially if the company is not directly interested in them (European Commission, 2015). On the other hand, in larger companies, ethics has a central role because they are more exposed and they have to save their reputation (Lawal et al., 2016). Thus, our first hypothesis states that:

H1.1: Larger SMEs are more willing to develop CE activities.

It has been shown in the literature that SMEs' age affects the will of implementation of CE practices, even if not with a linear relation (Hoogendoorn et al., 2015). Thus, our second hypothesis is:

H1.2: Oldest SMEs have more interest in undertaking a CE business model, also very recently founded SMEs show this behavior.

The will for SMEs to undertake sustainable activities as well as adopting attitudes towards green policies depends on the activity sector. Particularly, SMEs from more tangible sectors (manufacturing, construction, agriculture and waste management) are more prone to begin CE activities (Brand and Dam, 2009). In these sectors, the production process tends to generate more waste and it requires a greater quantity of raw materials; in addition, the process is rigidly screened by environmental parameters established by national and international institutions (Uhlaner et al., 2012). Thus, our third hypothesis states that:

H1.3: SMEs in more tangible sectors are more likely to make "green" investments.

The role in the production chain – Business-to-Business (B2B) and Business-to-Consumer (B2C) companies – may also lead to heterogeneous behavior with reference to CE activities. B2C companies have stronger motivations to apply sustainable activities than B2B ones; the formers sell products or services to final consumers so that they are exposed and must satisfy customers' needs to achieve competitive advantage (Källman, 2016). Thus, our fourth hypothesis establishes that:

H1.4: B2C SMEs are more willing to implement CE business models than B2B ones.

Investment in Research and Development (R&D) is fundamental to implement CE business models, i.e. without innovative technologies it is almost impossible to develop environmental sustainable ideas; for example, aggregate expenditure on R&D is very significant for a company that wants to apply sustainable actions such as reduction of CO_2 (Fernàndez et al., 2018). For these reasons, it is hypothesized that:

H1.5: An SME that invests more money in R&D is more willing to undertake CE activities.

The process of CE implementation is constrained by country-level factors that define different stages for the development of the concept of sustainability. In 1995 the Commission of Sustainable Development (CSD) created a set of indicators for studying the progress towards sustainability and established specific targets (Bartelmus, 1994). Based on the CSD approach, sustainable development contains four dimensions: social, economic, environmental and institutional (Spangenberg, 2002). The application of indicators at country level can be fundamental to understand

and solve problems of sustainable development (Diaz-Chavez, 2003). These country-level dimensions may impact the implementation of CE business model at company level. Our hypotheses with reference to country-level factors are the following:

H2.1: Country-level social dimension has a positive impact on the undertaking of CE activities.
H2.2: Country-level economic dimension has a positive impact on the undertaking of CE activities.
H2.3: Country-level environmental dimension has a positive impact on the undertaking of CE activities.
H2.4: Country-level institutional dimension has a positive impact on undertaking of CE activities.

3. Data

Data come from the Flash Eurobarometer 441 conducted in April 2016 and it contains 10,618 CATI interviews in the 28 countries of the EU. The number of interviews is almost the same in all countries (400), except for smaller countries. Complex survey weighting is taken into account to make the survey results representative of the EU population of SMEs. The data set contains five ordinal indicators on the implementation of CE practices in the past three years: 1. Re-plan of the way water is used to minimize use and maximize re-use; 2. Use of renewable energies; 3. Re-plan energy usage to minimize consumption; 4. Minimize waste by recycling or reusing waste or selling it to another company; and 5. Redesign products and services to minimize the use of materials or use recycled materials.

The survey asks the five items under the following question: “Has your company undertaken any of the following activities in the last 3 years?” The answer to each of the five items contains four ordinal categories: No, and we do not plan to do so, No, but we plan to do so, Yes, activities are underway, and Yes, activities have been implemented. The survey collects various company characteristics: the number of employees, total turnover in 2015, the age, the sector of economic activity, the type of goods and services sold, and the percentage of company’s turnover in 2015 invested in Research and Development.

All dimensions created by the CSD and described in Section 2 are useful to identify the country-level covariates. Despite several indicators for each dimension, the focus is on the most important ones related to CE and business world, according to the literature. Values of these variables are available on the Eurostat website. A number of indicators is available for each one of four the country-level dimensions; however, there exists a problem of multicollinearity among indicators related to the same dimension. For this reason, we selected only one variable per dimension: illiteracy rate (less than primary, primary, and lower secondary education), per capita GDP in euros, generation of waste per GDP unit (Kg per 1,000 euros) and corruption perception index (highly corrupted, very clean). We considered values of the covariates with reference to 2016.

4. Methods

In our models, the ordinal nature of the items is considered, data is weighted so to reproduce the distribution of SMEs in each country and the Maximum Likelihood (ML) method with Gaussian integration is used for parameter estimation.

Figure 1 summarizes the conceptual model. Two latent variables represent the willingness to undertake CE activities: one at company level (f^W) and one at country level (f^B). The H items Y_h correspond to the dependent variables of interest, the K variables X_k are the company-level covariates and the set of M variables Z_m are the country-level covariates. Estimation of this model allows to test the hypotheses presented in the previous section. The multilevel factor model (MFM) is an extension of confirmatory factor analysis (CFA) for hierarchical data: the first level uses company related variables to explain the latent construct while the second level measures the impact of country-level covariates. Since companies of the same country share characteristics, the assumption of independence is not valid and the nesting structure of the data must be considered (Costa and Dias, 2015).

In our two-level data, y_{ijh} denotes the response of company i in country j on item h on an ordinal scale. The MFM is usually estimated with the hypothesis of continuous observed variables. Being our variables of interest on an ordinal scale, a new continuous latent variable y^*_{ijh} is introduced, that is the propensity of company i in country j to be in category l of item h - according to the underlying variable approach (UVA). The relationship between the original variables and the latent variable y^*_{ijh} is determined by equation (1):

$$y_{ijh} = l \dots if \dots \pi_{h,l-1} < y^*_{ijh} \leq \pi_{h,l} \qquad (1)$$

where $\pi_{h,l}$ is the threshold of item h that separates the categories $l = 1, \dots, 4$ with $\pi_{h,0} = -\infty$ and $\pi_{h,5} = +\infty$.

The factor model at company level is given by equation (2):

$$y^*_{ijh} = \mu_{jh} + \lambda^W_h f^W_{ij} + \upsilon_{ij} \qquad (2)$$

where μ_{jh} is the random intercept of item h for country j, λ^W_h is the loading at company level for item h, and f^W_{ij} is the score of the latent variable at company level. Finally, υ_{ij} is the residual random variable, with distribution $\upsilon_{ij} \sim N(0, \sigma^2_w)$, where σ^2_w corresponds to the variability within groups.

The random intercept measures between-country variability and is given by equation (3):

$$\mu_{jh} = \mu_h + \lambda^B_h f^B_j + u_j \qquad (3)$$

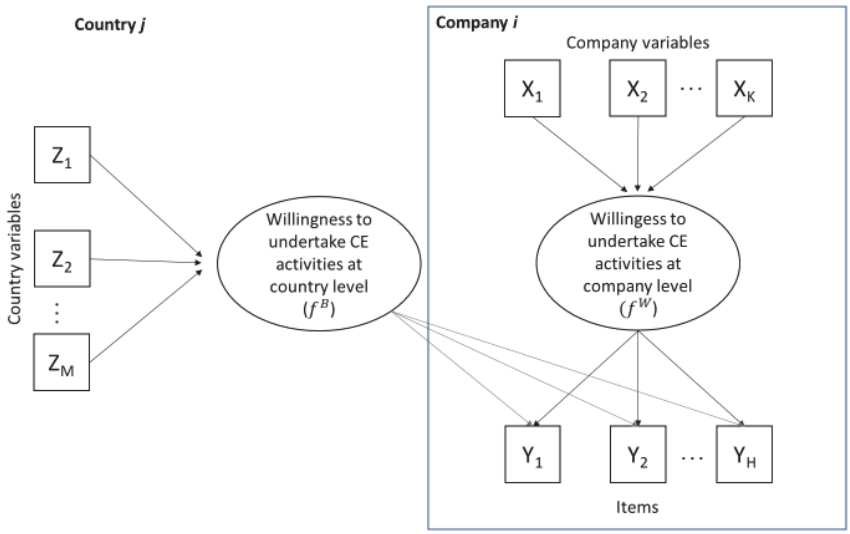


Figure 1: *Conceptual model*

where μ_h is the intercept for each item (set to zero for the thresholds), λ_h^B is the loading at country level for item h, and f_j^B is the score at country level. Finally, it is assumed that $u_j \sim N(0, \sigma_B^2)$, with σ_B^2 corresponding to the variance between groups. Residual random variables υ_{ij} and u_j are assumed to be independent. The two models (2) and (3) can be merged in a single equation.

The comparison of the factorial structures across distinct groups or population needs more attention (Moksnes et al., 2014). For example, for each country, all items are present; secondly, there is scale invariance because loadings, λ_h^B and λ_h^W, are defined as invariant across countries. Consequently, for all countries the addition of one unit in the latent variable has the same measure. The final assumption is that intercepts are invariant across countries (Dias and Trindade, 2016). To study the impact of the characteristics of the company on the latent variable at the individual level (f_{ij}^W), the MIMIC (Multiple Indicators and Multiple Causes) structure is applied.

In summary, the final model combines a Structural Equation Model (SEM) with observed and unobserved variables embedded in a multilevel structure. Model parameters are estimated using the Maximum Likelihood method with Mplus 6.1.

The intra-class correlation coefficient (ICC) is the proportion of variability related to different countries. It represents the correlation of two companies of the same country due to the fact that they share observed characteristics and some other non-directly observable values. If the value of the ICC is high, a great part of the variability is due to the different characteristics of the countries, then a multilevel approach is justified. Otherwise, with a low ICC, countries are not properly heterogeneous, then a hierarchical approach does not add value to the analysis. It was established that the minimum value of ICC to justify the multilevel approach is 0.05 (Hox et al., 2010).

5. Results

This section presents the results of hypotheses testing and heterogeneity analysis. First, a confirmatory factor analysis with ordinal variables is estimated, assuming a priori the presence of only one latent variable. Then we include the covariates through the estimation of a multilevel factor model. Two different models are estimated: the first one with covariates at company level (Model I), the second one with covariates at both levels (Model II). The comparison between these two models helps understanding the relevance of the information introduced using macro variables.

CFA with one latent variable and five items has a good fit to the data: the RMSEA has a value lower than 0.05 (0.039); indexes CFI and TLI are greater than 0.95 (0.977 and 0.953, respectively). Factor loadings at company- and country-level are all greater than 0.7 (Hair et al., 2010). At company level, the average weight assigned to the latent variable related to the willingness to undertake CE activities increases more than proportionally with the value of the variable representing CE activity of re-planning energy usage to minimize consumption. On the other hand, for the variables related to re-planning the way water is used, using renewable energy, minimizing waste, and redesign products and services, the increase is less than proportional. At country level, the only loading greater than 1 is related to the variable describing the action of minimizing waste. The variances of the latent variables at the company and country level are both statistically significant, accounting for the presence of variability within and between countries. As expected, the variance of the latent variable at the company level (2.427) is higher than at country level (0.471): heterogeneity within countries is higher than between countries. ICC is equal to 0.163, heterogeneity between countries amounts to 16.3% of the total variance.

Table 1: *Company-level covariates effects*

	Model I		**Model II**	
	Estimate	SE	Estimate	SE
Number of employees (ref. 1 to 9)				
10 to 49	0.329*	0.069	0.334*	0.069
50 to 250	0.620*	0.097	0.631*	0.097
Company's total turnover 2015 (ref. < 25,000 euros)				
[25,000-50,000) euros	-0.022	0.121	-0.025	0.120
[50,000-100,000) euros	0.010	0.106	0.003	0.003
[100,000-250,000) euros	0.094	0.121	0.086	0.118
[250,000-500,000) euros	0.257*	0.123	0.245*	0.120
[500,000-2,000,000) euros	0.308*	0.122	0.293*	0.120
[2,000,000-10,000,000) euros	0.418*	0.122	0.400*	0.120
≥ 10,000,000 euros	0.719*	0.165	0.700*	0.163
Company foundation (ref. before 1/1/2010)				
1/1/2010 to 1/1/2015	0.012	0.058	0.012	0.058
after 1/1/2015	-0.097	0.206	-0.101	0.207
Sector of activity (ref: Manufacturing)				
Retail	-0.239*	0.091	-0.242*	0.091
Services	-0.331*	0.098	-0.336*	0.098
Industry	-0.072	0.068	-0.075	0.068
Company sells (multiple choice)				

products directly to consumers	0.254*	0.058	0.253*	0.059
products to companies	0.18*	0.072	0.183*	0.072
services directly to consumers	0.488*	0.03	0.490*	0.053
services to companies	0.001	0.054	0.000	0.054
Company's total turnover 2015 invested in R&D (ref. < 5%)				
5%-9%	0.595*	0.076	0.596*	0.076
10%-14%	0.708*	0.057	0.709*	0.057
15%-19%	0.834*	0.166	0.834*	0.165
≥ 20%	0.666*	0.129	0.668*	0.129
Variance	1.962*	0.215	1.967*	0.216

*statistically significant

In general, it is possible to note that the differences between the two models (Table 1) in terms of the values of estimated coefficients and standard errors are negligible, as expected. The introduction of country-level variables in the model does not affect the estimates. The estimated variance does not vary significantly as a result of the introduction of the upper level variables in the model.

Values in Table 1 suggest that company size affects positively the willingness to undertake CE activities. Considering the number of employees, small (10-49 employees) and medium (50-249 employees) companies have a higher probability to undertake CE activities than micro ones (1-9 employees). Moreover, there is a strictly positive relationship between the number of employees and the latent variable: the highest slope refers to medium-sized companies. The relationship between company's turnover in 2015 and undertaking of CE activities is linear and positive as well. Thus, hypothesis 1.1, suggesting that larger SMEs are more willing to develop CE activities, is supported by the data. Undertaking CE activities does not depend on the age of the company. Thus, the hypothesis 1.2, stating that older and new SMEs have more interest in undertaking a CE business model, is not supported.

The implementation of CE practices is more prevalent in companies that belong to more tangible sectors. Table 1 displays that all slopes are negative, which means that companies in the manufacturing sector (the reference category) are the most active in undertaking CE activities. At the same time, the effect of the economic sector on the latent variable is the same as that of the manufacturing one, as the estimated slope is not significant. The category related to companies that sell services directly to consumers has the highest slope. The second highest and significant slope is related to companies that sell products directly to consumers. Consequently, companies that sell directly to consumers are more likely to undertake CE activities. The slope related to companies that sell services to companies or other organizations resulted not statistically significant. In conclusion, we can state that hypothesis 1.3 (i.e. SMEs from more tangible sectors are more likely to undertake CE activities) is partially supported, while hypothesis 1.4, B2C SMEs are more likely to implement CE business models than B2B companies, is supported.

Table 1 indicates a positive relationship between the percentage of turnover invested in R&D and the willingness to undertake CE practices: estimated coefficients are positive and significant. On the other hand, it seems that there is not a strictly positive relationship between them: the estimated coefficient related to the category from 15% to 19.9% is higher than that for the category 20% or more. Furthermore,

we can conclude that hypothesis 1.5 (SMEs which invest more in R&D are more likely to undertake CE activities) is only partially supported.

Table 2: *Country-level covariates effects*

	Model I		**Model II**	
	Estimate	SE	Estimate	SE
Illiteracy rate			0.024	0.007
GDP per capita (ln transformed)			0.230	0.318
Waste generation per GDP			-0.002	0.001
Corruption perception index			0.005	0.013
Variance	0.435	0.141	0.214	0.105
ICC	0.181		0.098	

Table 2 lists estimated coefficients for the two models without (Model I) and with covariates at country-level (Model II). We compare estimated coefficients with the hypotheses formulated in Section 2. Illiteracy rate has a significant and positive but low slope. This is probably due to the scale of the variable. We can state that an increase in illiteracy rate corresponds to an increase in the willingness to undertake CE activities, consequently, the social dimension has a negative impact and hypothesis 2.1 is not supported. Per capita GDP (log-transformed) has a slope with a positive sign, but it is not statistically significant, therefore, hypothesis 2.2, supposing that the economic dimension has a positive impact on CE activities, is not supported. The increase in the variable measuring generation of waste per GDP has a negative impact on the latent variable. On the other hand, the estimated slope shows a very low value: hypothesis 2.3 related to the environmental dimension is not supported. Finally, the Corruption Perception Index has a negative but not statistically significant impact on undertaking CE activities: hypothesis 2.4 (country-level institutional dimension positively affects undertaking CE activities) is not supported.

The most important focus of the analyses is directed to evaluate how these country-level variables explain the hierarchical structure of the model, through the change of the variances between groups and consequently of the ICCs. In the model without the upper level covariates, the between variance is equal to 0.435 and the ICC is 0.181. In this case, 18.1% of total variability of the latent variable – willingness to undertake CE activities – is due to the upper level of the data structure. As expected, in the model with country-level covariates, the between variability and the ICC are both lower (respectively 0.214 and 0.098). This means that the heterogeneity between countries is 9.8% of the total variance. In conclusion, the differences between these two estimates reveals that information from country-level variables explains heterogeneity between countries.

Comparing the estimated factor scores at both levels of the analysis, in order to study the impact of the company and country on the willingness to undertake CE activities it emerges all countries have a positive average willingness to undertake CE activities. However, ordering countries by increasing values of the mean of company-level factor scores, the following evidences emerge. Focusing on the first half of the list, we can notice that nine countries out of 14 are geographically located in the Eastern side of Europe (Bulgaria, Estonia, Slovakia, Latvia, Czech Republic, Romania, Lithuania, Hungary, and Poland). On the other hand, countries with higher

values are mostly developed countries in Central and Western Europe (Belgium, Portugal, France, the Netherlands, Finland, Luxembourg, Austria, Germany, and Denmark). Dispersion is high in Scandinavian countries (Finland and Sweden) and in United Kingdom. In general, we can conclude that differences between countries are not significant principally because of high dispersion.

Considering the estimated country-level factor scores, we can notice that the ranking above described is still valid, although with some countries that have completely different positions: for example, while from a company-level point of view Great Britain and Ireland have a lower value of the factor score, their estimates of factor scores at the upper level are the highest ones. This is another reason to justify the hierarchical structure of the model and the heterogeneity between countries. In conclusion, in both cases rankings are more or less as expected.

6. Conclusions

This study focused on the willingness of SMEs in the EU to undertake CE practices. The dataset comes from a survey involving more than 10,000 SMEs in the EU. The dataset provided five items about the possible activities related to CE, the first focus of the analysis was to synthesize the information of these variables creating a latent variable –willingness to undertake CE practices – through a CFA using ordinal data.

The next step was to analyze the hierarchical structure of the dataset with a multilevel factor model in order to study the heterogeneity between countries. We included covariates at the company and country-level and studied their impact on the latent variable, testing the hypotheses introduced in Section 2. While the slopes of the covariates at the first level are almost in line with the hypotheses (with the exception of the age of the company, which is not significant), at the upper level, none of the formulated hypotheses is supported. This was due to the strict ranges of some variables, which lead to small slopes, and other coefficients being not significant. On the other hand, the introduction of the macro-variables considerably reduces the intra-class correlation coefficient, which means that these variables give us important information about the differences between countries. We concluded that there might be other upper-level variables that can better explain the heterogeneity across countries.

Finally, we studied the factor scores at both levels, establishing that although there seems to be a division between Western and Eastern countries (with some exceptions), the differences between them are not significant. This conclusion is also supported by the non-significance of the country-level variables that show the EU countries tend to function as a homogeneous block at country level.

One possible development of this research involves the introduction of micro-variables related to company's perception about the access to information and resources for possible CE activities.

References

1. Bartelmus, P. (1994). Towards a Framework for Indicators of Sustainable Development. UN, New York.
2. Bianchi, R., Noci, G. (1998). "Greening" SMEs' Competitiveness. Small Bus. Econ. 11, 269–281.
3. Brand, M. J., Dam, L. (2009). Corporate social responsibility in small firms – Illusion or big business? Empirical evidence from the Netherlands. RENT 2009 Conference, Budapest, Hungary.
4. Costa, L. P., Dias, J. G. (2015). What do Europeans believe to be the causes of poverty? A multilevel analysis of heterogeneity within and between countries. Soc. Ind. Res. 122, 1-20.
5. Dias, J. G., Trindade, G. (2016). The Europeans' expectations of competition effects in passenger rail transport: A cross-national multilevel analysis. Soc. Ind. Res. 129, 1383–1399.
6. Diaz-Chavez, R. (2003). Sustainable Development Indicators for Peri-Urban Areas. A Case Study of Mexico City. PhD Thesis. EIA Unit IBS. University of Wales Aberystwyth: UK.
7. European Commission (2003). Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises 2003/361/EC. Official Journal of EU.
8. European Commission (2015). Closing the Loop – An EU Action Plan for the CE. Available on https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52015DC0614 (19 November 2019).
9. Geissdoerfer, P., Savaget, P., Bocken, N. M. P., Hultink, E. J. (2017). CE – A new sustainability paradigm? J. Clean. Prod. 143, 757–768.
10. Geng, Y., & Doberstein, B. (2008). Developing the CE in China: Challenges and opportunities for achieving 'leapfrog development'. Int. J. Sustain. Dev. World Ecol. 15(3), 231–239.
11. Hair, J., Black, W. C., Babin, B. J., Anderson, R. E. (2010). Multivariate Data Analysis (7th ed.). Upper saddle River, New Jersey: Pearson Education International.
12. Hollins, O. (2011). The Further Benefits of Business Resource Efficiency. Department for Environment. Food and Rural Affairs. London, UK.
13. Hoogendoorn, B., Guerra, D., van der Zwan, P. (2015). What drives environmental practices of SMEs? Small Bus. Econ. 44(4), 759–781.
14. Hox, J. J., Maas, C. J., Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. Stat. Neerl. 64(2), 157-170.
15. Johnston, P., Everard, M., Santillo, D., Robert, K. H. (2007). Reclaiming the definition of sustainability. Environ. Sci. Pollut. Res. Int. 14(1), 60-66.
16. Källman, M. (2016). Opportunities and barriers for CE-business models. Comparing conditions for rental in markets dominated by sales. Master Thesis in Sociology, University of Gothenburg, Gothenburg.
17. Lawal, F. A., Worlu, R. E., Ayoade, O. E. (2016). Critical success factors for sustainable entrepreneurship in SMEs: Nigerian perspective. Mediterr. J. Soc. Sci. 7, 338.
18. Lieder, M., Rashid, A. (2016). Towards CE implementation: A comprehensive review in context of manufacturing industry. J. Clean. Prod. 115, 36–51.
19. Moksnes, U. K., Løhre, A., Byrne, D. G., Haugan, G. (2014). Satisfaction with life scale in adolescents: evaluation of factor structure and gender invariance in a Norwegian sample. Soc. Indic. Res. 118, 657-671.
20. Papadopoulos, G., Rikama, S., Alajääskö, P., Salah-Eddine Z., Airaksinen, A., Luomaranta, H. (2018). Statistics on Small and Medium-sized Enterprises. Available on https://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_small_and_medium-sized_enterprises#General_overview (26 November 2019).
21. Spangenberg, J. H. (2002). Institutions for Sustainable Development: Indicators for Performance Assessment. Cologne, Austria: SERI Sustainable Europe Research Institute, 133–162.
22. The Standing Committee of the National People's Congress China. (2008). CE Promotion Law of the People's Republic of China. Available on http://www.lawinfochina.com/display.aspx?id=7025&lib=law (18 November 2019).
23. Uhlaner, L. M., Berent-Braun, M. M., Jeurissen, R. J. M., de Wit, G. (2012). Beyond size: Predicting engagement in environmental management practices of Dutch SMEs. J. Bus. Ethics, 109, 411–429.
24. United Nations (2015). Transforming Our World: The 2030 Agenda for Sustainable Development, New York.
25. Yuan, Z., Bi, J., Moriguichi, Y. (2006). The CE: A new development strategy in China. J. Ind. Ecol. 10, 4–8.

The employment effects of Italian Jobs Act. An ex-post impact evaluation.

Gli effetti occupazionali del Jobs Act. Una valutazione ex-post.

Alessandro Zeli♦ , Leopoldo Nascia♣

Abstract: In this paper we conducted an investigation of how the Jobs Act previsions affects employment. We estimate the impact of Jobs Act relief on social security contributions and the effect of new firing rules on employment using a large sample of Italian firms and by applying a two-step procedure: propensity score matching and Difference-in-Difference estimation. The outcomes of this model do not signal a strong effect of these measures both for the employment changes and for flexible workers changes. The employment changes seem to be benefited more from new dismission rules than from de-contribution incentives.

Sommario: Questo lavoro è finalizzato a capire quali siano stati gli effetti del Jobs Act sull'occupazione e sul lavoro flessibile. Si sono condotte, quindi, delle stime per quantificare l'impatto reale di due provvedimenti contenuti nel Jobs Act: la deduzione contributiva per i neo-assunti e le nuove regole sul licenziamento. A questo scopo si è applicata una metodologia a doppio stadio: il *propensity score matching* e il modello *Difference-in-Difference*. I risultati ottenuti dalle stime individuano deboli effetti di tali provvedimenti sulla dinamica dell'occupazione e del numero dei lavoratori flessibili impiegati. La dinamica dell'occupazione sembra beneficiare maggiormente dalle nuove regole sui licenziamenti piuttosto che dagli incentivi contributivi.

Keywords: Jobs Act, employment, flexible workers, propensity score matching, difference-in-difference.

♦ Correspondent author, Orcid: 0000-0003-0744-1557. Istat, Directorate for analysis and development of economic statistics, Rome, Italy. Email: zeli@istat.it.
♣ Istat, Rome, Italy. Email: nascia@istat.it.

1 Introduction

Employment tax deduction is one of the public economic policy tools most commonly used for growth and recovery. In many countries tax deductions are implemented to help firms hire new employees. A frequently recurring question related to employment tax deductions is whether they lead to the creation of jobs that otherwise would not have been created. This reflects the need to determine whether employment tax deductions generate additional effects on job creation. We contribute to this discussion by conducting a systematic investigation of how the Jobs Act previsions affects employment. In 2015 Italy adopted a new labor market policy: the so-called Jobs Act (JA) aimed at fostering employment and reducing costs in firing employees, in particular an important hiring subsidy was granted in the form of relief on social security contributions and new regulations lowering firing costs and making them less uncertain were approved, making open-end contracts more similar to fixed terms contracts.

Different studies have been conducted to determine the effect of social deductions and other JA provisions on employment. Among the others it deserves to mention the research of Sestito and Viviano (2016), they found, by using microdata for Veneto, that the new firing rules make the firms less reluctant to offer permanent contacts to new untested workers. Cirillo, Fana and Guariascio (2017) highlighted that monetary incentives matter a lot in explaining the dynamics of employment after 2015 and that new permanent contracts essentially came from transformation of flexible contract in permanent ones. The authors found that another effect of the new firing rules is the arise of part-time contracts.

Our contribution is to estimate the impact of Jobs Act relief on social security contributions on employment using a large sample of Italian firms and by applying a reduced form approach to empirical modelling. We explore the effects of de-contributions on employment growth from the following perspectives: we try to verify the permanent contract employment increase for beneficiaries, if this change is durable and what effects it had on decreasing in non-fixed term contracts employees. We, also, investigate the effect of new firing rules on employment.

To address these questions, we carried out a two-step procedure: propensity score matching and Difference-in-Difference (DiD) estimation. The methodology used in estimating differential effects permits to understand if the differential change of employment level is due to the social security de-contribution and if there are also effects for flexible workers and, finally, if the changes are permanent or decreasing over time.

2 The Jobs Act provisions

The Job Acts combined disposal is quite complex and it is structured on 8 Law Decrees concerning different aspects of labor market and labor law. In 2015 the Italian Government implemented a set of measures aimed to impact the labor market and improve employment. These provisions called Jobs Act (JA) were included in the Law Decree n.23/2015, in particular under the new JA regime we can highlight three main features. First the JA introduced a notable permanent hiring subsidy in the form of relief on social security contributions. Beneficiaries were granted with a total exemption of social security contribution for 2015 (but excluding compulsory employment insurance) for each worker hired with a permanent job contract and for a duration of three years. In 2016 they were granted for a reduction in social security contribution of 40 per cent. Another provision included in the JA was a reformation of the so-called Workers' Statute (*Statuto dei Lavoratori*) in particular the reform of one of the its most important part: the article 18. The article 18 provided for the reintegration into the workplace for workers unfairly dismissed. The new revision, on one hand, limits the dismissed worked reintegration to only few very particular cases, on the other hand, provides only a money compensation calculated on the basis of the worker length of service, starting from a minimum of 4 monthly wages up to 24 monthly wages, this mechanism was called "increasing protection" (*a tutele crescenti*). The aim of this provision is to increase the hiring tendency of the firm, and it is based on the assumption that firm have in mind a trade-off between hiring and

workplace protections (i.e., the more the firms have the possibility of firing the more the firms hire).
H1: The first research hypothesis is a classical test on the impact of a policy (in this case de-contribution) on the firms hiring behavior. In particular if there is positive employment change differential between beneficiary firms (treated) and non-beneficiary ones (non-treated). **H2:** The second research hypothesis consists in a test on possible additional employment due to the weakness of Article 18 and in particular to possibility to dismiss more easily workers even when hired with a permanent contract. In this case we can assume the hypothesis that firms, that in years before 2015 have always hired workers (treated), continue to hire workers, while firms that did not hire workers in the previous years began to hire after 2015, because of the possibility to dismiss the new hired workers (non-treated). In this case we have to verify a negative differential after 2015 between the employment changes of always growing firms and the others ones.

3 Methodology

The empirical problem was to evaluate whether there is a causal effect of exploiting JA provisions on firms' employment behavior. Following the approach and notation used by Blundell and Costas (2002), Bandik and Karpaty (2011) and Zeli (2018), we adopted a two-stage strategy: we first constructed a sample of matched beneficiary and non-beneficiary firms, and we then estimated a DiD coefficient regarding this matched sample. Let $TC \in \{0,1\}$ be an indicator of whether a firm i is treated (i.e., is exploiting JA social security relief or the JA possibility to easily dismiss the new hired workers) in a time period t, and let $y^1_{i,t+s}$ be the employment at time $t+s$; $s>0$, after the first de-contribution year. If firm i does not exploit the JA benefit, its outcome is denoted as $y^0_{i,t+s}$. The causal effect on employment of being an JA beneficiary for firm i at time t can be defined as:

(1) $$y^1_{i,t+s} - y^0_{i,t+s}$$

It is now possible to observe $y^1_{i,t+s}$ while $y^0_{i,t+s}$ is not observable, this is the primary problem in the estimation of causal effects. Therefore, it was possible to define the average effect of exploiting JA benefits as:

(2) $E\{y^1_{i,t+s} - y^0_{i,t+s} | TC_{it} = 1\} = E\{y^1_{i,t+s} | TC_{it} = 1\} - E\{y^0_{i,t+s} | TC_{it} = 1\}$

The last term of equation (2) is the counterfactual; the difficulty is now to construct this. In other words, we must estimate what the outcome in JA-exploiting firms would have been, on average, had they not exploited JA. Our approach implies to employ matching techniques. Matching involves pairing beneficiary with non-beneficiary firms with similar pre-provision characteristics X, that is, debt ratio, cash, sales growth, size, and class of economic activity. Using such techniques, we could build a sample of non-beneficiary twin firms to beneficiary firms to better approximate the non-observed counterfactual event in the equation (2) (Vella and Verbeek, 1999). We used the Rosenbaum and Rubin (1983) propensity score matching methodology, and used a Probit model to estimate the probability (or propensity score) of being a beneficiary firm, which was the first step toward implementing propensity score matching. In particular, the following equation (3) explicitly indicates the variables included in the Probit model.

(3) $p(TC_{it} = 1) = \beta_0 + \beta_1\, prod + \beta_2\, capital - labor\ ratio + \beta_3\, age + \beta_4\, average\ labor\ cost + \beta_4\, flexible\ intensity + \beta_5\, total\ assets + \delta_1 Industry + \delta_2 Territory$

TC_{it}=1 denotes a non-beneficiary firm in year *t-1* that benefits from JA provision in year *t*, X_{it-1} is a vector of relevant firm-specific variables in year t-1 that may influence

the firm's probability of being a beneficiary in the year t. D_j controls for other effects, such as industry or area effects.
Calculating and obtaining the propensity scores, after the Probit model estimation, made it possible to select the nearest control firms for which the propensity score determined the smallest distance from a treated firm. We utilized the Stata procedure PSMATCH2 (Leuven, Sianesi, 2003) to match treated and control firms. In order to identify the counterfactual, we adopted the nearest neighbor matching estimation method, with only one match per treatment observation and no replacement (aus dem Moore; 2014. According to Wooldridge (2002), this can be obtained by estimating the following regression:

$$(4)\ \Delta Empl_{t+s} = \beta_0 + \beta_1 TC_i + \beta_2 After_{t+s} + \beta_3 TC_i * After_{t+s} + \beta_4 X_i + \varepsilon$$

where Empl_{t+s} is the target outcome variable, TC is a dummy variable equal to 1 for beneficiary (treated) firms T, and equal to 0 for non-beneficiary firms C. It controls for constant differences between target firms and firms in the control group before the tax credit facilities. The dummy variable After_{t+s} takes the value of 1 in the post-tax deduction year *t+s* and 0 in the year before JA provisions introduction. This dummy variable captured the aggregate period effects that are common to the two groups T and C. The last term $\mathrm{TC_i} * \mathrm{After}_{t+s}$ represents the interaction between $\mathrm{TC_i}$ and After_{t+s}. The coefficient of this last term (β_3) represents the DiD estimator of the effect of being a beneficiary of treated firm T; in other words, the $\beta_3 = \gamma_{t+s}$. X_{t+s}, includes relevant covariates that explain the individual employment Empl_{t+s} , such as: profitability (GOS on value added), capital/labour ratio, sales, average cost of labor and the valued added growth by industry. To estimate the H1 and provide a measure of how much the employment reacts to the implementation of JA we modified the model as follows:

$$(5)\ \Delta Empl = \beta_0 + \beta_1 TC_i + \beta_2 After_{t+s} + \beta_3 TCa_i * After_{t+s} + \beta_4 X_i + \varepsilon$$
$$(5')\ \Delta Flex = \beta_0 + \beta_1 TC_i + \beta_2 After_{t+s} + \beta_3 TCa_i * After_{t+s} + \beta_4 X_i + \varepsilon$$

Dependent variables are the yearly change in employment and in flexible workers, while *TCa* is obtained by multiplying the previous variable TC by the amount of JA de-contributions (Angrist and Pischke, 2009). A positive and significant coefficient β_3 signals an effective impact of JA on employment.
As regards the H2 we estimated the equation (4) by using as dependents both the Employment yearly changes and the Flexible workers yearly changes. In this case a decrease in the employment gap between the treated and control group (i.e., a negative sign of β_3) may give a signal of an effective impact of the policy).

4 Data and variables

In this paper we exploited three main sources of microdata: the social security database OROS that includes data coming from all firms with dependents in Italy, the Istat register on business employment (ASIA *Occupazione*) and the balance sheet data of limited enterprises, a database acquired (by Istat) from the Chamber of Commerce. We only considered firms that remained in the panel in the period 2012-2017; hence we obtained a balanced panel of over 190,000 firms limited firms (a great part of limited firms has dependent workers and almost all limited firms are included in ASIA register). This panel is well representative of incorporated enterprises.

5 Results

In this section are presented first the result of propensity score matching both for social security de-contribution beneficiaries and for always growing firms. After are presented the results of DiD models.

5.1 Matched sample - Propensity score matching

Differences in the characteristics of beneficiary and non-beneficiary firms before deduction could bias estimates of the causal effects of access to the de-contribution, for instance. The reason is that it is difficult to distinguish whether firms' performances in those post de-contribution years is attributable to the de-contribution itself or to the fact that firms with a high performance tend to be beneficiaries. In order to overcome this problem, we applied a matching approach. We conducted propensity score estimation to match the samples of treated and untreated observations, with respect to all relevant firms' characteristics, using the 2014 data to separate the estimates from any possible anticipation effects.
Therefore, we estimated the propensity score, the conditional probability of requesting the benefit, by using the Probit model (3), and we then applied the propensity score matching method, as described above. In order to verify the quality of matching, t-tests for equality of means in the treated and non-treated groups, both before and after matching, were carried out: to ensure good balancing; these should be non-significant after matching. As shown in Table 1, almost all of the covariates were well balanced, so we can be confident that we obtained an effective control group both for de-contribution beneficiaries and non-beneficiaries and for always growing firms.

Table 1. Balance checking statistics – De-contribution beneficiaries and non-beneficiaries - Always growing firms and others

		De-contribution		***Always growing***	
	Unmatched	***t-test***		**t-test**	
Variable	***Matched***	***t***	***p>t***	**t**	**p>t**
Prod	U	44.2	0.000	18.4	0.000
	M	-1.0	0.330	-0.4	0.657
Capital to labor ratio	U	-6.0	0.000	12.2	0.000
	M	1.4	0.150	-0.6	0.532
Age	U	-18.9	0.000	-57.2	0.000
	M	4.2	0.000	-2.1	0.037
Average cost of labor	U	50.6	0.000	11.4	0.000
	M	0.0	0.972	0.0	0.994
Flex_int	U	8.8	0.000	-0.2	0.847
	M	2.9	0.004	-0.3	0.757
Total assets	U	90.7	0.000	32.6	0.000
	M	2.4	0.015	-2.5	0.013

A sub-sample consisting of only matching units was also considered, reducing the sample size to around 82,000 firms for de-contribution estimation and around 80,000 firms for new firing rules (always growing firms) These sets of firms was utilized to estimate the following DiD model.

5.2 DiD model – H1: Social security de-contribution effects

In order to study whether JA de-contributions had any effects on employment in the following years, we estimated the regression models in equation (5) and (5'). The dependent variables were the yearly change in employment and flexible workers at the firm level and the key estimate was the DiD estimator β_3. Table 2 presents the effects of JA on post deduction employment and numbers of flexible workers. The DiD estimator $\beta_3=inter=(TC*After_{t+s})$ for employment change (top panel) is positive, and indicates that, on average, JA de-contributions had a positive effect on employment in the years for which de-contributions were granted. However, the coefficient is weakly significant (10 per cent) an its value is quite low. If the estimated effects of interest remain substantially unchanged and significant after the inclusion of the individual specific trend, it indicates that we can accept the results obtained by the DiD procedure (parallelism assumption, Angrist and Pischke, 2009). When this individual specific trend is included in the FE base model, we can observe that the *inter* coefficient (our DiD effects) maintains its significancy and the value of the coefficient is substantially unchanged; this confirms the effects found with our model. Another question that we have to face is the possibility that the increases in employment were due to a positive economic cycle (i.e., an overall increase in GDP). To verify this hypothesis, we introduced the variable VAG (Value added growth) namely the yearly changes in value added by industry as calculated by National Account. However, the introduction of this variable in the base model neither yields significant parameter nor changes the value of *inter* coefficient. Hence, we can state that the little de-contribution effect is not affected by positive economic cycle.

Table 2. The effects of JA on post-de-contribution employment –the 2012–2017 panel. Std. Err. clustered in firm's code. Robust standard errors are reported in Italics. * Significant at 10%, ** at 5%, and*** at 1%. Significant interaction effects (β3) in bold. Firms controls variables are: profitability, capital to labor ratio, sales, average cost of labor. No of firms = 82,519.

	after_s		Inter		VAG	Individ. specific trend	Firms contr.	Industry, territorial and dim. dummies	intert+1		intert+2		Individ specific trend	Firms contr.	Industry, territorial and dim. dummies
D empl	-1.01	*	0.09	*			yes	yes	0.1	**	0.04	*		yes	yes
	0.58		0.05						0.04		0.02				
D empl	-1.05	*	0.09	*		yes	yes	yes	0.1	**	0.04	*	yes	yes	yes
	0.63		0.05						0.04		0.02				
D empl	-1.06	*	0.09	*	yes	yes	yes	yes						yes	yes
	0.64		0.05												
D flex	-0.22	***	0.02	***			yes	yes	0.05	***	0.03	***		yes	yes
	0.04		0						0		0				
D flex	-0.48	***	0.02	***		yes	yes	yes	0.04	***	0.02	***	yes	yes	yes
	0.05		0						0		0				
D flex	-0.48	***	0.02	***	yes	yes	yes	yes						yes	yes
	0.05		0												

In order to investigate the dynamic pattern of the post-deduction employment effects, the interaction variable for the whole post de-contribution period *inter* = (TCA_i * $After_{t+s}$) with year-by-year interaction variables in the fourth column was replaced i.e., $inter_{t+1}$ = (TCA_i * $After_{t+1}$) starting from the first year after deduction year onward (year-specific effects model) (Bandik and Karpaty, 2011). The coefficients on these interactions are significant (right grey-shadowed part of Table 2), beginning from the first year after the year of initial de-contribution, and they remain significant for the second year. We can note, however, a decrease in the value of the coefficient indicating a decrease in the effects of de-contributions. As regards the impact on flexible workers changes, we can observe (Table 2, lower panel) that, also in this case, the *inter* coefficient is positive and significant, even if it is more significant than the employment change one it presents a very tiny value signaling a little (very close to zero) effect of JA de-contribution on the flexible workers hiring. The *inter* coefficient maintains its significancy and value of the coefficient when the individual specific trend is included in the FE model. The dynamic effects present a decrease of de-contribution impact on number of flexible workers hired over time.

5.3 DiD model – H2: New dismissions rules effects

The same approach was carried out to analyze the JA new dismission rules on firms' employment behavior, we estimated the regression models in equation (4) both having the employment changes and flexible workers changes as dependents (Table 3).

Table 3: The effects of JA new dismission rules on employment and flexible workers – the 2012–2017 panel. Std. Err. clustered in firm's code. Robust standard errors are reported in Italics. * Significant at 10%, ** at 5%, and*** at 1%. Significant interaction effects (β3) in bold. Firms controls variables are: profitability, capital to labor ratio, sales, average cost of labor. No of firms = 80,797.

	after_s		Inter		VAG	Individ. specific trend	Firms contr.	Industry territ. dimens. dummies	intert+1		intert+2		intert+3		Individ. specific trend	Firms contr.	Industry, territ. dimens. dummies
D empl	0.88	***	-2.08	***			yes	yes	-1.00	***	-1.24	***	-1.40	***		yes	yes
	0.04		0.06						0.04		0.06		0.07				
D empl	0.88	***	-1.86	***		yes	yes	yes	-1.51	***	-2.09	***	-2.60	***	yes	yes	yes
	0.04	***	0.07						0.08		0.15		0.21				
D empl	0.89	***	-1.86	***	yes	yes	yes	yes								yes	yes
	0.04		0.07														
D flex	0.48	***	0.03				yes	yes	0.56	***	0.56	***	0.40	***		yes	yes
	0.02		0.03						0.02		0.03		0.03				
D flex	0.48	***	0.19	***		yes	yes	yes	0.57	***	0.57	***	0.41	***	yes	yes	yes
	0.02		0.04						0.03		0.03		0.04				

D flex	0.48	***	0.19	***	yes	yes	yes	yes					yes	yes
	0.02		0.04											

The DiD estimator *inter* (TC^*After_{t+s}) for employment change (top panel) is negative and highly significant, and indicates that, on average, JA new dismission rules seem have had a positive effect on the firms' employment behavior. In other words, the differential between the employment policies of firms which always had an occupational increase in the years previous of JA introduction and the employment policies of the others firms was being reducing. Also, in this case when the individual specific trend is included in the FE base model, the *inter* coefficient maintains its significancy and the value. On the contrary, when dynamics effects are being evaluated estimation yield coefficients quite different and this could make arise some doubts on the goodness of our model to estimate the per-year effects. As concerns the impact on flexible workers changes (Table 3 lower panel), we can observe that, in this case, the *inter* coefficient is positive but, even if it is significant, it presents a very tiny value, signaling a little effects of JA new firing rules on the flexible workers hiring. The *inter* coefficient varies notably its value when the individual specific trend is included in the FE model. The dynamic effects present a decrease of de-contribution impact on number of flexible workers hired over time.

6 Conclusions

In this paper we analyze the employment's effects of a part of JA provision introduced by the DL 23/2015 and the Financial Laws in 2015 and 2016 on limited companies. In particular we investigated two important provisions: the social security de-contributions granted for three years to companies hiring permanent workers in 2015 and 2016 and the new dismission rules for new hired workers. To estimate the effects of the provisions we utilized a well-tested in literature two-stages model: propensity score matching plus DiD model approach. The outcomes of this model do not signal a strong effect on employment of these measures both for the employment changes and for flexible workers changes. The employment changes seem to be benefited more from new dismission rules than from de-contribution incentives, as regards the flexible workers (the JA measures were aimed to convert the temporary contracts to permanent ones) there is not the presence of strong estimated coefficients (signaling a stabilization in the number of flexible workers hired in the period), however, no negative coefficients were found that would have indicated a large conversions from temporary to permanent contracts. More in-depth analysis should be carried out, first to complete the analysis of the provisions implemented by JA. Moreover, from methodological point of view a robustness check should also be carried out to confirm the results obtained here.

References

Angrist, J.A., Pischke, J.S.: Mostly Harmless Econometrics: An Empiricist's Companion. University Press, Princeton (2009)

aus dem Moore, N.: Corporate Taxation and Investment – Evidence from the Belgian ACE Reform. Ruhr Economic Papers n.534 (2014)

Bandik, R., Karpaty, P.: Employment Effects of Foreign Acquisition. Int. Rev. of Econ. Fin. 20(2): 211-224 (2011)

Blundell, R., Costa Dias, M.: Evaluation methods for Non-Experimental data. Fisc. Stud. 21:427-468 (2000)

Cirillo, V., Fana, M., Guarascio, D.: Labour market reforms in Italy: evaluating the effects of the Jobs Act. Econ. Pol. 34(2):211-232 (2017)

Leuven, E., Sianesi, B.: PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Boston College Department of Economics: Statistical Software Components. Downloadable from http://ideas.repec.org/c/boc/bocode/s432001.html. (2003)

Rosenbaum, P., Rubin, D.B.: The central role of propensity score in observational studies for causal effects. Biom. 70:41-55 (1983)

Sestito, P., Viviano, E.: Hiring incentives and/or firing cost reduction? Evaluating the impact of the 2015 policies on the Italian labour market. Occasional Papers-Banca d'Italia n.325 (2016)

Vella, F., Verbeek, M.: Two steps estimation of panel data models with censored endogenous variables and selection bias. J. Econ. 90:239-264 (1999)
Wooldridge J.M.: Econometric analysis of cross-section and panel data. MIT-press, Cambridge (MA) (2002)
Zeli A.: The impact of ACE on investment: the Italian case. Econ. Pol. (35)3 741–762 (2018)

Statistics for finance: new models, new data

The News–Jumps Relationship in the Cryptocurrency Market

La Relazione tra Notizie e Salti nei Prezzi nel Mercato delle Criptovalute

Ahmet Faruk Aysan, Massimiliano Caporin, Oguzhan Cepni, and Francesco Poli

Abstract We aim to decipher the relationship between price jumps and news sentiment in cryptocurrencies. Using one-minute coin-specific sentiment data, we detect jumps at intraday level and correlate them with news events through logistic regressions. We show that the release of information increases the jump probability, especially in the next one hour, and that topics limited to emotions, such as *optimism* and *anger*, and to market fundamentals are identified as possible jump causations. Jump sensitivity to news sentiment varies across coin characteristics, such as adoption (community vs. firm-driven) and market cap (big vs. small).

Abstract *Studiamo la relazione tra i salti nei prezzi delle criptovalute e il sentiment delle loro notizie. Usando dati su singole criptovalute a frequenza di un minuto, rileviamo i salti a livello infragiornaliero e li correliamo con la diffusione di notizie utilizzando regressioni logistiche. Dimostriamo che il rilascio di informazioni aumenta la probabilità di salti, specialmente nell'ora successiva, e che temi relativi ad emozioni e a fondamentali di mercato sono identificati come possibili cause di salti. La sensitività dei salti alle notizie varia a seconda delle caratteristiche delle criptovalute, come essere gestite da una comunità o da un'azienda e la capitalizzazione di mercato. Studiamo inoltre l'effetto del fine settimana in questa relazione.*

Key words: Cryptocurrency; jumps; jump spillover; logistic regression; news content; sentiment analysis.

Ahmet Faruk Aysan
Hamad Bin Khalifa University, Doha, Qatar, e-mail: aaysan@hbku.edu.qa

Massimiliano Caporin
University of Padova, via Battisti 241, Padova 35121, Italy, e-mail: massimiliano.caporin@unipd.it

Oguzhan Cepni
Copenhagen Business School, Porcelænshaven 16A, Frederiksberg DK-2000, Denmark, e-mail: oce.eco@cbs.dk

Francesco Poli
University of Padova, via Battisti 241, Padova 35121, Italy, e-mail: francesco.poli@unipd.it

1 Intro

Whether cryptocurrencies are eventually judged as financial innovations or speculative bubbles, they have attracted increased interest and scrutiny from market participants, policymakers, regulators, and investors. They continue to exhibit extreme volatility relative to fiat currencies, forcing market participants to monitor and study this new market closely.

In this paper, we analyze the relationship between price jumps and news sentiment on cryptocurrencies. We use comprehensive one-minute sentiment data on 16 coins from the Thomson Reuters MarketPysch Indices (TRMIs) database, which uses a proprietary algorithm that identifies news stories from several thousand traditional news and social media sources. TRMIs are based on advanced machine learning techniques that score online media sources specific to cryptocurrencies using specially-selected lexicon to utilize the content derived from news and social media. TRMIs are available at the cryptocurrency level and at high frequency. In contrast to previous media-based constructed sentiment indices, TRMIs capture the multiple dimensions of sentiment related to a spectrum of emotions, uncertainty, regulatory issues, and market fundamentals, rather than one single dimension of sentiment at aggregate level. We first detect jumps in cryptocurrency returns at the intraday level and then correlate their occurrence with TRMI-scored events through logistic regressions. We analyze the relevance of different sentiment themes for jumps in the coin returns. We also examine the relationship between sentiment and return jumps through the lens of cryptocurrency characteristics such as adoption (community vs. firm-driven) and market cap (big vs. small).

Combining the sentiment-themes (topics) with non-parametric identification of jumps, we are able to show that some content has a greater impact than others, or that "not all words are equal". In particular, we detect that sentiment on topics linked to emotions, such as *optimism* and *anger*, and on a more extensive set of topics related to market fundamentals, such as *market risk*, *price direction*, *price forecast*, and *volatility*, can be identified as the main trigger of price jumps. Our findings indicate that, while topics related to market fundamentals affect both positive and negative jumps, *optimism*, under the category *emotional*, is more closely related to the occurrence of positive jumps and *violence*, under the category *risks*, is more related to the occurrence of negative jumps. We note several additional results. First, we find evidence that the release of information, as monitored by the TRMIs, increases the probability of price jumps, especially within 60 minutes of the TRMI event happening. Second, we further examine in the cross-section how cryptocurrency characteristics determine the extent of the relation between jump occurrence and TRMI events. Our results suggest that the jump sensitivity to sentiment is higher for community-driven coins, such as Bitcoin, NEO, and Litecoin, where the communities significantly impact the coins' success and price direction. The reason for this might be that, unlike firm-driven coins, community members are more willing to post in social media channels to stimulate the success of particular projects [7].

2 The effect of news sentiment on jumps

We filter the returns from their intraweekly and intradaily periodicity using the Weighted Standard Deviation (WSD) estimator of [3], which has been shown to be robust to price jumps. To compute the WSD estimator, we consider only those days where at least 25% of the intradaily returns is not null. Subsequently, we compute the jump test according to eq. (19) of [1] for each intradaily interval using the WSD standardized returns. We set the significance level of the test to 0.001%.

We use logistic regressions to investigate how the jump intensity is linked to the multiple dimensions of sentiment characterizing the news and the volume of the information flow. To evaluate the role of information accumulation, we consider dynamic models where the occurrence and size of jumps and TRMIs are evaluated on different time windows:

1. Binary variables for jumps and TRMI occurrence

$$\pi_{j,t} = \alpha_j + \beta_{1,j}\mathscr{J}_{j,t-1} + \beta_{5,j}\sum_{i=1}^{5}\mathscr{J}_{j,t-i} + \beta_{30,j}\sum_{i=1}^{30}\mathscr{J}_{j,t-i} + \beta_{60,j}\sum_{i=1}^{60}\mathscr{J}_{j,t-i} + \\ + \gamma_{1,j}\mathscr{TR}^{l}_{j,t-1} + \gamma_{5,j}\sum_{i=1}^{5}\mathscr{TR}^{l}_{j,t-i} + \gamma_{30,j}\sum_{i=1}^{30}\mathscr{TR}^{l}_{j,t-i} + \gamma_{60,j}\sum_{i=1}^{60}\mathscr{TR}^{l}_{j,t} \quad (1)$$

2. Absolute size of jumps and absolute value of TRMI

$$\pi_{j,t} = \alpha_j + \beta_{1,j}J_{j,t-1} + \beta_{5,j}\sum_{i=1}^{5}J_{j,t-i} + \beta_{30,j}\sum_{i=1}^{30}J_{j,t-i} + \beta_{60,j}\sum_{i=1}^{60}J_{j,t-i} + \\ + \gamma_{1,j}TR^{l}_{j,t-1} + \gamma_{5,j}\sum_{i=1}^{5}TR^{l}_{j,t-i} + \gamma_{30,j}\sum_{i=1}^{30}TR^{l}_{j,t-i} + \gamma_{60,j}\sum_{i=1}^{60}TR^{l}_{j,t-i} \quad (2)$$

Where $\mathscr{J}_{j,t}$ is the binary variable detecting jump occurrence of coin j at time t, $J_{j,t}$ is the jump absolute size observed when $\mathscr{J}_{j,t}$ is not null, $\mathscr{TR}^{l}_{j,t}$ is the occurrence of an observation (i.e., new signed information is released) in TRMI l for coin j at time t, and $TR^{l}_{j,t}$ is the corresponding absolute value. $\pi_{j,t}$ is the logit for the dependent variable, i.e., the occurrence of a jump in coin j at time t. Both equations are considered for each coin $j \in \mathscr{I}$ and each TMRI $l \in \mathscr{L}$, where $\mathscr{I}$ and $\mathscr{L}$ represent the set of coins and the set of TRMIs, respectively.

We report a summary of findings with respect to the cross-section of estimates. Table 1, Panel A reports the effect of the TRMI occurrence and absolute value on the future jump occurrence in returns, equations (1) and (2). The categories of the TRMI indices and selected TRMIs—those associated to a *net* number of positive significant coefficients (significant positive minus significant negative) at least equal to 6 in column 5 (effect of occurrence of the index during the last hour on future jumps)—are reported in the first column, and the explanatory variables in the other columns.

Columns 2–5 report the TRMI results when the explanatory variables are the jump occurrence and the TRMI occurrence and columns 6–9 when the explanatory variables are the jump absolute value and the TRMI absolute value.

The entries in the table show, for each TRMI category, the percentage of significant positive and significant negative coefficients across all coins and, for each combination of TRMI index and explanatory variable, the number of coins, over a total of 16, for which the estimated coefficient is both positive and significant using a two-sided test at the 5% level ($^{+}$) and the number of coins for which the estimated coefficient is both negative and significant using a two-sided test at the 5% level ($^{-}$).

Firstly, we find evidence that both the occurrence and the absolute value of TRMIs increase the probability of jumps, especially in the 60 min after the TRMI occurrence (the percentage of coins for which the coefficients are positive and significant is, respectively, 24.7% and 13.2% for the occurrence and for the absolute value). This indicates that news sentiment is incorporated only slowly into prices, as investors react to news up to one hour after the event. We relate the delayed response of returns to news sentiment to the inattention hypothesis resulting from bounded rationality. There is extensive literature suggesting that investors have limited attention capacity and finite cognitive ability, and this prevents them from immediately processing the information in the news flow to update their portfolio positions [5, 4, 6, 2].

Secondly, we notice that the TRMI categories *Emotional* and *Market Fundamentals* are more important than *Innovative Aspect* and *Risks*. Hence, our results suggest that specific emotions might have different effects on the decision making process of individual investors.

We also consider other cross-type event interactions, such as how jump occurrence in returns affects the probability of future occurrences of TRMI events. In doing so, we aim at further uncovering the directions of the information flow between return and sentiment. Table 1, Panel B suggests that the occurrence of jumps increases the probability of TRMIs occurring, especially in the next 5, 30, and 60-minute time scale. The jump absolute size, as proxied by the absolute standardized return, is even more relevant than the jump occurrence in explaining future news releases and social media activity. Intuitively, this shows that extreme movements in returns lead to the formation of further sentiment since financial news releases stories about jump events that just occurred. Another possible explanation is that extreme returns can raise the attention of social media users, and news providers react by increasing their activity. For instance, if a positive jump is detected and traders start to chase positive returns (creating speculative responses), this situation could shape certain expectations about the market, thereby forming a positive sentiment.

Table 1 Summary results on the effect of TRMIs/jumps on jumps/TRMIs

Panel A: Effect of TRMIs occurrence on jumps occurrence - Dep. variable: logit of $\mathcal{J}$

	Regressors: TRMIs occurrence								*Regressors: TRMIs abs. value*							
	$\mathcal{TR}_1$		$\mathcal{TR}_5$		$\mathcal{TR}_{30}$		$\mathcal{TR}_{60}$		TR_1		TR_5		TR_{30}		TR_{60}	
All	5.1%$^+$	0.9%$^-$	5.4%$^+$	1.5%$^-$	9.3%$^+$	2.2%$^-$	24.7%$^+$	3.1%$^-$	8.1%$^+$	0.1%$^-$	9.7%$^+$	1.6%$^-$	5.8%$^+$	3.3%$^-$	13.2%$^+$	5.5%$^-$
Emotional	6.8%$^+$	2.3%$^-$	6.8%$^+$	0.6%$^-$	14.8%$^+$	2.8%$^-$	33.0%$^+$	5.1%$^-$	10.2%$^+$	0.6%$^-$	12.5%$^+$	1.7%$^-$	8.5%$^+$	2.8%$^-$	17.0%$^+$	8.5%$^-$
Market Fundam.	5.1%$^+$	0.0%$^-$	6.2%$^+$	2.8%$^-$	9.7%$^+$	1.7%$^-$	31.8%$^+$	1.1%$^-$	10.8%$^+$	0.0%$^-$	9.7%$^+$	2.3%$^-$	6.2%$^+$	4.5%$^-$	14.8%$^+$	6.2%$^-$
Innovative Asp.	4.2%$^+$	0.7%$^-$	4.9%$^+$	1.4%$^-$	6.9%$^+$	3.5%$^-$	13.9%$^+$	2.8%$^-$	6.9%$^+$	0.0%$^-$	6.9%$^+$	2.8%$^-$	5.6%$^+$	2.8%$^-$	11.8%$^+$	3.5%$^-$
Risks	2.8%$^+$	0.6%$^-$	3.4%$^+$	0.6%$^-$	3.4%$^+$	1.1%$^-$	16.5%$^+$	2.3%$^-$	2.8%$^+$	0.0%$^-$	6.8%$^+$	0.0%$^-$	2.3%$^+$	3.4%$^-$	9.1%$^+$	2.3%$^-$
sentiment	**3**$^+$ **0**$^-$		**1**$^+$ **1**$^-$		**5**$^+$ **0**$^-$		**7**$^+$ **2**$^-$		**4**$^+$ **0**$^-$		**6**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**2**$^+$ **3**$^-$	
optimism	**0**$^+$ **1**$^-$		**1**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**7**$^+$ **1**$^-$		**1**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**1**$^+$ **2**$^-$		**2**$^+$ **2**$^-$	
anger	**0**$^+$ **0**$^-$		**3**$^+$ **0**$^-$		**2**$^+$ **2**$^-$		**8**$^+$ **1**$^-$		**2**$^+$ **0**$^-$		**0**$^+$ **2**$^-$		**1**$^+$ **0**$^-$		**3**$^+$ **1**$^-$	
marketRisk	**2**$^+$ **0**$^-$		**3**$^+$ **1**$^-$		**2**$^+$ **0**$^-$		**7**$^+$ **0**$^-$		**3**$^+$ **0**$^-$		**3**$^+$ **0**$^-$		**3**$^+$ **3**$^-$		**4**$^+$ **2**$^-$	
priceDirection	**3**$^+$ **0**$^-$		**2**$^+$ **1**$^-$		**4**$^+$ **0**$^-$		**6**$^+$ **0**$^-$		**5**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**3**$^+$ **1**$^-$		**2**$^+$ **2**$^-$	
priceForecast	**1**$^+$ **0**$^-$		**3**$^+$ **0**$^-$		**0**$^+$ **0**$^-$		**9**$^+$ **0**$^-$		**1**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**0**$^+$ **1**$^-$		**3**$^+$ **1**$^-$	
volatility	**2**$^+$ **0**$^-$		**2**$^+$ **0**$^-$		**0**$^+$ **0**$^-$		**6**$^+$ **0**$^-$		**4**$^+$ **0**$^-$		**5**$^+$ **0**$^-$		**1**$^+$ **0**$^-$		**4**$^+$ **0**$^-$	

Panel B: Effect of jumps occurrence on TRMIs occurrence - Dep. variable: logit of $\mathcal{TR}$

	Regressors: jumps occurrence								*Regressors: jumps abs. size*							
	$\mathcal{J}_1$		$\mathcal{J}_5$		$\mathcal{J}_{30}$		$\mathcal{J}_{60}$		J_1		J_5		J_{30}		J_{60}	
All	4.4%$^+$	1.9%$^-$	20.3%$^+$	0.4%$^-$	27.3%$^+$	1.2%$^-$	30.7%$^+$	8.1%$^-$	4.2%$^+$	0.9%$^-$	17.3%$^+$	0.4%$^-$	9.4%$^+$	0.1%$^-$	60.0%$^+$	0.4%$^-$
Emotional	3.4%$^+$	2.3%$^-$	31.2%$^+$	0.0%$^-$	44.9%$^+$	1.7%$^-$	40.3%$^+$	12.5%$^-$	4.0%$^+$	0.6%$^-$	25.6%$^+$	0.6%$^-$	17.6%$^+$	0.0%$^-$	84.7%$^+$	0.0%$^-$
Market Fundam.	5.7%$^+$	1.7%$^-$	27.3%$^+$	0.0%$^-$	34.1%$^+$	1.1%$^-$	33.0%$^+$	9.1%$^-$	6.2%$^+$	0.6%$^-$	22.7%$^+$	0.0%$^-$	13.6%$^+$	0.0%$^-$	75.0%$^+$	0.0%$^-$
Innovative Asp.	2.8%$^+$	0.7%$^-$	10.4%$^+$	1.4%$^-$	10.4%$^+$	0.0%$^-$	22.2%$^+$	4.2%$^-$	3.5%$^+$	0.7%$^-$	8.3%$^+$	1.4%$^-$	2.1%$^+$	0.0%$^-$	26.4%$^+$	0.7%$^-$
Risks	4.5%$^+$	2.8%$^-$	8.0%$^+$	0.6%$^-$	13.1%$^+$	1.7%$^-$	24.4%$^+$	5.7%$^-$	2.3%$^+$	1.7%$^-$	8.5%$^+$	0.0%$^-$	1.1%$^+$	0.6%$^-$	44.3%$^+$	1.1%$^-$

Panel A: we estimated eq. (1) and (2) for each coin and for each TRMI index. The dependent variable is the log-odds of the jumps occurrence. Columns 2–5 report the results of eq. (1), and columns 6–9 report the results of eq. (2). The explanatory variables are reported on each column. The table reports, for each explanatory variable, the percentage of significant positive and significant negative coefficients across all coins and the following categorizations of TRMIs: all, emotional, market fundamentals, innovative aspect, and risks (the Buzz is not included). Selected TRMI indices (those associated to a *net* number of positive significant coefficients (significant positive minus significant negative) at least equal to 6 on colums 5) are reported on the first column. The entries of the table show, for each combination of TRMI index and explanatory variable: the number of coins, over a total of 16, for which the estimated coefficient is both positive and significant using a two sided test at the 5% level ($^+$), the number of coins for which the estimated coefficient is both negative and significant using a two sided test at the 5% level ($^-$); the cross-sectional median of the estimated coefficient.
Panel B: we estimated eq. (1) and (2) for each coin and for each TRMI index. The dependent variable is the log-odds of the TRMI occurrence. Columns 2–5 report the results of eq. (1), and columns 6–9 report the results of eq. (2). The explanatory variables are reported on each column. The table reports, for each explanatory variable, the percentage of significant positive and significant negative coefficients, using a two sided test at the 5% level, across all coins and the following categorizations of TRMIs: all (the Buzz is not included), emotional, market fundamentals, innovative aspect, and risks. For readibility, zero values are replaced with blank spaces.

3 The role of cryptocurrency characteristics on the news–jumps relationship

To better understand the variations in the impact of news sentiment on jumps, we split our set of coins into two groups: firm-based vs. community-based. In this regard, we aim to explain the sensitivity of the relationship between jumps and news sentiment controlling for cryptocurrency characteristics. Unreported results show that news sentiment matters more for community-driven coins. The reason for this might be that community-driven coins are more exposed to sentiment shocks through social media and investor blogs than firm-driven ones are.

We adopt a second grouping criterion, by sorting the coins based on their market cap, and then we classify those that appear above the top 25th percentile as *big* and similarly those below the bottom 25th percentile as *small*. The effect of TRMIs on jump occurrence are stronger for big coins compared to small coins. This might be due to the greater level of interest in big coins since they regularly receive mainstream media attention, and also, many individuals are likely to be unaware of the existence of small coins.

4 Conclusion

Our findings show that specific themes of the news sentiment are significantly related to the jump intensity and explain a significant fraction of variations in the jumps across cryptocurrencies. For instance, we find that sentiment on topics limited to emotions such as *optimism* and *anger* and on a more extensive set of topics related to market fundamentals such as *market risk*, *price direction*, *price forecast*, and *volatility* is identified as the main causation of price jumps. We also shed some light on the potential determinants of the cross-sectional news–jumps relationships, especially as they relate to cryptocurrency characteristics, by classifying them as firm-driven vs. community-driven and small vs. big coins. The results suggest that news sentiment matters more for community-driven and big coins.

Considering that jumps in asset prices are an essential input for many financial and economic decisions, such as derivatives pricing, volatility forecasting, and risk management, our study could enrich the modeling approach of return dynamics by explicitly incorporating news flows or indexes that summarize the informative content of news and social media while accounting, at the same time, for their sentiment.

References

1. Andersen, T.G., Bollerslev, T. and Dobrev, D.: No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and iid noise: Theory and testable distributional implications, Journal of Econometrics **138**, 125–180 (2007)
2. Barberis, N.: Psychology-based models of asset prices and trading volume, Handbook of Behavioral Economics: Applications and Foundations 1 **1**, 79–175 (2018)
3. Boudt, K., Croux, C. and Laurent, S.: Robust estimation of intraweek periodicity in volatility and jump detection, Journal of Empirical Finance **18**, 353–367 (2011)
4. DellaVigna, S. and Pollet, J.M.: Investor inattention and Friday earnings announcements, The Journal of Finance **64**, 709–749 (2009)
5. Hirshleifer, D., Lim, S.S. and Teoh, S.H.: Driven to distraction: Extraneous events and underreaction to earnings news, The Journal of Finance **64**, 2289–2325 (2009)
6. Louis, H. and Sun, A.: Investor inattention and the market reaction to merger announcements, Management Science **56**, 1781–1793 (2010)
7. Lu, C.-T., Xie, S., Kong, X. and Yu, P.S.: Inferring the impacts of social media on crowdfunding, Proceedings of the 7th ACM international conference on Web search and data mining , 573–582 (2014)

A weighted quantile approach to Expected Shortfall forecasting

Un approccio alla previsione dell'Expected Shortfall basato sui quantili pesati

Giuseppe Storti and Chao Wang

Abstract We present a novel semi-parametric Expected Shortfall (ES) forecasting framework. The proposed approach is theoretically motivated and is based on a two-step estimation procedure. The first step involves the estimation of Value-at-Risk (VaR) at different quantile levels through a set of quantile time series regressions. Then, the ES is computed as a weighted average of the estimated quantiles. The quantiles weighting structure is parsimoniously parameterized by means of a Beta weight function whose coefficients are optimized by minimizing a joint VaR and ES loss function of the Fissler-Ziegel class.The properties of the proposed approach are first evaluated with an extensive simulation study using two data generating processes. We then present the results of an application to a set of stock market indices in which the performances of the WQ estimation are compared to those of a range of parametric and semi-parametric models. The results of the forecasting experiments provide clear evidence in support of the proposed approach.

Abstract *Viene presentato un nuovo approccio semi-parametrico alla stima ed alla previsione dell'Expected Shortfall (ES). L'approccio proposto è teoricamente fondato ed è basato su una procedura di stima a due stadi. Il primo stadio richiede la stima del Value-at-Risk (VaR) a diversi livelli attraverso un set di regressioni quantiliche dinamiche. L'ES viene quindi calcolato come una media mobile dei quantili stimati. La struttura di ponderazione dei quantili viene parametrizzata in maniera parsimoniosa attraverso una funzione Beta i cui coefficienti sono ottimizati minimizzando una funzione di perdita congiunta per VaR ed ES, scelta nell'ambito della classe definita da Fissler-Ziegel. Le proprietà dell'approccio proposto vengono innanzitutto investigate attraverso un esteso studio di simulazione nel quale vengono considerati due distinti processi generatori dei dati. Quindi, vengono presentati i risultati di una applicazione ad un insieme di indici azionari nella quale la performance dell'approccio WQ viene confrontata con quella di un insieme di modelli*

Giuseppe Storti
University of Salerno, 84084 Fisciano (SA), e-mail: storti@unisa.it

Chao Wang
University of Sydney, NSW 2006 Sydney, e-mail: chao.wang@sydney.edu.au

parametrici e semi-parametrici. I risultati degli esperimenti di previsione forniscono chiara evidenza in favore dell'approccio proposto.

Key words: Value-at-Risk, Expected Shortfall, quantile regression, Beta weights, joint loss.

1 Extended abstract

The literature on ES modelling and forecasting is closely related to previous research on VaR. The dynamic quantile regression type model, e.g. the Conditional Autoregressive Value-at-Risk (CAViaR) model of Engle and Manganelli (2004), is a popular semi-parametric approach to estimate and forecast VaR. However, CAViaR type models cannot be used to directly estimate and forecast ES since the quantile loss is not consistent for the ES.

Fissler and Ziegel (2016) develop a family of joint loss functions (or "scoring functions") that are strictly consistent for the true VaR and ES, i.e. their expectations are uniquely minimized by the true VaR and ES series.

Patton et al. (2019) then propose new dynamic models for VaR and ES, through adopting the generalized autoregressive score (GAS) framework (Creal et al. 2013) and utilizing the loss functions in Fissler and Ziegel (2016). More or less at the same time, Taylor (2019) proposes a joint ES and quantile regression framework (ES-CAViaR) which relies on the Asymmetric Laplace (AL) density to build a likelihood function whose Maximum Likelihood Estimates (MLEs) coincide with those obtained by minimisation of a strictly consistent joint loss function for the couple (VaR, ES). In particular, under specific choices of the functions involved in the joint loss function of Fissler and Ziegel (2016), it can be shown that the negative of the AL log-likelihood function, presented in Taylor (2019), can be derived as a special case of the Fissler and Ziegel (2016) class of loss functions.

The frameworks in Taylor (2019) assume that the difference or ratio between VaR and ES follow specific dynamics, also in order to guarantee that VaR and ES do not cross with each other. Essentially, this implies additional assumptions on ES dynamics.

In this paper, a new ES estimation and forecasting framework is proposed where the ES is modelled as an affine function of tail quantiles. Hence, we refer to our approach as the *Weighted Quantile* estimator. The quantiles are produced from the CAViaR model of Engle and Manganelli (2004) by grid search of a range of equally spaced quantile levels below the target VaR level, i.e. 2.5%. For large grid sizes, the weighting pattern of the selected quantiles is based on a two parameter Beta weight function. The Beta weight function is a parsimonious but yet flexible choice and is able to reproduce a variety of different behaviours such as declining, increasing or hump shaped patterns. For less dense grids, direct estimation of the weights can be also entertained. We estimate the parameters of the Beta weight function by minimizing strictly consistent VaR and ES joint loss functions of the class defined in

Fissler and Ziegel (2016). In particular we focus on the AL loss in Taylor (2019). It is worth noting that the proposed estimator does not require any additional assumption on the dynamics of the ES process, but it only relies on the natural definition of ES as the tail expectation of the conditional distribution of returns, so reducing model uncertainty and risk of potential mis-specification on the ES. Our method has some interesting connections with the existing literature. First, there are some evident affinities between the WQ approach and the Conditional Autoregressive Expectile (CARE) models proposed by Taylor (2008). Namely, both our framework and CARE models involve a two-step estimation procedure and a grid search process. We show that our framework can produce more accurate ES forecasting results than CARE.

Further, the proposed framework has also some connections with the literature on forecasts combination. Taylor (2020) has recently proposed to use a forecast combination of different VaR&ES models of the same order. However, our strategy is substantially different since we are combining forecasts of a list of VaR models (CAViaR) of different quantile orders, instead of a list of different models.

The properties of the WQ approach are first evaluated with an extensive simulation study using two data generating processes. Two forecasting studies with different out-of-sample sizes are then conducted, one of which focuses on the 2008 Global Financial Crisis (GFC) period. The proposed models are applied to a set of stock market indices and their forecasting performances are compared to those of a range of parametric and semi-parametric models, including GARCH, Conditional AutoRegressive Expectile (CARE), joint VaR and ES quantile regression models and simple average of quantiles. The results of the forecasting experiments provide clear evidence in support of the proposed WQ approach.

References

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Engle, R. F. and S. Manganelli (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *J. of Bus. & Econ. Stat. 22*(4), 367–381.

Fissler, T. and J. F. Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics 44*(4), 1680–1707.

Patton, A. J., J. F. Ziegel, and R. Chen (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics 211*(2), 388 – 413.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics 6*(2), 231–252.

Taylor, J. W. (2019). Forecasting var and es using a semiparametric approach based on the asymmetric laplace distribution. *J. of Bus. & Econ. Stat. 37*(1), 121–133.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting 36*(2), 428–441.

Smooth and Abrupt Dynamics in Financial Volatility: the MS-MEM-MIDAS

Componenti Persistenti e Cambi di Regime nella Dinamica della Volatilità: il MS-MEM-MIDAS

Giampiero M. Gallo, Edoardo Otranto and Luca Scaffidi Domianello

Abstract In this paper we remark that the evolution of the realized volatility is marked by a combination between high–frequency dynamics and a smoother persistent dynamics evolving at a lower–frequency level. We suggest a new Multiplicative Error Model which combines the mixed frequency features of a MIDAS with Markovian dynamics. When estimated in–sample on the realized kernel volatility of the S&P500 index, this model dominates other simpler specifications, especially when monthly aggregated realized volatility is used. The same pattern is confirmed in the out–of–sample forecasting performance which suggests that adding an abrupt change in the average level of volatility better helps in tracking extreme episodes of volatility and a relative quick absorption of the shocks.

Abstract *L'evoluzione della volatilità realizzata è generalmente caratterizzata da movimenti ad alte frequenze e dinamiche più persistenti riferite a frequenze più basse. In questo lavoro proponiamo un nuovo Multiplicative Error Model che combina l'utilizzo di frequenze miste, tipiche di un MIDAS, con dinamiche di tipo markoviano. La stima in–sample della serie della volatilità realizzata dell'indice S&P500 dimostra la superiorità di questo modello su altre specifiche più semplici, soprattutto quando viene utilizzata la volatilità realizzata aggregata mensile. Lo stesso risultato è confermato in termini di previsione out–of–sample, che dimostra come l'aggiunta di un brusco cambiamento nel livello medio di volatilità aiuti a catturare episodi estremi di volatilità e un relativamente rapido assorbimento degli shock.*

Key words: Realized Volatility, Multiplicative Error Model, Markov switching, MIDAS, Short– and Long–Run Components

---

Giampiero M. Gallo
New York University in Florence, Italy, e-mail: giampiero.gallo@nyu.edu

Edoardo Otranto
University of Messina, Italy, e-mail: eotranto@unime.it

Luca Scaffidi Domianello
University of Messina, Italy, e-mail: lscaffidi@unime.it

1 Introduction

The recent global economic and financial crises have renewed the interest in studying the relationship between the real economy and the financial market volatility. Starting from [12] and [14], several authors document the economic sources of volatility and, in particular, the increase during a recession and the decrease during expansion phases (known as the *countercyclical* pattern of stock market volatility). [6] gave a new perspective within this strand of literature by introducing the GARCH-MIDAS model, a multiplicative component model in which the conditional variance is decomposed into short–run and long–run components. While the former follows a GARCH dynamics aimed at capturing volatility clustering and daily fluctuations, the latter represents a slow–moving average level of volatility, driven by macroeconomic and/or financial variables. Since different frequencies are involved (typically, monthly frequency of economic variables versus daily frequency of financial variables, possibly aggregated) another appeal of the model is the capability of mixing them in the same analysis.

[1] extended MIDAS volatility models to the class of Multiplicative Error Models (MEM, [4, 5]) suggesting a MEM-MIDAS. Due to its smooth pattern, the long–run component of the MEM-MIDAS model is not able to capture abrupt shifts in the average level of volatility and this suggests a further extension, in which a Markovian dynamics is added to the short–run and long–run component (we call it MS-MEM-MIDAS)[1]. We thus offer an insight on the contribution of variables observed at a lower frequency, when a Markov switching component allows for the sudden adjustment of the average level of volatility. As shown by an application on the realized kernel volatility of the S&P500 index, the MS-MEM-MIDAS offers improvements both in– and out–of–sample relative to a Markov Switching MEM (without the MIDAS component) and to a MEM-MIDAS without switching behavior. The paper is organized as follows: Section 2 describes the new model proposed, Section 3 illustrates the empirical analysis, with some concluding remarks following.

2 A New Model in the MEM Class

The MEM is a class of time series models for non–negative processes $\{x_t\}$ describing the evolution of phenomena related to financial market activity (e.g. volatility, durations [7], volumes [11], number of trades, etc.) that, in its asymmetric structure, is specified as follows:

$$\begin{aligned} & x_t = g_t \tau \varepsilon_t, \\ & \varepsilon_t \sim Gamma(a_1, 1/a_1) \quad \forall t \\ & g_t = (1 - \alpha_1 - \beta_1 - \gamma_1/2) + \alpha_1 \tfrac{x_{t-1}}{\tau} + \beta_1 g_{t-1} + \gamma_1 D_{(r_{t-1}<0)} \tfrac{x_{t-1}}{\tau}. \end{aligned} \tag{1}$$

[1] See [13] in a GARCH framework.

The specification in Eq. (1) implies that $\mu_t = g_t \tau$ is the expectation of x_t, conditional on the information set at the previous period, $\mathscr{I}_{t-1}$, i.e. $E(x_t|\mathscr{I}_{t-1}) = \mu_t$, given that the error term ε_t follows a *Gamma* distribution with a unit mean[2]. D is a dummy variable equal to 1 when the returns at time t is negative, 0 otherwise, and the coefficient γ_1 captures the so–called leverage effect, whereby a negative return impacts subsequent volatility more than a positive one. Moreover, to ensure the positiveness and the stationarity of the process, we apply the usual sufficient constraints: $\alpha_1 \geq 0$, $\beta_1 \geq 0$, $\gamma_1 \geq 0$ and $\alpha_1 + \beta_1 + \gamma_1/2 < 1$. Under the given stationarity, the unconditional mean is equal to τ.

In order to accommodate variables observed at different frequencies, let us now define a double time index for the variable of interest. With a slight abuse of notation, let $\{x_{i,t}\}$ be the same non–negative process, where now we isolate the i-th day within the low-frequency period t (be it a week, a month, or a quarter). The relevant conditioning set becomes then $\mathscr{I}_{i-1,t}$.

The MEM-MIDAS model is specified as a multiplicative component model:

$$\begin{aligned}
&x_{i,t} = g_{i,t}\tau_t\varepsilon_{i,t} \\
&\varepsilon_{i,t} \sim Gamma(a, 1/a) \quad \forall\, i = 1,...,N_t \quad and \quad t = 1,...,T \\
&g_{i,t} = (1 - \alpha_1 - \beta_1 - \gamma_1/2) + \alpha_1 \tfrac{x_{i-1,t}}{\tau_t} + \beta_1 g_{i-1,t} + \gamma_1 D_{(r_{i-1,t}<0)} \tfrac{x_{i-1,t}}{\tau_t} \\
&\tau_t = exp\left\{\omega_1 + \theta \textstyle\sum_{k=1}^{K} \varphi_k(\lambda_1,\lambda_2) X_{t-k}\right\} \\
&\varphi_k(\lambda_1,\lambda_2) = \tfrac{(k/K)^{\lambda_1-1}(1-k/K)^{\lambda_2-1}}{\sum_{j=1}^{K}(j/K)^{\lambda_1-1}(1-j/K)^{\lambda_2-1}}
\end{aligned} \tag{2}$$

where $g_{i,t}$, the short-run component[3], follows a unit mean MEM process and τ_t is a slow-moving component driven by a low frequency stationary variable, X_t. The MIDAS filter is based on $\varphi_k(\lambda_1,\lambda_2)$, a weighting function of the past K values of X_t, with the weights that sum up to one. This filter, based on the beta function, is quite flexible, allowing us to link variables sampled at a different frequency. We set $\lambda_1 = 1$ and $\lambda_2 > 1$, to ensure a monotonically decreasing pattern, as far as λ_2 increases, that is the most recent observations have more influence on the long-run component.

In order to allow for an abrupt shift in the average level of volatility, we suggest the novel Markov Switching MEM-MIDAS model (MS MEM-MIDAS) as a multiplicative model with several components where we add a Markovian dynamics[4]:

[2] In order to ensure the non–negativeness of x_t the error term is defined on a positive support. Parameters are identified by a 1 subscript in order to allow the comparison with other models presented below.

[3] Notice that when $i = 1$, then $(i-1,t) = (N_{t-1}, t-1)$.

[4] See [9] for a comprehensive description of the MS MEM.

$$
\begin{aligned}
&x_{i,t} = g_{i,t,s_{i,t}} \tau_{i,t} \varepsilon_{i,t} \\
&\varepsilon_{i,t} | s_{i,t} \sim Gamma\left(a_{s_{i,t}}, 1/a_{s_{i,t}}\right) \quad \forall\, i = 1, ..., N_t \quad and \quad t = 1, ..., T \\
&g_{i,t,s_{i,t}} = (1 - \alpha_{s_{i,t}} - \beta_{s_{i,t}} - \gamma_{s_{i,t}}/2) + \alpha_{s_{i,t}} \frac{x_{i-1,t}}{\tau_{i-1,t}} + \beta_{s_{i,t}} g_{i-1,t,s_{i-1,t}} + \gamma_{s_{i,t}} D_{(r_{i-1,t}<0)} \frac{x_{i-1,t}}{\tau_{i-1,t}} \\
&\tau_{i,t} = exp\left\{\omega_{s_{i,t}} + \theta \sum_{k=1}^{K} \varphi_k(\lambda_1, \lambda_2) X_{t-k}\right\} \\
&\varphi_k(\lambda_1, \lambda_2) = \frac{(k/K)^{\lambda_1 - 1}(1-k/K)^{\lambda_2 - 1}}{\sum_{j=1}^{K}(j/K)^{\lambda_1 - 1}(1-j/K)^{\lambda_2 - 1}}.
\end{aligned}
\tag{3}
$$

In this specification, coefficients in the short–run component depend on a regime represented by a discrete time latent variable, $s_{i,t}$ which varies as a first–order Markov chain at the higher frequency according to transition probabilities:

$$
P\{s_{i,t} = j | s_{i-1,t} = l\} = p_{lj} \quad \forall\, l, j = 1, \ldots, J, \tag{4}
$$

with p_{lj} the transition probability and J the number of states (with the usual constraints). In this model, also the low–frequency component is allowed to change within period t according to a constant $\omega_{s_{i,t}}$ which changes with the same regimes.

Notice that the short–run component suffers the path dependence problem, that is it depends on the whole history of the latent variable $s_{i,t}$, then we use the collapsing procedure adopted by [9], based on [10]:

$$
\hat{g}_{i,t,s_{i,t}} = \frac{\sum_{l=1}^{J} P\{s_{i,t} = j, s_{i-1,t} = l | \mathscr{I}_{i,t}\} \hat{g}_{i,t,s_{i,t},s_{i-1,t}}}{P\{s_{i,t} = j | \mathscr{I}_{i,t}\}}, \tag{5}
$$

i.e. by averaging the J^2 possible values of the short-run component $g_{i,t,s_{i,t}}$, with the weights equal to the corresponding filtered probabilities.

3 Empirical Analysis

We select as dependent variable the S&P 500 annualized Realized kernel volatility[5] (RV), while we choose three different low frequency variables: industrial production (IP) growth rate, monthly RV, and the Equity Market volatility (EMV) indicator for Macroeconomics News and Outlook of [2][6]. Overall, we estimate 8 models: the MEM, MS(3) MEM, the MEM-MIDAS and the MS(3)-MEM-MIDAS for the three low frequency variables indicated above. The first estimation period spans between 2, January 2003 and 31, December 2014, and the results are presented in Table (1).

The estimated parameters are in line with the previous studies: the coefficient θ (which translates the MIDAS filter of the low–frequency variable on the high–

[5] From the realized variance taken from Oxford-Man institute's Realized Library (https://realized.oxford-man.ox.ac.uk/data/download), we derive realized volatility as its annualized square root in percentage terms.

[6] The data for IP and EMV are available at https://fred.stlouisfed.org/series, while monthly RV is the aggregation of the daily RV for each month.

frequency one) is negative when the forcing variable of the long run component is the IP rate, the known *countercyclical* pattern of volatility, while it is positive when we consider monthly RV or EMV tracker. In addition, we have a value of λ_2 very high, especially for the financial variables, then the most recent observations of the low–frequency variable have more influence on the long-run component. For what concern the Markov Switching models, we consider three regimes (like [9]) that allows us to discriminate among low–, mid–, and high–volatility periods. By looking at the Figure 1 we can notice the correspondence between high–volatility regime and market downturn periods (the bankruptcy of Lehman Brothers in September 2008, flash crash in May 2010, and the credit rating downgrade of the United States sovereign debt in the second half of the year 2011).

The in–sample statistics reported in Table 2 show us that MS(3) MEM-MIDAS models are the best ones, especially that based on monthly RV, according to the Information Criteria and the statistical losses we use. Graphically, we can see that the models with Markovian dynamics offers a better pattern of the long–run component, that is the average level around which conditional volatility fluctuates (Figure 2). Finally, we conduct on out–of–sample exercise with a rolling window scheme for the period between January 2, 2015 and December 31, 2020, by generating the one step ahead forecasts for the year 2015 based on the first in–sample period, then shifting forward the estimation period by one year and re-estimating the model to produce the forecasts for the following year, and so on.

Comparatively speaking, the forecasting performance of the models can be evaluated by calculating the Diebold and Mariano (DM – [3]) test statistics, using QLIKE as the loss function of reference. In Table 3, we report the results for the in–sample period: as it often happens, the more parameterized models outperform the simpler ones: in particular, the MS(3)-MEM-MIDAS model with RV performs really well, beating all models, with indistinguishable performance relative to the one with EMV. When turning to out–of–sample (see Table 4), MIDAS models using industrial production fare poorly, overall, and the base MEM does not fare worse than some other richer models. The MS(3)-MEM-MIDAS with RV maintains the satisfactory performance (never dominated by others), showing that the addition of the Markov switching behavior adds relevant value in forecasting, although we find ourselves unable to reject the null of equal performance with respect to the MIDAS model with RV, the MS(3) MEM, as well as the MS(3)-MEM-MIDAS with EMV.

4 Conclusion

In this paper, we have introduced a new class of (asymmetric) Multiplicative Error Models which combine, on the one hand, the possibility of combining variables sampled at different frequency (MIDAS) and a Markovian dynamics (Markov switching). The novelty lies in particular in the behavior of the low–frequency component (monthly in the case of our application) which is allowed to assume different values according to the latent regime prevailing at the daily level. The forecast-

Table 1 Parameter estimates with standard errors in parentheses for the eight models considered for annualized realized kernel volatility. Sample: Jan. 2, 2003 – Dec. 31, 2014.

	MEM	MEM MIDAS IP	MEM MIDAS EMV	MEM MIDAS RV	MS(3) MEM	MS(3) MEM MIDAS IP	MS(3) MEM MIDAS EMV	MS(3) MEM MIDAS RV
α_1	0.119	0.107	0.094	0.096	0.041	0.038	0.044	0.027
	(0.016)	(0.016)	(0.015)	(0.015)	(0.023)	(0.016)	(0.014)	(0.019)
β_1	0.768	0.769	0.750	0.751	0.768	0.775	0.789	0.774
	(0.017)	(0.017)	(0.017)	(0.017)	(0.042)	(0.029)	(0.023)	(0.066)
γ_1	0.128	0.135	0.147	0.149	0.164	0.164	0.157	0.166
	(0.012)	(0.012)	(0.012)	(0.012)	(0.020)	(0.019)	(0.014)	(0.034)
α_2					0.000	0.000	0.000	0.000
					(0.000)	(0.000)	(0.000)	(0.000)
β_2					0.843	0.832	0.733	0.880
					(0.033)	(0.031)	(0.184)	(0.036)
γ_2					0.103	0.101	0.045	0.054
					(0.034)	(0.027)	(0.047)	(0.057)
α_3					0.155	0.051	0.054	0.181
					(0.040)	(0.039)	(0.165)	(0.081)
β_3					0.724	0.800	0.690	0.673
					(0.048)	(0.040)	(0.242)	(0.073)
γ_3					0.117	0.171	0.240	0.156
					(0.033)	(0.032)	(0.054)	(0.037)
ω_1	2.518	2.578	1.403	1.867	2.219	2.364	1.418	1.815
	(0.036)	(0.031)	(0.079)	(0.050)	(0.042)	(0.029)	(0.076)	(0.047)
ω_2					2.627	2.763	1.744	2.060
					(0.087)	(0.042)	(0.101)	(0.141)
ω_3					3.157	3.366	2.330	2.367
					(0.148)	(0.072)	(0.113)	(0.127)
a_1	6.874	7.020	7.203	7.196	7.655	7.540	7.488	7.568
	(0.276)	(0.244)	(0.202)	(0.201)	(0.305)	(0.235)	(0.210)	(0.344)
a_2					8.018	9.318	10.84	11.97
					(1.075)	(0.744)	(1.476)	(2.699)
a_3					8.012	7.456	7.695	6.436
					(0.728)	(0.644)	(1.065)	(0.800)
θ		-0.223	0.074	0.045		-0.207	0.066	0.038
		(0.051)	(0.005)	(0.003)		(0.019)	(0.005)	(0.004)
λ_2		2.691	5.824	9.934		2.770	4.490	7.610
		(1.850)	(1.102)	(1.737)		(0.352)	(0.742)	(2.676)
p_{11}					0.994	0.996	0.995	0.997
					(0.004)	(0.002)	(0.002)	(0.002)
p_{22}					0.973	0.977	0.956	0.968
					(0.010)	(0.009)	(0.017)	(0.017)
p_{33}					0.987	0.971	0.972	0.983
					(0.006)	(0.011)	(0.018)	(0.017)
p_{12}					0.006	0.004	0.005	0.001
					(0.004)	(0.002)	(0.002)	(0.004)
p_{21}					0.016	0.012	0.033	0.015
					(0.009)	(0.007)	(0.015)	(0.013)
p_{32}					0.013	0.029	0.028	0.017
					(0.007)	(0.011)	(0.018)	(0.017)
p_{13}					0.000	0.000	0.000	0.002
p_{23}					0.011	0.011	0.011	0.017
p_{31}					0.000	0.000	0.000	0.000

[a] The elements of the transition probability matrix p_{13}, p_{23}, p_{31} are derived as the complement to one of the sum of the other elements by row and hence do not have a standard error. To facilitate the comparison with the parameters of the other models, we reparameterize τ in Eq. (1) with $exp(\omega_1)$.

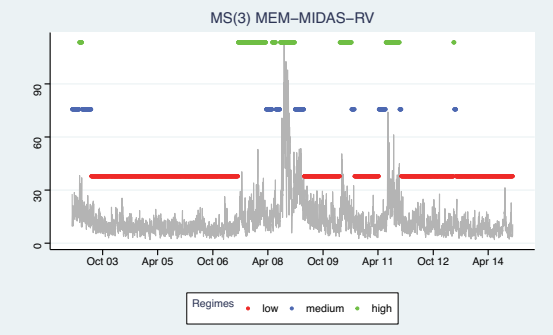


Fig. 1 Smoothed inference and annualized realized kernel volatility (gray line). Each day is assigned to a volatility regime based on the value of the smoothed probability.

ing performance appears to be satisfactory, especially in what concerns the MS(3)-MEM-MIDAS which uses monthly realized volatility as the driving variable for the low–frequency component.

In this version of the Markov switching MEM-MIDAS the Markov chain regulating the dynamics of the realized volatility is the same for both the low– and the high–frequency components. The extension to a specification allowing the Markovian dynamics to be different across components (along the lines of [8] in a different context) seems to be a venue to be investigated.

Fig. 2 Estimated conditional volatility of four models. Realized Volatility (gray line), conditional volatility (blue line), long–run component (green line). Volatility Proxy: annualized Realized kernel Volatility.

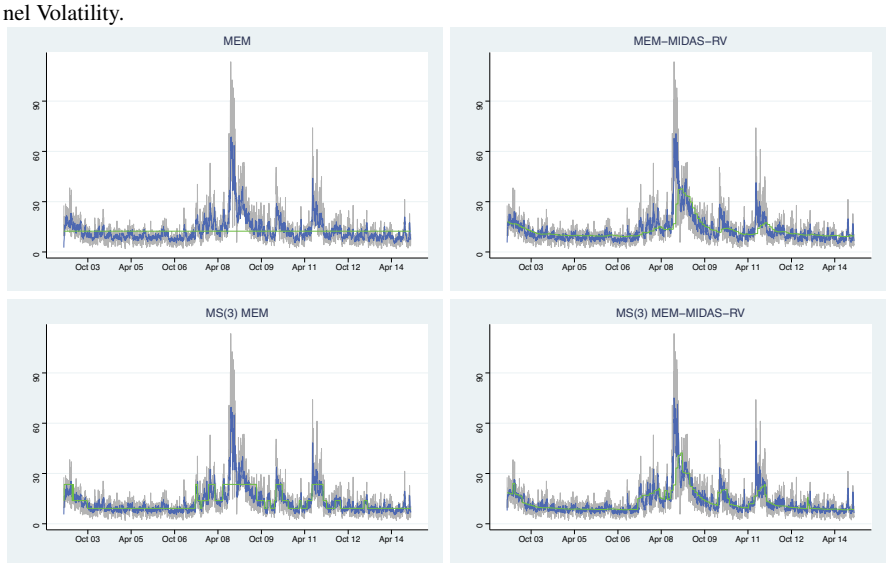


Table 2 In–sample performance of the estimated models: Sample: Jan. 2, 2003 – Dec. 31, 2014.

	MEM	MEM MIDAS IP	MEM MIDAS EMV	MEM MIDAS RV	MS(3) MEM	MS(3) MEM MIDAS IP	MS(3) MEM MIDAS EMV	MS(3) MEM MIDAS RV
LOGLIK	-8611.61	-8578.42	-8537.78	-8539.33	-8509.89	-8489.82	-8483.89	**-8477.97**
AIC	5.714	5.693	5.666	5.667	5.657	5.645	5.641	**5.637**
BIC	5.724	5.707	**5.680**	5.681	5.699	5.691	5.687	5.683
QLIKE	7.450	7.291	7.102	7.109	7.044	6.964	6.928	**6.912**
MSE	34.116	33.453	33.684	34.312	32.886	31.535	**31.094**	32.055

[a] Loglik: maximized value of the log-likelihood. AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion. QLIKE: Quasi-Likelihood function (multiplied by 100). MSE: mean squared error. Boldface for the best models by row.

Table 3 Diebold-Mariano test: p-value under the null hypothesis of equal performance of the in–sample forecasts. Sample: Jan. 2, 2003 – Dec. 31, 2014.

QLIKE	MEM	MEM MIDAS IP	MEM MIDAS EMV	MEM MIDAS RV	MS(3) MEM	MS(3) MEM MIDAS IP	MS(3) MEM MIDAS EMV
MEM-MIDAS-IP	0.013						
MEM-MIDAS-EMV	0.017	0.031					
MEM-MIDAS-RV	0.026	0.054	0.594				
MS(3)-MEM	0.010	0.019	0.137	0.118			
MS(3)-MEM MIDAS-IP	0.007	0.009	0.010	0.006	0.030		
MS(3)-MEM-MIDAS-EMV	0.006	0.006	0.002	0.001	0.411	0.016	
MS(3)-MEM-MIDAS-RV	0.005	0.006	0.001	0.000	0.007	0.085	0.333

[a] p-value of the Diebold and Mariano test. H_0 : QLIKE (row) = QLIKE (column); H_a : QLIKE (row) $<$ QLIKE (column). In red p–values < 0.1 (model by row "wins" against model by column); in blue p–values >0.90 (model by column "wins" against model by row).

References

1. Amendola, A., Candila, V., Cipollini, F., & Gallo, G. M.: Doubly multiplicative error models with long–and short–run components. Technical Report (2021).
2. Baker, S. R., Bloom, N., Davis, S. J., & Kost, K.: Policy News and Stock Market Volatility. NBER Working Paper No. 25720 (2019).
3. Diebold, F. X., & Mariano, R. S.: Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), pp. 253-263, (1995).

Table 4 Diebold-Mariano test: p-value under the null hypothesis of equal performance of the out of sample forecasts. Forecasting Period: Jan. 2, 2015 – Dec. 31, 2020.

QLIKE	MEM	MEM MIDAS IP	MEM MIDAS EMV	MEM MIDAS RV	MS(3) MEM	MS(3) MEM MIDAS IP	MS(3) MEM MIDAS EMV
MEM-MIDAS-IP	1.000						
MEM-MIDAS-EMV	0.764	0.014					
MEM-MIDAS-RV	0.018	0.000	0.003				
MS(3)-MEM	0.130	0.001	0.073	0.548			
MS(3)-MEM-MIDAS-IP	0.793	0.005	0.547	0.980	0.957		
MS(3)-MEM-MIDAS-EMV	0.092	0.000	0.014	0.452	0.411	0.016	
MS(3)-MEM-MIDAS-RV	0.023	0.000	0.004	0.165	0.113	0.000	0.142

[a] p-value of the Diebold and Mariano test. H_0 : QLIKE (row) = QLIKE (column); H_a : QLIKE (row) $<$ QLIKE (column). In red p–values < 0.1 (model by row "wins" against model by column); in blue p–values >0.90 (model by column "wins" against model by row).

4. Engle, R. F.: New frontiers for ARCH models. Journal of Applied Econometrics, 17(5), pp. 425–446 (2002).
5. Engle, R. F., & Gallo, G. M.: A multiple indicators model for volatility using intra-daily data. Journal of Econometrics, 131(1-2), pp. 3–27 (2006).
6. Engle, R. F., Ghysels, E., & Sohn, B.: Stock market volatility and macroeconomic fundamentals. The Review of Economics and Statistics, 95(3), pp. 776–797 (2013).
7. Engle, R. F., & Russell, J. R.: Autoregressive conditional duration: A new model for irregularly spaced transaction data. Econometrica, 66(5), pp. 1127–1162 (1998).
8. Gallo, G. M., & Otranto, E.: Volatility spillovers, interdependence and comovements: A Markov Switching approach, Computational Statistics & Data Analysis, 52(6), pp. 3011-3026 (2008).
9. Gallo, G. M., & Otranto, E.: Forecasting realized volatility with changing average levels. International Journal of Forecasting, 31(3), pp. 620–634 (2015).
10. Kim, C.-J.: Dynamic linear models with Markov-switching. Journal of Econometrics, 60(1-2), pp. 1–22 (1994).
11. Manganelli, S.: Duration, volume and volatility impact of trades. Journal of Financial Markets, 8(4), pp. 377–399 (2005).
12. Officer, R. R.: The variability of the market factor of the New York Stock Exchange. the Journal of Business, 46(3), pp. 434–453 (1973).
13. Pan, Z., Wang, Y., Wu, C., & Yin, L.: Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. Journal of Empirical Finance, 43, pp. 130–142 (2017).
14. Schwert, G. W.: Why does stock market volatility change over time? The Journal of Finance, 44(5), pp. 1115–1153 (1989).

The tail index and related quantities for volatility models

Indice di coda e grandezze associate per modelli di volatilitá

Fabrizio Laurini

Abstract The tail index provides useful information for assessing the speed of decay of models for rare events. Related to the marginal tail behavior there is associated also the extremogram, a conditional probability that an extreme event occurs some lags after one big value has been recorded. When stochastic volatility models have regularly varying tails there is a convenient manipulation that allows to exploit a variety of result to obtain Monte Carlo methods to derive the tail index. We present a simplified way to derive the tail index even for high order GARCH processes with infinite marginal variance.

Abstract *L'indice di coda fornisce informazione utile alla comprensione della velocità con cui alcuni modelli gestiscono gli eventi rari. Abbinato all'indice di coda c'è anche l'estremogramma che indica la probabilità condizionata di osservare un evento estremo qualche lag dopo aver osservato un primo evento estremo. Quando un modello ha code a variazione regolare ci sono manipolazioni utili che permettono di ottenere, via Monte Carlo, l'indice di coda in modo piuttosto agevole. Viene presentato un insieme di metodi per ottenere l'indice di coda nel caso di modelli GARCH di ordine elevato, anche nel caso in cui la varianza non esista.*

Key words: Regular variation, Til index, Volatility models

1 Introduction

Volatility models for risk management are designed to handle log-returns, defined as $X_t = \log P_t - \log P_{t-1}$, where P_t, $t = 1, 2, \ldots$, is the price of a generic asset. Since

Fabrizio Laurini
University of Parma
Department of Economics and Management and Ro.S.A.
Via J.F. Kennedy, 6, 43125, Parma, Italy
e-mail: fabrizio.laurini@unipr.it

losses can be amplified during periods of large volatility, many risk managers routinely need models to predict the volatility, as isolated extreme values can often be managed, but there is major risk when there is a clustering of these extreme values.

Probably the most popular model, among practitioners, for $\{X_t\}$ is the generalised autoregressive conditionally heteroskedastic (GARCH) defined as $X_t = \sigma_t Z_t$ where $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$. For the process to be strictly stationary, the parameters need to satisfy some constraints.

Properties of GARCH(p,q) processes are often determined by $\phi := \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j$. Two important special cases of GARCH(p,q) processes arise when $p = 0$ or $\phi = 1$, corresponding to ARCH(q) and IGARCH(p,q) processes respectively. The IGARCH(p,q) process is strictly stationary but not second-order stationary, due to $E(X_t^2) = \infty$ for all t. This process is of particular importance as often estimated parameters have $\hat{\phi} \approx 1$.

These models are capable of capturing heavy tails and clustering of extreme values or "threshold-limit" evaluation of $\Pr(X > x)$ and $\chi_X(\tau) = \Pr(X_{t+\tau} > x \mid X_t > x)$ with the latter named the extremogram. In GARCH(p,q) models all depend on the tail behaviour of the squared process X_t^2 which is,

$$\Pr(X_t^2 > ux \mid X_t^2 > u) \to x^{-\kappa}.$$

with $u \to \infty$, fixed $x > 1$ and $\kappa > 0$. Direct computation using the tail expression gives very poor numerical performance, with the exception of simple special cases like ARCH(1) and GARCH(1,1).

We provide empirical evidence that is mainly ϕ influencing κ, with the special case of $\kappa = 1$ being valid for all IGARCH(p,q). We obtain the tail index from the simulation of the so called tail process; this is guaranteed to be efficient for evaluation of limiting properties, as we start the Monte Carlo directly from the limit.

2 Recurrences and regular variation of GARCH(p,q) processes

Focusing on the squared GARCH process, X_t^2, we write the process as a stochastic recurrence equation (SRE) and benefit of some interesting byproducts. Let the $(p+q)$ vector $\mathbf{Y}_t$, the $(p+q)\times(p+q)$ matrix $\mathbf{A}_t$ and the $(p+q)$ vector $\mathbf{B}_t$ be

$$\mathbf{Y}_t = \begin{pmatrix} X_t^2 \\ \vdots \\ X_{t-q+1}^2 \\ \sigma_t^2 \\ \vdots \\ \sigma_{t-p+1}^2 \end{pmatrix}, \quad \mathbf{A}_t = \begin{pmatrix} \alpha^{(q-1)} Z_t^2 & \alpha_q Z_t^2 & \beta^{(p-1)} Z_t^2 & \beta_p Z_t^2 \\ I_{q-1} & 0_{q-1} & 0_{(q-1)\times(p-1)} & 0_{q-1} \\ \alpha^{(q-1)} & \alpha_q & \beta^{(p-1)} & \beta_p \\ 0_{(p-1)\times(q-1)} & 0_{p-1} & I_{p-1} & 0_{p-1} \end{pmatrix}, \quad \mathbf{B}_t = \begin{pmatrix} \alpha_0 Z_t^2 \\ 0_{q-1} \\ \alpha_0 \\ 0_{p-1} \end{pmatrix}$$

then, the squared GARCH(p,q) processes satisfy the SRE

$$\mathbf{Y}_t = \mathbf{A}_t \mathbf{Y}_{t-1} + \mathbf{B}_t, \qquad t \in \mathbb{Z}, \tag{1}$$

where $\{\mathbf{A}_t\}$ and $\{\mathbf{B}_t\}$ are i.i.d. sequences. The stationary solution requires that the top Lyapunov exponent $\gamma < 0$, where $\gamma = \lim_{t\to\infty} \frac{1}{t} E\left(\ln \|\mathbf{A}_t \mathbf{A}_{t-1} \cdots \mathbf{A}_1\|\right)$. This expression is not an ideal starting point for evaluating γ. The numerical evaluation of γ is required whenever $\phi > 1$ and $\sum_{j=1}^{p} \beta_j < 1$.

From Basrak et al (2002) the stationary solution to the SRE (1) exhibits a multivariate regular variation property, i.e., for any t, any vector norm $\|\cdot\|$ and all $r > 0$,

$$\frac{\Pr(\|\mathbf{Y}_t\| > rx, \mathbf{Y}_t/\|\mathbf{Y}_t\| \in \cdot)}{\Pr(\|\mathbf{Y}_t\| > x)} \xrightarrow{v} r^{-\kappa} \Pr(\hat{\mathbf{D}}_t \in \cdot), \qquad \text{as } x \to \infty, \tag{2}$$

where $\xrightarrow{v}$ denotes vague convergence, $\kappa \geq 0$, and $\hat{\mathbf{D}}_t$ is a $(p+q)$-dimensional random vector on the unit sphere (with respect to a norm $\|\cdot\|$) defined by $\mathbb{S}^{p+q} \subset \mathbb{R}^{p+q}$.

The distribution of $\hat{\mathbf{D}}_t$ is termed the spectral measure of the vector $\mathbf{Y}_t$. A consequence of property (2) for GARCH(p,q) processes is that all the marginal variables of $\mathbf{Y}_t$ have regularly varying tails with index $\kappa > 0$, so both X_t^2 and σ_t^2 have regularly varying tails of index κ.

When $\max(p,q) \geq 2$, through use of the multivariate regular variation structure, Basrak and Segers (2009, Propositions 3.3, 5.1) show that

$$E(\|\mathbf{A}\hat{\mathbf{D}}_t\|^{\kappa}) = 1, \tag{3}$$

where $\mathbf{A}$ is i.i.d. to $\mathbf{A}_t$, and uniquely $\Pr(\hat{\mathbf{D}}_t \in \cdot) = E(\|\mathbf{A}\hat{\mathbf{D}}_t\|^{\kappa}; \mathbf{A}\hat{\mathbf{D}}_t/\|\mathbf{A}\hat{\mathbf{D}}_t\| \in \cdot)$, where the notation $E(X;Y) := E(X\mathbf{1}(Y))$ and $\mathbf{1}(Y)$ is the indicator of the event Y.

These results will be key in our subsequent development. To start with,

$$H_{\mathbf{D}_t}(\mathbf{w}) = E(\|\mathbf{A}\hat{\mathbf{D}}_t\|^{\kappa}; (\mathbf{A}\hat{\mathbf{D}}_t/\|\mathbf{A}\hat{\mathbf{D}}_t\|) \leq \mathbf{w}). \tag{4}$$

Additionally, $\kappa > 0$ has a further characterization: it is the unique positive solution of the equation

$$\lim_{t\to\infty} \frac{1}{t} \ln E\left(\|\mathbf{A}_t \mathbf{A}_{t-1} \cdots \mathbf{A}_1\|^{\kappa}\right) = 0. \tag{5}$$

It looks natural to try to evaluate κ by solution of the equation (5), but such an approach is impossible due to numerical instabilities of the limit product of random matrices.

From condition (3) we know that the κ-th moment of $\|\mathbf{A}\hat{\mathbf{D}}_t\|$ is equal to 1, where $\hat{\mathbf{D}}_t \sim H_{\hat{\mathbf{D}}_t}$ on the space $\mathbb{S}^{p+q}$. It is possible to show that there is an equivalent characterization to κ given by

$$\int_{\mathbb{S}^{p+q}} E\left[\|\mathbf{A}\mathbf{w}\|^{\kappa}\right] H_{\kappa}(d\mathbf{w}) = 1, \tag{6}$$

and that the unit measure H_{κ} must be $H_{\hat{\mathbf{D}}_t}$. Thus if we can find, or simulate from, $H_{\hat{\mathbf{D}}_t}$ we can find κ.

To evaluate κ all that is required is to define a class of unit measures H_k, over $k \in (0,\infty)$, which contains within it as an interior point $H_{\hat{\mathbf{D}}_t}$, and then vary k until property (6) is found. This can be done even if Z_t has unbounded support.

3 Evaluating the spectral measure and the tail index

This section gives the details of our algorithm for sampling from the limit distribution $H_{\hat{\mathbf{D}}_t}$ and then uses this algorithm repeatedly to find κ. The algorithm requires no assumptions on the support for Z_t. Throughout we take $t = 0$ as we start the tail process at that time. We will first assume that κ is known and present the idea for generating from $H_{\hat{\mathbf{D}}_0}$ and then discuss the case when κ is unknown.

To simulate from the spectral measure $H_{\hat{\mathbf{D}}_0}$, defined via (4), our approach is to introduce a series of distributions related via a recursion, and whose invariant distribution is $H_{\hat{\mathbf{D}}_0}$, using a particle filtering scheme for fixed point distributions.

Denote the series of random variables whose distribution we are simulating from by $\widetilde{\mathbf{D}}_s$, for iteration $s \geq 0$. The recursion relating the distributions of these random variables for $s \geq 1$ is

$$\Pr(\widetilde{\mathbf{D}}_s \in \cdot) = \frac{E(\|\mathbf{A}\widetilde{\mathbf{D}}_{s-1}\|^\kappa; \mathbf{A}\widetilde{\mathbf{D}}_{s-1}/\|\mathbf{A}\widetilde{\mathbf{D}}_{s-1}\| \in \cdot)}{E(\|\mathbf{A}\widetilde{\mathbf{D}}_{s-1}\|^\kappa)}. \tag{7}$$

By construction, the invariant distribution of this process is $H_{\hat{\mathbf{D}}_0}$, since if $\widetilde{\mathbf{D}}_{s-1}$ is drawn from $H_{\hat{\mathbf{D}}_0}$, then the right-hand side of (7) is equal to

$$\frac{E(\|\mathbf{A}\hat{\mathbf{D}}_0\|^\kappa; \mathbf{A}\hat{\mathbf{D}}_0/\|\mathbf{A}\hat{\mathbf{D}}_0\| \in \cdot)}{E(\|\mathbf{A}\hat{\mathbf{D}}_0\|^\kappa)}.$$

As $E(\|\mathbf{A}\hat{\mathbf{D}}_0\|^\kappa) = 1$, this distribution is equal to the definition of $H_{\hat{\mathbf{D}}_0}$ given by expression (4).

This involves first simulating a value for $\widetilde{\mathbf{D}}_s$ via $\widetilde{\mathbf{D}}_s = \mathbf{A}\widetilde{\mathbf{D}}_{s-1}/\|\mathbf{A}\widetilde{\mathbf{D}}_{s-1}\|$, and assigning this value a weight proportional to $\|\mathbf{A}\widetilde{\mathbf{D}}_{s-1}\|^\kappa$. Thus we can use sequential importance sampling to generate samples of $\widetilde{\mathbf{D}}_s$ for $s \geq 1$ from an initial sample of $\widetilde{\mathbf{D}}_0$.

Sampling from $H_{\mathbf{D}_0}(\mathbf{w})$ is achieved, until convergence, following these s steps:

1. Generate from any distribution in $\mathbb{S}^{p+q}$. Even a multiple uniform could be a valid choice, despite not ideal for speed of convergence.
2. Generate J independent copies of $\mathbf{A}$, denote these as $\mathbf{A}_s^{(j)}$ for $j = 1,\ldots,J$.
3. Generate J equally weighted particles at time $s-1$ by sampling independently from our approximation to the distribution of $\widetilde{\mathbf{D}}_{s-1}$. Denote these particles as $\mathbf{D}_{s-1}^{\star(j)}$ for $j = 1,\ldots,J$.
4. Generate J particles at time s, $\widetilde{\mathbf{D}}_s^{(j)} = \mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}/\|\mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}\|, \quad j = 1,\ldots,J$

5. Assign each particle a weight, $m_s^{\star(j)} = \|\mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}\|^\kappa$ for $j = 1,\ldots,J$, and normalise these via

$$m_s^{(j)} = \frac{m_s^{\star(j)}}{\sum_{j=1}^J m_s^{\star(j)}}. \tag{8}$$

The weighted particles, $\{\widetilde{\mathbf{D}}_s^{(j)}, m_s^{(j)}\}_{j=1}^J$ gives our approximation to the distribution of $\tilde{\mathbf{D}}_s$.

Now consider the situation when κ is unknown. For a trial value of k (for κ), apply the above scheme until convergence and use these particles to approximate the expectation $E(\|\mathbf{A}\widetilde{\mathbf{D}}_0\|^k)$. We repeat this evaluation over $k > 0$ until we find, iteratively, the value of k for which the estimate $E(\|\mathbf{A}\widetilde{\mathbf{D}}_0\|^k) = 1$ is reliable. This value is $k = \kappa$.

Here we attempt to develop insight showing how ϕ influences the value of κ. If we have a strictly stationary GARCH(p,q) process, then $\phi < 1,\ \phi = 1,\ \phi > 1$ if and only if $\kappa > 1,\ \kappa = 1,\ \kappa < 1$ respectively, with $\kappa = 1$ always for any IGARCH(p,q) process. Similarly, knowing $\phi > 1$ or $\phi < 1$ provides valuable information on κ which is quite independent of the heaviness of the tails of the innovations Z.

There are two stages to be used in our approach: initialisation and propagation. For the initialisation stage, we consider the behaviour of the squared GARCH process conditional on it being in an extreme state at time $t = 0$, so we require that $\hat{X}_0^2 > 1$.

The propagation stage uses results of Basrak et al (2002), with the tail chain for $\hat{\mathbf{D}}$, denoted $\{\hat{\mathbf{D}}_t^{TC}\}_{t\geq 0}$, given by $\hat{\mathbf{D}}_t^{TC} = \mathbf{A}_t\hat{\mathbf{D}}_{t-1}^{TC}$ for $t \geq 1$ and $\hat{\mathbf{D}}_0^{TC} = \hat{\mathbf{D}}_0$, with $\hat{\mathbf{D}}_0$ being the vector generated in the initialisation step.

4 Some numerical examples: Evaluation of κ and the extremogram

We take the distribution of the innovation process Z_t as scaled Student-t_ν with zero mean and unit variance. In practice it is impossible to solve equation (5) directly. To calculate κ we iterate over different values of k by taking 10^6 samples from the starting distribution and iterating.

Figure 1 illustrates that ϕ has an impact on the value of κ. When $\phi \neq 1$ no explicit relationship appears to hold between ϕ and κ, as κ changes markedly with the innovation distribution.

We illustrate that the derived value of κ is consistent with the GARCH(p,q) process' observed marginal tail. The observable tail can be derived from long run simulations. Over a range of $r > 1$, we compare the limiting probabilities $\Pr(\hat{X}_t^2 > r \mid \hat{X}_t^2 > 1) = r^{-\kappa}$ with the empirical estimate of the probabilities $\Pr(X_t^2 > rx \mid X_t^2 > x)$ for very large x. Figure 1 shows this comparison for x being the 0.99998 marginal quantile on a log-scale. If κ is correct and x is large enough, the log-probabilities should be proportional with gradient κ. The results show that the limit tail is con-

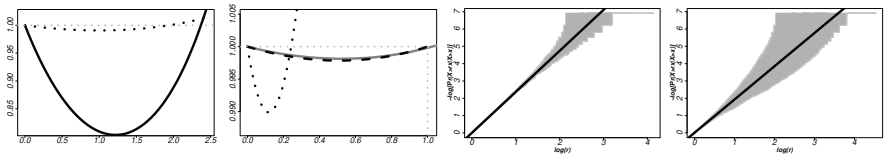


Fig. 1 Plots of κ in some GARCH models: far left, two GARCH(2,2) models (— and $\cdots$) which are second order stationary with $\phi < 1$; middle left for models an IGARCH(1,1) (black dashed) an IGARCH(2,2) (grey solid), that both have $\phi = 1$ and an ARCH(2) taken with $\phi > 1$ (black dotted). Grey dotted lines represent horizontal and vertical lines set at 1. Middle right and far right are the associated QQ plots.

sistent with the empirical distribution subject to Monte Carlo noise, and hence the κ value seems appropriate.

We also contrast the empirical estimate of the extremogram by including the cases of asymmetric Z for an IGARCH(2,2) and an ARCH(2) with $\phi > 1$ in Figure 2 In every comparison black lines are our algorithm's value of the true limit values $\chi_{X^U}(\tau)$ and $\chi_{X^L}(\tau)$ with the three grey lines are empirical extremogram estimates $\tilde{\chi}_{X^U}(\tau,u)$ and $\tilde{\chi}_{X^L}(\tau,u)$ based on a sample of size $n = 5 \times 10^7$, at u corresponding to 0.99 (continuous solid light grey), 0.999 (dashed grey) and 0.9999 (dotted dark grey) quantiles of X_t^U and X_t^L respectively. These comparisons show that as along as a high enough threshold is used then there is very strong agreement between empirical estimates and our evaluation of the limit values for $\chi_X(\tau)$ at different lags for the upper and lower GARCH processes.

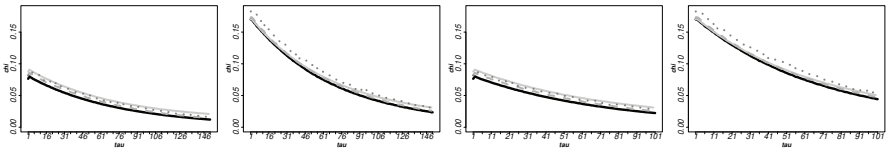


Fig. 2 Extremograms $(\tau, \chi_{X^U}(\tau))$ and $(\tau, \chi_{X^L}(\tau))$ (black solid lines) with $Z_t \sim ST(0,1,1,3)$ and empirical estimates (grey lines) for for an IGARCH(2,2) and an ARCH(2) with $\phi > 1$.

References

Basrak B, Segers J (2009) Regularly varying multivariate time series. Stoch. Proc. App. 119:1055–1080

Basrak B, Davis RA, Mikosch T (2002) Regular variation of GARCH processes. Stoch. Proc. App. 99(1):95–115

Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Mathematica 131:207–248

Bayesian inference for complex random structures

Bayesian nonparametric modeling of mortality curves via functional Dirichlet processes

Modellazione Bayesiana non-parametrica per curve di mortalitá tramite processi di Dirichlet funzionali

Emanuele Aliverti and Bruno Scarpa

Abstract There has been growing interest on modeling mortality curves for multiple nations. We focus on modeling mortality through the age-at-death distribution, a function which characterizes the probability of dying at a specific age given the number of individuals living at that age. Such a measure shows substantially different patterns across nations, motivating the use of flexible models to account for different shapes and levels of smoothing. We rely on a Bayesian nonparametric model leveraging a functional basis decomposition to model country-specific shapes. Model-based clustering of the functional trajectories is induced assigning a Dirichlet-process prior to the basis coefficients, and letting the number of clusters to be inferred from the data. We analyze mortality data from 1991 and 2018, showing interesting differences in terms of functional centroids and clustering.

Abstract *Negli ultimi anni è cresciuto l'interesse vesto la modellazione di curve di mortalità per più nazioni. In questo lavoro, ci concentriamo sulla modellazione della mortalità attraverso la distribuzione dell'età alla morte. Tali funzioni risultano molto eterogenee tra nazioni, motivando l'impiego di modelli flessibili per tenere conto di forme variegate. Per modellare le traiettorie specifiche di ogni paese, si utilizza un modello Bayesiano non parametrico basato su scomposizione in basi funzionali. Il raggruppamento delle traiettorie funzionali viene indotto assegnando un processo di Dirichlet come distribuzione a priori per i coefficienti delle basi, stimando il numero ottimale di gruppi basandosi sui dati. Il modello viene applicato per confrontare le curve di mortalità del 1991 con quelle del 2018, individuando differenze interessanti in termini di raggruppamento e forma dei gruppi funzionali.*

Key words: Bayesian nonparametrics, B-splines, clustering, Dirichlet-process, mortality.

Emanuele Aliverti
Università Ca' Foscari di Venezia, e-mail: emanuele.aliverti@unive.it

Bruno Scarpa
Università degli Studi di Padova, e-mail: scarpa@stat.unipd.it

1 Introduction

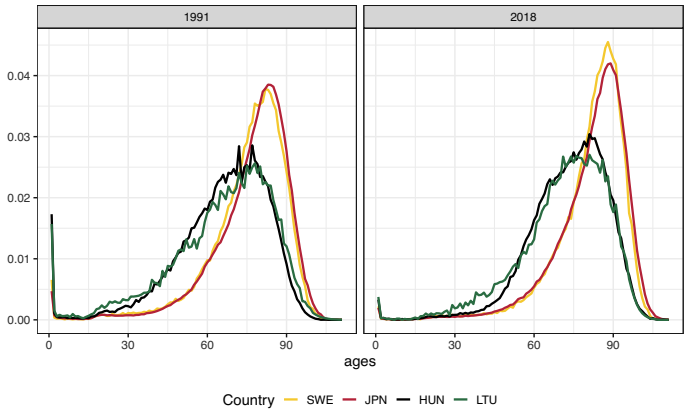


Fig. 1: Age-at-death distribution across four illustrative countries (Sweden, Japan, Hungary and Lithuania) in 1991 and 2018.

Recent changes in life expectancy, population growth and morbidity have stimulated an increased interest in flexible models for mortality [e.g., 1, 11]. For illustration, Figure 1 focuses on four demonstrative countries (Sweden, Japan, Hungary and Lithuania) across two years (1991 and 2018), depicting the year- and country-specific age-at-death distribution. Such quantities provide, for each calendar year, an indication on the country-specific probability of dying at a specific age, based on individuals currently living at that age. Therefore, analysis on these mortality curves can deliver insights on many facets of a population, allowing to compare mortality levels across nations, to investigate recent demographic transitions and to evaluate the overall socio-economic level of a country. In fact, the age-at-death distribution is employed to investigate several demographic trends, such as the compression of old-age-mortality, the evolution of lifespan variability and the reduction of infant and perinatal mortality [e.g., 2].

When interest is on modeling multiple countries, it is important to induce sufficient flexibility to characterize different shapes of the mortality curves. For example, from Figure 1 we observe different patterns across nations, with Central- and Eastern-European countries characterized by larger levels of adult mortality and an anticipation of the modal age at death. In particular, differences in mortality curves involve (a) the overall shape, which is driven by differences in the mortality structure (b) the level of smoothing, with countries with larger population (e.g., Japan) and higher quality data (e.g., Sweden) characterized by smoother curves and (c) the progression across time, which are driven by country-specific transitions.

In this work, we account for these aspects relying on a Bayesian non-parametric model for mortality curves based on a functional Dirichlet process. This model al-

lows to flexibly characterize various shapes of the mortality curves, borrowing information across different countries and smoothing the age-at-death distributions. In addition, the Dirichlet process induces a model-based clustering of similar curves, allowing to group countries according to the shape of mortality and letting the number of clusters to be inferred from the data as a part of the estimation process.

2 Data and methods

We focus on data retrieved from the Human Mortality Database [7], comparing Australia (AUS), Austria (AUT), Belgium (BEL), Belarus (BLR), Canada (CAN), France (FRA), Hong Kong (HKG), Switzerland (CHE), Czechia (CZE), Denmark (DNK), Spain (ESP), Estonia (EST), Finland (FIN), Greece (GRC), Hungary (HUN), Iceland (ISL), Italy (ITA), Japan (JPN), Lithuania (LTU), Luxembourg (LUX), Latvia (LVA), Netherlands (NLD), Norway (NOR), Poland (POL), Portugal (PRT), Slovakia (SVK), Slovenia (SVN), Sweden (SWE), Taiwan (TWN), United Kingdom (U.K.) and United States of America (USA).

Focusing for simplicity on calendar-year data, we denote as $y_i(t)$ the value of the age-at-death distribution for country $i = 1,\dots n$ at age $t = 1,\dots,T$; in our example, $n = 31$ and $T = 111$. Following [6, Chapter 7], we model this quantities as functional data, letting

$$y_i(t) = \eta_i(t) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0,\sigma^2), \tag{1}$$

where $\varepsilon_i(t)$s are independent Gaussian errors and $\eta_i(t)$ is an underlying smooth trajectory. To include smoothness in the model, while avoiding strong assumptions on the functional form of $\eta_i(t)$, we decompose such trajectories via B-splines [4], letting

$$\eta_i(t) = \textstyle\sum_{k=1}^{K} \beta_{ik}\mathbf{B}_k(t), \quad \beta_{i1},\dots,\beta_{1K} \sim Q, \tag{2}$$

where $[\mathbf{B}_1(t),\dots,\mathbf{B}_K(t)]$ denotes a set of fixed cubic B-splines basis functions, while $[\beta_{i1},\dots,\beta_{iK}]$ denotes the country-specific coefficients. Furthermore, we place a Dirichlet process prior on Q, letting

$$Q \sim \text{DP}(\alpha, Q_0), \quad Q_0 \sim N(0,1) \tag{3}$$

where $\text{DP}(\alpha, Q_0)$ denotes a Dirichlet process with concentration parameter α and base measure Q_0, assigned to a standard Gaussian; we refer to [6] for an introduction to the Dirichlet process. The constructive representation of the Dirichlet process lets

$$Q = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \overset{\text{iid}}{\sim} Q_0, \tag{4}$$

were $\pi_h = v_h \prod_{l<h}(1-v_l)$ is a probability weight with $v_h \overset{\text{iid}}{\sim} \text{beta}(1,\alpha)$, and δ_θ denotes a Dirac mass at θ, an atom randomly drawn from Q_0 [9]. This representation as an infinite mixture assigns each observations to a distinct cluster $h = 1,2,\dots,$

with subjects in the same clusters assigned to the same value δ_{θ_h}. In our application, this implies that each country will be allocated to one group, with subjects in the same cluster having the same value for the basis coefficients and therefore same trajectory $\eta_i(t)$.

Prior specification proceeds with an Inverse-Gamma prior on σ^2 with shape and rate parameters equal to $1/100$, while posterior inference relies on truncating the stick-breaking representation at n (the maximum number of groups) and letting $v_n = 0$. This truncation leads to a finite mixture model with stick-breaking weights, allowing conjugate Beta updates for v_h, Gaussian updates for β_i and Inverse-Gamma updates for σ^2. Lastly, we assign a Gamma$(2, 1/4)$ prior on α and update the concentration parameter following [5].

3 Results

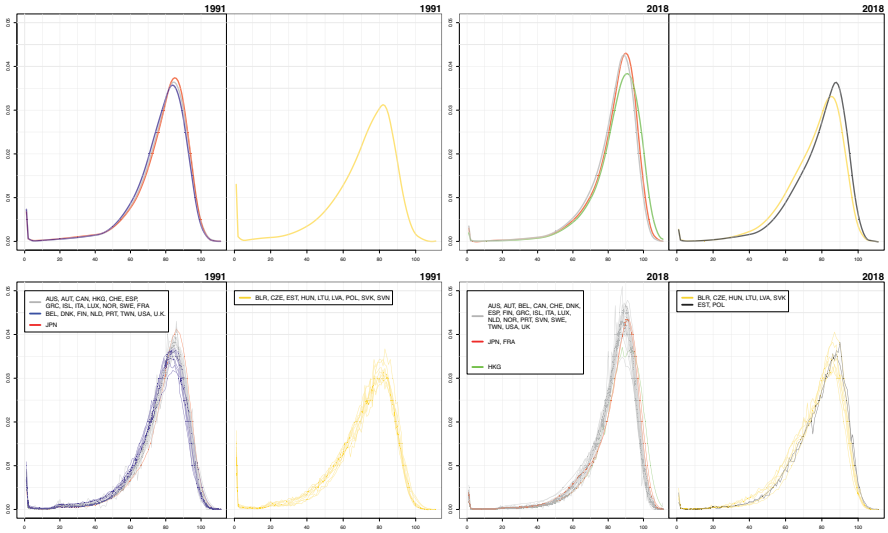


Fig. 2: Top panels: estimated functional centroids for the non-empty groups. Bottom panels: observed data and cluster allocation. Data from the same calendar year have been divided into two panels – according to the cluster allocation – to improve the graphical visualization.

Posterior inference relies on 5000 iterations collected after a burn-in of 2000; effective sample size and mixing were satisfactory for all the parameters of interest. We conduct inference on the cluster assignment of the curves and the functional centroids of each non-empty group, estimated via posterior mean; graphical representations of these quantities are reported in the bottom and top panel of Figure 2, respectively. In 1991 (first and second panels from the left), data provide evidence

of 4 clusters. The first panel depicts results from the 3 groups that characterize western countries; we observe two main clusters (grey and blue curves in the first panel) with similar shapes, with the former characterized by larger modal age at death. Japan (red curve) occupies a singleton; this results is not surprising, considering that it is the country with the most compressed old-age mortality and the most delayed modal age at death in 1991.

Results from 2018 show some remarkable differences. In Central- and Eastern-European countries, posterior inference provides evidence of two clusters (fourth panel). Indeed, Estonia and Poland are assigned to a separate trajectory (black curves), which is characterized by lower adult mortality than the cluster with Belarus and Hungary (yellow curves), among others. This result confirms how these countries have experienced substantially different demographic transitions in the last 30 years [8]. Additionally, we observe that Slovenia has been assigned to the large block of western countries (third panel, grey curves), showing a mortality which is more similar to western-European countries. These nations are characterized by a compression of old-age mortality, low levels of infant mortality and a modal age at death close to 88 years. Most countries are assigned to the same cluster, as a result of global convergence trends in terms of mortality [e.g., 3, 10]. However, posterior inference identifies two more shifted clusters: one contains Japan and France (red curves), while the other is a singleton containing Hong-Kong (green curve), well known for being the country with largest life expectancy.

4 Discussion

In this work, we used a Bayesian nonparametric functional model to characterize mortality curves in different countries, comparing the age-at-death distribution across 1991 and 2018 in terms of functional clustering.

Some future directions are worth to be explored. For example, it would be interesting to consider male and female population separately, as mortality is notoriously characterized by sex differences [e.g., 1]. From a modeling perspective, the formal inclusion of the temporal dependence across years is desirable, in order to model the time series of functional curves and characterize the country-specific evolution across time; also, we expect that such transitions are similar for neighbors countries, and therefore the inclusion of proximity information could improve inference.

5 Acknowledgments

A preliminary version of these analysis been conducted during the master thesis of Sofia Curzio, MSc in Statistical Sciences at the University of Padova.

References

[1] Emanuele Aliverti, Stefano Mazzuco, and Bruno Scarpa. Dynamic modelling of mortality via mixtures of skewed distribution functions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, (in press), 2021.

[2] Ugofilippo Basellini and Carlo Giovanni Camarda. Modelling and forecasting adult age-at-death distributions. *Population Studies*, 73(1):119–138, 2019.

[3] Marie-Pier Bergeron-Boucher, Vladimir Canudas-Romo, Jim Oeppen, and James W. Vaupel. Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, 37:527–566, 2017.

[4] Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

[5] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. 90(430):577–588, 1995.

[6] Nils L. Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. *Bayesian nonparametrics*. Cambridge University Press, 2010.

[7] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).

[8] France Meslé. Mortality in central and eastern europe: long-term trends and recent upturns. *Demographic Research*, 2:45–70, 2004.

[9] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[10] Chris Wilson. On the scale of global demographic convergence 1950–2000. *Population and Development Review*, 27(1):155–171, 2001.

[11] Lucia Zanotto, Vladimir Canudas-Romo, and Stefano Mazzuco. A mixture-function mortality model: illustration of the evolution of premature mortality. *European Journal of Population*, pages 1–27, 2020.

Bayesian nonparametric clustering of spatially-referenced spike train data

Raggruppamento bayesiano nonparametrico di serie di attivazioni neuronali georeferenziate

Laura D'Angelo

Abstract Spike trains are a representation of the activity of neurons: they indicate the sequence of recorded firing events, and they are usually expressed as binary time series which, at each time, indicate the active/resting state of neurons. Clustering of these series is a relevant task in neuroscience, as it allows for identification of groups of co-activating cells. We propose a Bayesian nonparametric mixture model that clusters neurons with a similar activation pattern. The model relies on a latent continuous process that describes the evolution of the spike probabilities to identify similar time series. Moreover, the spatial location of each neuron is used to inform the mixture weights: this favors clustering of neighboring cells, following the neuroscience assumption that close neurons often activate together.

Abstract *Gli spike train sono rappresentazioni dell'attività neuronale: essi descrivono le attivazioni dei singoli neuroni nel tempo, e sono generalmente espressi come serie binarie che a ogni istante temporale indicano lo stato di attivazione o riposo. A partire da queste serie è di interesse identificare cellule con modelli di attivazione simili. In questo contributo si propone un modello mistura bayesiano nonparametrico che permette di raggruppare neuroni a partire dagli spike train. Il modello fa uso di un processo latente continuo che descrive l'evoluzione delle probabilità di attivazione e permette di identificare serie simili. Inoltre, i pesi della mistura sono funzione della posizione dei neuroni, in modo da favorire il raggruppamento di neuroni spazialmente vicini, come suggerito da studi nelle neuroscienze.*

Key words: Calcium imaging, Dirichlet process, Gaussian process, Mixture models, Probit stick-breaking.

Laura D'Angelo
Department of Economics, Management and Statistics
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan
e-mail: laura.dangelo@unimib.it

1 Introduction

Calcium imaging is a microscopy technique that allows for visualization of the activity of populations of neurons over time at a neuronal level. The output of this technique is a movie of time-varying fluorescence intensities, and a complex preprocessing phase aims to extract the calcium traces for each observable cell in the region of interest. Then, a deconvolution phase is employed to extract the *spike trains*, which are binary time series that describe the presence or absence of a spike at each time point. The spikes correspond to activations of the cell, usually in response to some particular stimulation. In some areas of the brain as, for example, the hippocampus, it is of interest to investigate the existence of groups of co-activating cells, as the functional properties of this area are topic of research and discussion (Eichenbaum et al., 1989; Redish et al., 2001). However, identification of such groups is a difficult task, as it requires clustering binary time series that exhibit a similar pattern over long periods of time (Bittner et al., 2017). In many areas of the brain the functional properties of neurons are linked with their anatomical structure, hence the spatial location of the cells can be a relevant piece of information in the clustering procedure in order to favor grouping of neighboring neurons.

Our interest is in clustering similar binary time series: a difficulty in this context is the presence of erratic spikes, which make the series somehow different, even if the overall pattern matches. We propose to model each series as independent realizations of Bernoulli random variables, whose probabilities however depend on a latent continuous mixture, which describes the temporal evolution of the spike probability. This process also allows us to easily introduce a temporal dependence structure between spikes, as they are usually not uniformly distributed in time: it is a known phenomenon, often observed in calcium imaging studies, the occurrence of multiple consecutive spikes, which lead to longer observed calcium transients (Dombeck et al., 2010; D'Angelo et al., 2022).

As motivating application we considered a dataset containing the calcium traces of neurons located in the hippocampus of a mouse. First, we performed a preprocessing phase to extract the spike trains using the deconvolution method of Jewell et al. (2019). In Section 2 we describe the proposed mixture model to perform clustering of these binary time series; in Section 3 we apply the proposed approach to the hippocampal data.

2 Model specification

Let $\boldsymbol{s}_i = \{s_{i,1}, \ldots, s_{i,T}\}$ be the spike train of neuron $i = 1, \ldots, n$, meaning that for all time points $t = 1, \ldots, T$, $s_{i,t} \in \{0, 1\}$ describes the absence/presence of a spike. Moreover, assume that for each neuron we also have information on the position in the brain, expressed through the spatial coordinates $\boldsymbol{l}_i \in \mathscr{L} \subset \mathbb{R}^2$.

Similarly to D'Angelo (2022), we assume that, for each t, the spike train $s_{i,t}$ is the realization of independent Bernoulli random variables whose probabilities depend

on an underlying mixture of Gaussian processes through a probit transformation. Denoting with $\tilde{\boldsymbol{s}}_i = \{\tilde{s}_{i,1}, \ldots, \tilde{s}_{i,T}\}$ the realization of this latent process, we write

$$s_{i,t} \sim \text{Bernoulli}(\Phi(\tilde{s}_{i,t}))$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution. On the continuous process $\tilde{\boldsymbol{s}}_i$ we specify a nonparametric mixture prior to induce a clustering of neurons. The atoms of the mixture are realizations of a Gaussian process over time: denoting with $\{\tilde{\boldsymbol{s}}_1^*, \ldots, \tilde{\boldsymbol{s}}_K^*\}$ the set of distinct values, because of the almost sure discreetness of draws from this prior, there is a positive probability of observing ties, hence giving rise to a clustering structure. When multiple neurons are associated with the same value $\tilde{\boldsymbol{s}}_k^*$, for $k = 1, \ldots, K$, they will share the same spike probabilities, thus inducing similarities in the observed spike trains.

Moreover, to introduce information on the spatial location of neurons, we make use of the probit stick-breaking process of Rodríguez et al. (2008). Denoting with $\Sigma = \Sigma(\boldsymbol{l}_i)$ the proximity matrix between neurons, taking value 1 on the diagonal (corresponding to a distance equal to zero), and with off-diagonal values that exponentially decrease with the distance between cells, the weights of the mixture are built through a stick-breaking construction starting from random variables that depend on Σ. This favors clustering of neurons located close to each other.

A similar model was used in D'Angelo (2022) to simultaneously deconvolve and cluster fluorescence traces. Here, however, we focus on the clustering task, and we consider directly the extracted spike trains. This two-stage procedure is computationally more efficient, as the deconvolution phase can be performed through efficient optimization strategies, as, for example, those described in Pnevmatikakis et al. (2016), Friedrich and Paninski (2016), Friedrich et al. (2017), Jewell and Witten (2018), and Jewell et al. (2019).

3 Analysis of hippocampal spike train data

We considered a subset of 20 neurons sampled from a large dataset containing the calcium traces of hundreds of neurons located in the hippocampus of a mouse. We considered the first 1000 time points of the series, and we applied the model of Section 2 to obtain a clustering of these spike trains.

Figure 1 shows the clustering resulting from application of our model. Each row of the plot represents the extracted spike trains: in correspondence of each time point, a vertical segment indicates the presence of a spike. For clarity, we also reported the observed (raw) fluorescent traces, represented with continuous lines. Each line is colored according to its estimated cluster membership. The represented partition is the posterior point estimate obtained by minimizing the variation of information loss (Wade and Ghahramani, 2018). We estimated the presence of 4 groups of co-activating neurons. For some traces, it is evident the presence of co-

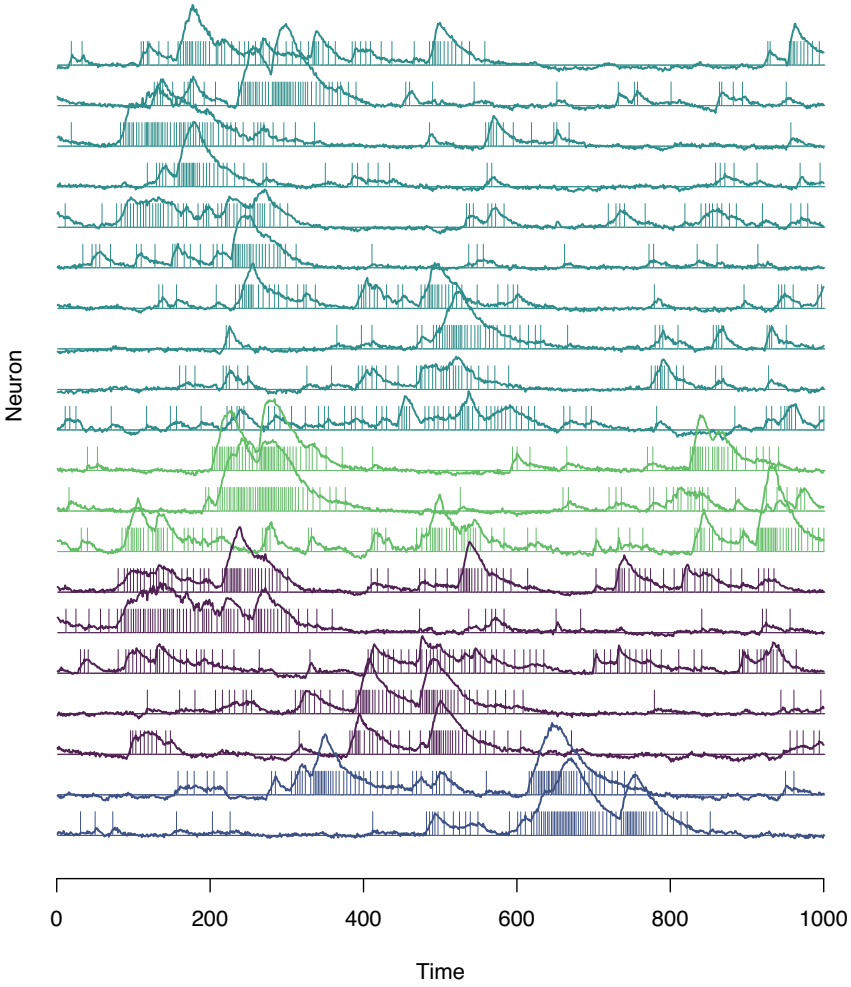


Fig. 1 Observed fluorescence traces (continuous lines) and extracted spike trains (0/1 vertical segments). The colors correspond to the estimated cluster.

occuring spikes as, for example, for the two bottom neurons, which have an intense activity between time 600 and 700.

Our model makes use of the spatial location of each neuron to inform the mixture weights, to encourage neighboring neurons to be allocated to the same cluster. It is then interesting to analyze how the estimated groups of co-activating neurons are spread in the region of interest. Figure 2 shows the spatial location of the considered

cells, and the colors again correspond to the estimated cluster. It is possible to notice that many neurons allocated to the same group are quite close one another; however, these groups of co-activating neurons can also be quite scattered. This is consistent with the results of other studies, which found that co-activating neurons are often close to each other, but they can also be found in regions quite far from the "center" of the cluster.

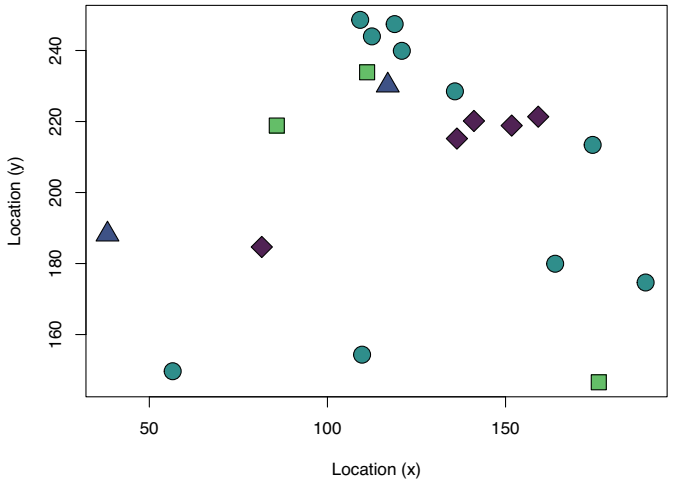


Fig. 2 Location of the neurons in the hippocampus. The colors correspond to the estimated cluster.

References

1. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S., and Magee, J. C.: Behavioral time scale synaptic plasticity underlies CA1 place fields. Science **357**, 1033–1036 (2017)
2. D'Angelo, L.: Bayesian modeling of calcium imaging data. PhD Thesis, University of Padova (2022)
3. D'Angelo, L., Canale, A., Yu, Z., and Guindani, M.: Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. Biometrics (2022) DOI: 10.1111/biom.13626.
4. Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L., and Tank, D. W.: Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. Nat. Neurosci. **13**, 1433–1440 (2010)
5. Eichenbaum, H., Wiener, S. I., Shapiro, M., and Cohen, N. J.: The organization of spatial coding in the hippocampus: a study of neural ensemble activity. J. Neurosci. (1989)
6. Friedrich, J. and Paninski, L.: Fast active set methods for online spike inference from calcium imaging. NIPS. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Barcelona, Spain, 1984–1992 (2016)

7. Friedrich, J., Zhou, P., and Paninski, L.: Fast online deconvolution of calcium imaging data. PLoS Comput. Biol. **13**(3), 1–26 (2017)
8. Jewell, S. and Witten, D.: Exact spike train inference via L0 optimization. Ann Appl Stat. **12**(4), 2457–2482 (2018)
9. Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M.: Fast nonconvex deconvolution of calcium imaging data. Biostatistics **21**(4), 709–726 (2019)
10. Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., et al.: Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron **89**(2), 285–299 (2016)
11. Redish, A. D., Battaglia, F. P., Chawla, M. K., Ekstrom, A. D., Gerrard, J. L., et al.: Independence of firing correlates of anatomically proximate hippocampal pyramidal cells. J. Neurosci. **21**(5) (2001)
12. Rodríguez, A., Dunson, D. B., and Gelfand, A. E.: The nested Dirichlet process. J. Am. Stat. Assoc. **103**(483), 1131–1154 (2008)
13. Wade, S. and Ghahramani, Z. . Bayesian cluster analysis: point estimation and credible balls (with discussion). Bayesian Anal. **13**(2), 559–626 (2018)

Bayesian Analysis of Mortality in Iceland via Locally Adaptive Splines

Analisi Bayesiana della Mortalità in Islanda tramite Spline Localmente Adattive

Federico Pavone and Sirio Legramanti

Abstract Despite its long history, mortality data analysis is still a very active field of research and has witnessed interesting recent developments. In this work, we focus on Iceland, which has a remarkably long record of mortality data. We analyze age-specific mortality data from 1838 to 2018, adopting the approach recently proposed in Pavone et al. (2022). This model consists of locally-adaptive splines, which guarantee interpretable inference alongside flexible time dynamics. The local adaptivity allows modelling gradual changes in mortality rates together with possible sudden shocks, arguably due to dramatic events such as wars and epidemics.

Abstract *Nonostante la sua lunga storia, l'analisi dei dati di mortalità è ancora un ambito di ricerca molto attivo e ha recentemente mostrato diversi sviluppi interessanti. In questo lavoro, ci focalizziamo sull'Islanda, che ha una storia di dati di mortalità notevolmente lunga. Analizziamo dati di mortalità suddivisi per età, dal 1838 al 2018, usando l'approccio proposto recentemente in Pavone et al. (2022). Tale modello consiste in processi spline localmente adattivi, che garantiscono un'inferenza interpretabile oltre ad una dinamica temporale flessibile. L'adattività locale permette di modellare sia cambiamenti graduali nei tassi di mortalità che possibili shock improvvisi, presumibilmente dovuti ad eventi drammatici quali guerre ed epidemie.*

Key words: Demography, Gaussian process, Time series

Federico Pavone
Università Bocconi, Via Röntgen 1, 20136 Milano (Italia),
e-mail: federico.pavone@phd.unibocconi.it

Sirio Legramanti
Università degli Studi di Bergamo, Via dei Caniana 2, 24127 Bergamo (Italia),
e-mail: sirio.legramanti@unibg.it

1 Introduction

The study of mortality has been of interest for a very long time in different research fields including actuary, demography, and statistics. The most developed countries have been collecting demographic data since at least the 17th century. The analysis of mortality data is aimed at either projecting mortality trends into the future, or at studying the impact of both endogenous and exogenous events. Projecting mortality trends in the near future is paramount to design policies for social security, health, and economics and to plan business strategies for some private companies. For example, insurance companies base part of their business strategies on analyses of mortality data aimed at estimating life expectancies of population subgroups.

The study of historical data allows understanding how epidemics and wars have affected mortality. Moreover, when further information - e.g. causes of death - are available, mortality data are used to guide policies related to, for example, drug consumption or endemic diseases.

In this work, we apply the model proposed by [6] to Icelandic mortality records. Iceland has an extraordinary long record of mortality data, making it of particular interest from a demographic standpoint. Moreover, this country has undergone several dramatic events, ranging from measles epidemics to extreme weather conditions, which arguably impacted age-specific mortality rates.

The rest of the paper is structured as follows: in Section 2 we describe the adopted model, while in Section 3 we analyze Icelandic mortality rates from 1838 to 2018.

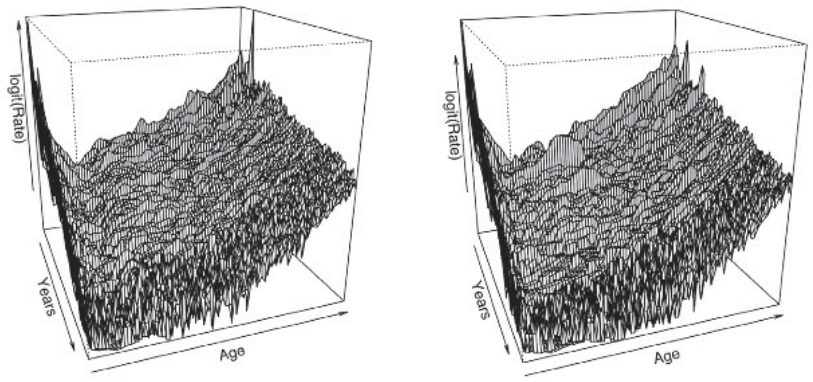


Fig. 1: Mortality rates in Iceland from 1838 to 2018, for males (left) and females (right); ages range from 0 to 80.

2 Model specification

The motivating Icelandic mortality rates, grouped by sex, are plotted in Figure 1 as a function of age and year. The plotted surfaces exhibit uneven levels of smoothness, with relatively smooth areas and sharper transitions. This motivates the adoption of the locally-adaptive approach proposed in [6].

Such a proposal defines locally-adaptive spline processes, which offer interpretable inference and flexible time dynamics. For each year, mortality rates as a function of age are modelled via spline processes. The evolution of spline coefficients across years instead follows a multivariate nested Gaussian process. The nested Gaussian process [10], and its multivariate extension [6], have a locally-adaptive smoothness property which makes them suitable to model sudden shocks, possibly related to external events such as epidemics and wars.

Let us denote with m_{tz} the observed mortality rate at age z during year t and with $\boldsymbol{m}_t = (m_{t1}, \ldots, m_{tZ})^T$ the vector collecting the rates over all ages. We rely on the large n approximation suggested by [6], which allows the model to be written as

$$\begin{aligned} \text{logit}(\boldsymbol{m}_t)|\boldsymbol{u}_t &\sim \text{N}_Z(S\boldsymbol{u}_t, \sigma_\varepsilon^2 I_Z) \\ \{\boldsymbol{u}_t\}_{t\in\mathscr{T}} &\sim \text{mnGP}(\sigma_u^2, \sigma_a^2, \rho, \phi, \kappa), \end{aligned} \tag{1}$$

where N_Z is the Z-dimensional multivariate normal distribution, S is the design matrix derived from the pre-specified M-spline basis functions, I_Z is the Z-dimensional identity matrix, and mnGP is the multivariate nested Gaussian process of parameters σ_u^2, σ_a^2, ρ, ϕ, κ. The latter three hyperparameters define the correlation structure among the components of $\boldsymbol{u}_t = (u_{t0}, \ldots, u_{tK})^T$. The first component u_{t0} models mortality within the first year of life, while the remaining latent variables $\{u_{t1}, \ldots, u_{tK}\}$ are the weights of the K basis functions. Due to the limited support of each basis function, the latent variables can be interpreted as the mortality rates of the age classes induced by the spline representation.

When restricted to the observed years, the model takes the convenient form of a Gaussian state space model, in which the latent states consist of a 3-dimensional vector $\boldsymbol{\theta}_{tk}$ for each term u_{tk}. The three components of $\boldsymbol{\theta}_{tk}$ are u_{tk} and its first and second derivatives, denoted with du_{tk} and a_{tk}. Looking at the smoothing distribution of du_{tk}, we can get information about how fast the mortality rates change at specific time instants and, for example, compare different shocks in terms of suddenness.

The model hyperparameters are estimated via maximum marginal likelihood, while the smoothing and filtering distributions of the latent states are obtained via Kalman filtering and smoothing [3, 4].

3 Analysis of Iceland age-specific mortality from 1838 to 2018

We consider Icelandic mortality data from 1838 to 2018. These data, originally from The National Statistical Institute of Iceland[1], are publicly available in the Human Mortality Database[2]. We restrict our analysis to ages ranging from 0 to $Z = 80$. We use M-splines of second degree and knots at ages 5, 15, 25, 35, 50, 60, 70, and 75, resulting in $K = 11$ spline functions. We estimate the model parameters using the Nelder-Mead optimization algorithm [5], while Kalman filtering is performed via the `R`-package `KFAS` [7, 9].

Figure 2 shows the smoothing distribution of some of the latent variables corresponding to five different age classes. We can observe similar mortality patterns for males and females with a higher signal-to-noise ratio for female data, reflected in a more oscillatory behaviour compared to males.

The dashed vertical lines in Figure 2 mark some specific years of Icelandic demographic history [1, 8]. In 1846 a measles epidemic hit the country, contributing to the bumps that we can observe in the mortality curves of almost all age classes. In 1869, Iceland suffered epidemics of diphtheria and croup, which however had a smaller impact on mortality rates. In 1882 there was a second outbreak of measles, and for age classes 10-26 and 22-42 the bumps around this year are higher than the ones of the 1846 epidemic. Part of this difference can be probably attributed to the extremely cold weather that characterized 1882 and 1883, due to a large amount of ice off the coast of Iceland. This caused extremely low temperatures, leading to severe problems with crops and famine, together with cholera epidemics.

The twentieth century witnessed an overall decay in mortality [1]. In particular, the 1-4 age class shows a constant decrease in the mortality rate since the beginning of the century. The 1918 Spanish flu has been the last big epidemic in the country, and the 22-46 age class presents the highest shock in that year. This is coherent with the fact that, also in other countries, the fatality of Spanish flu has been higher among young adults [2]. A few decades later, between 1940 and 1950, we notice a period of lower negative values of du_k for age classes 1-4, 10-26, and 22-42. This indicates a faster decrease of mortality rates in those years.

Acknowledgements The authors are grateful to Daniele Durante for his helpful comments on a first draft of this paper. The authors also acknowledge the support from MIUR-PRIN 2017 project 20177BRJXS.

[1] Statistics Iceland. URL www.statice.is

[2] Human Mortality Database. URL www.mortality.org

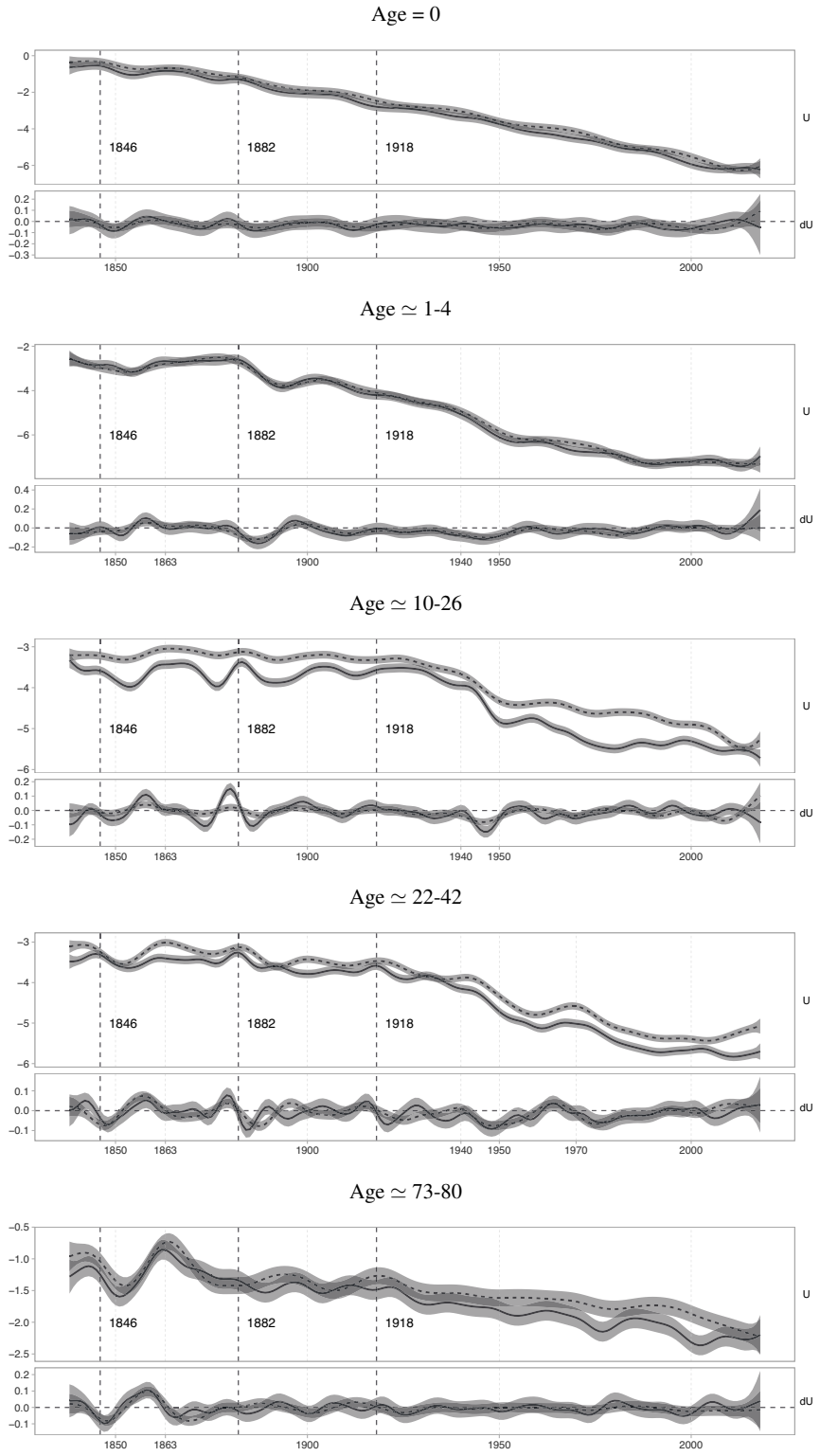


Fig. 2: Posterior smoothing distribution of u_0, u_1, u_4, u_5, and u_{10}. Female and male mortality respectively in solid and dashed line. Two standard deviation uncertainty is reported. The age range refers approximately to the splines support.

References

1. Andreeva M. About mortality data for Iceland. Document from the Human Mortality Database. URL www.mortality.org/hmd/ISL/InputDB/ISLcom.pdf
2. Gagnon A, Miller MS, Hallman SA, Bourbeau R, Herring DA, Earn DJ, Madrenas J. Age-specific mortality during the 1918 influenza pandemic: Unravelling the mystery of high young adult mortality. PloS one. 2013; 8(8):e69586.
3. Kalman RE. A new approach to linear filtering and prediction problems. Journal of Basic Engineering. 1960; 82(1):35-45.
4. Kalman RE, Bucy RS. New results in linear filtering and prediction theory. Journal of Basic Engineering. 1961; 83(1):95-108.
5. Nelder JA, Mead R. A simplex method for function minimization. The Computer Journal. 1965; 7(4):308-13.
6. Pavone F, Legramanti S, Durante D. Bayesian learning and forecasting of age-specific period mortality via locally adaptive spline processes. Working paper. 2022
7. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL www.r-project.org. 2020
8. Tomasson RF. A millennium of misery: The demography of the Icelanders. Population Studies. 1977; 31(3):405-27.
9. Villegas AM, Kaishev VK, Millossovich P. StMoMo: An R Package for Stochastic Mortality Modeling. Journal of Statistical Software. 2018; 84(3):1-38.
10. Zhu B, Dunson DB. Locally adaptive Bayes nonparametric regression via nested Gaussian processes. Journal of the American Statistical Association. 2013; 108(504):1445-56.

Advances in clustering

Two-step Latent Class Approach with Measurement Equivalence Testing

Approccio a due fasi per modelli a classi latenti, con test per equivalenza di misurazione

Zsuzsa Bakk, Roberto Di Mari, Jennifer Oser, Marc Hooghe[1]

Abstract In this study we introduce a two-step approach to latent class analysis using measurement equivalence testing, and apply the approach to cross-national data on adolescents in 14 countries in 1999, 2009 and 2016 to investigate competing expectations about changing citizenship norms. The findings exemplify how stepwise latent class modelling can be implemented to separate the measurement model (types of citizenship norms) from the structural model (change of norms over time, and the effect of covariates on the development of norms), while testing for measurement equivalence.

Abstract *Nel presente lavoro viene introdotto un approccio a due fasi per l'analisi delle classi latenti che include test per la cosiddetta measurement equivalence. La proposta viene applicata ad un data set, di tipo cross-section, relativo ad adolescenti di 14 diversi paesi, negli anni 1999, 2009 e 2016, al fine di fornire un quadro delle aspettative circa le norme di cittadinanza. I risultati illustrano come l'approccio a fasi per l'analisi delle classi latenti può essere efficacemente implementato per la separazione del modello di misurazione (tipologia normativa di cittadinanza) dal modello strutturale (dinamiche temporali del concetto di norma e effetto di variabili esplicative sullo sviluppo della stessa), consentendo di testare per la measurement equivalence.*

[1] Zsuzsa Bakk, Leiden University, z.bakk@fsw.leidenuniv.nl,
Roberto Di Mari, Catania University, roberto.dimari@unict.it
Jennifer Oser, Ben-Gurion University, oser@post.bgu.ac.il
Marc Hooghe University of Leuven, marc.hooghe@kuleuven.be

Key words: latent class analysis, two-step estimators, measurement equivalence, citizenship norms

1 Introduction

Political science literature about changing citizenship norms is divided between a "citizen engagement and a "democratic erosion" school of thought. The "citizen engagement" theory suggests an increasingly prevalent preference for an engaged citizenship norm that is common especially among young cohorts who are politically active in diverse ways [2]. In contrast, a more recent line of research has observed signs of cultural backlash and democratic erosion [3]. In order to examine the prevalence of the different norm types we propose the use of latent class (LC) analysis, a person-centered approach that allows for the classification of respondents into a set of (latent) groups with testing for the simultaneous presence of the different groups across countries. We propose a two-step estimator [1] that is conceptually and practically preferable over the classical one and three-step estimators when analysing complex datasets. The approach works as follows: in the first step, we fit a simple LC model based on the citizenship norms indicators. In an intermediary step, measurement equivalence of the first step model is established. In the second step, we fit a full LC model with the covariates of interest related to the LC variable and the measurement model parameters fixed at their first-step values.

2 Latent Class Modeling with Covariates: the Two-step Estimator

The LC model can be formalized as follows. Let $i = 1, \dots, N$ be the index of respondents, $j = 1, \dots, J$ be the index of items and $s = 1, \dots, S$ be the index of the latent classes. The value $X_i = s$ denotes that respondent i is a member of latent class s, and Y_{ij} indicates whether respondent i answered "yes" ($Y_{ij} = 1$) or "no" ($Y_{ij} = 0$) to the j-th item and $\boldsymbol{Y}_i$ denotes the full vector of responses.

The latent class measurement model can be formulated as follows:

$$P(\boldsymbol{Y}_i \,|X_i) = \sum_{s=1}^{S} P(X_i = s\,) \prod_{j=1}^{J} P(Y_{ij}|X_i = s) = \sum_{s=1}^{S} P(X_i = s\,) P(\boldsymbol{Y}_i|\, X_i = s) \quad (1)$$

Note that in Equation (1) by taking the product over the set of indicators J we assume that the indicators are independent of each other given the latent classes.
The conditional item probabilities can be expressed and estimated using logistic regression parametrization :

$$P\left(Y_j \,\middle|\, \boldsymbol{X} = \boldsymbol{s}\right) = \frac{\exp(\alpha_{js})}{1+\exp(\alpha_{js})} \quad (2)$$

Let us assume a K-vector of individual covariates is available, and let us denote with $\boldsymbol{Z}_i$ the observed values for respondent i on the K covariates. The structural model refers to the probability that respondent i belongs to class s, for $s = 1, \dots, S$, allowing these probabilities to depend on covariate values, that is $P(X_i = s \mid \boldsymbol{Z}\_i)$. A latent class model with covariates specifies the probability of having a specific set of responses for respondent i as follows:

$$P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) = \sum_{s=1}^{S} P(X_i = s \mid \boldsymbol{Z}_i) P(\boldsymbol{Y}_i \mid X_i = s) \quad (3)$$

Logistic regressions can be used to parametrize our latent class probabilities as follows:

$$P(X_i = s \mid \boldsymbol{Z}\_i) = \frac{\exp(\alpha_s + \boldsymbol{Z}_i' \boldsymbol{\beta}_s)}{1 + \sum_{t=2}^{S} \exp(\alpha_t + \boldsymbol{Z}_i' \boldsymbol{\beta}_t)} \quad (4)$$

where α_s and $\boldsymbol{\beta}_s$ are respectively the intercept term and a K-vector of regression coefficients, for $s = 2, \dots, S$.

Furthermore, the model formulated in Equation 3 assumes that the indicators $\boldsymbol{Y}_i$ are conditionally independent of the covariates $\boldsymbol{Z}_i$ given the LC variable X_i. This assumption can be relaxed by allowing direct effects of **Z** on **Y**, a situation known as measurement inequivalence. That is Equation 3 can be modified as:

$$P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) = \sum_{s=1}^{S} P(X_i = s \mid \boldsymbol{Z}_i) P(\boldsymbol{Y}_i \mid X_i = s, \boldsymbol{Z}_i) \quad (5)$$

Assuming a sample of N respondents, under the above specification the model log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i),$$

where $\boldsymbol{\theta}$ denotes the set of all model parameters. Using full information maximum likelihood, the model defined in Equation (3,5) can be estimated simultaneously. While statistically this is the most efficient estimator, in practice step-wise estimators are preferred, because of the possibility to separate the establishment of a measurement model from the modelling of antecedents and consequences of the clustering. This latter – referred to as structural model - is seen as a separate step, and the same measurement model (often after being validated) can be used in different structural models.

Let us consider partitioning $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$contains the parameters of the measurement model – i.e. all $S \times J$ item-response probabilities – and $\boldsymbol{\theta}_2$ contains the regression parameters α_s and $\boldsymbol{\beta}_s$, for $s = 2, \dots, S$. The two-step estimator first estimates $\boldsymbol{\theta}_1$fitting a simple latent class model without covariates. Given the ML estimate $\widehat{\boldsymbol{\theta}}_1$of $\boldsymbol{\theta}_1$, we estimate $\boldsymbol{\theta}_2$ by maximizing the following (pseudo) log-likelihood:

$$\boldsymbol{\ell}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_2) = \sum_{i=1}^{N} \log P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) .$$

The measurement model parameters are therefore kept fixed at their first-step values, and the log-likelihood is maximized only with respect to the structural model parameters. Formulas to compute correct standard errors for $\widehat{\boldsymbol{\theta}}_2$ that account also for

the variability of the uncertainty deriving from step 2 are available in closed form (see [1] for derivations).

2.1 Testing for Measurement Equivalence using the Two-step Estimator

Testing for measurement equivalence at both scale and item level is done over the first step model by reparametrizing the item-response probabilities to obtain the homogeneous, partially homogeneous and heterogeneous model – using [4]'s terminology.

The most general heterogeneous model implies that the measurement model is different in each group (e.g. country). This is obtained by modifying Equation 2 to allow for group specific latent class effect. By letting $\boldsymbol{Z_i}$ be the vector of $G-1$ country dummies (where the G-th has been taken as reference), we therefore have**:**

$$P(Y_j \mid \boldsymbol{X}=\boldsymbol{s},\boldsymbol{Z}) = \frac{\exp(\alpha_{js} + \boldsymbol{Z_i'}\boldsymbol{\beta}_{js})}{1+\exp(\alpha_{js} + \boldsymbol{Z_i'}\boldsymbol{\beta}_{js})} \quad (6)$$

A more restrictive model - the partially homogeneous model - is obtained by restricting the grouping variable to have a class-constant effect on the indicators:

$$P(Y_j \mid \boldsymbol{X}=\boldsymbol{s},\boldsymbol{Z}) = \frac{\exp(\alpha_{js} + \boldsymbol{Z_i'}\boldsymbol{\beta}_{j})}{1+\exp(\alpha_{js} + \boldsymbol{Z_i'}\boldsymbol{\beta}_{j})} \quad (7)$$

The most restrictive model of interest - the structural homogeneous model - is obtained by allowing only an indirect effect of the grouping variable via the latent classes.

In our analytical approach, we implement the following steps:

1. Establish the measurement model on the pooled data over the 3 timepoints by fitting a set of LC models with increasing number of classes. Select the optimal number of latent classes using fit measures such as BIC (Schwarz 1978).
2. On the final LC model, establish measurement equivalence across the participating countries using the steps proposed by [4] by selecting the best fitting model using BIC.
3. Fit the structural model conditioning on the final model from step 2.

3 Data analysis

We investigate changing citizenship norms by analysing comparative data on adolescents' conceptions of good citizenship (International Association for the Evaluation of Educational Achievement, IEA) data from: 1999, 2009 and 2016 [5].

The twelve identical items in all three waves range from obeying the law and voting in elections, to protecting the environment and defending human rights.

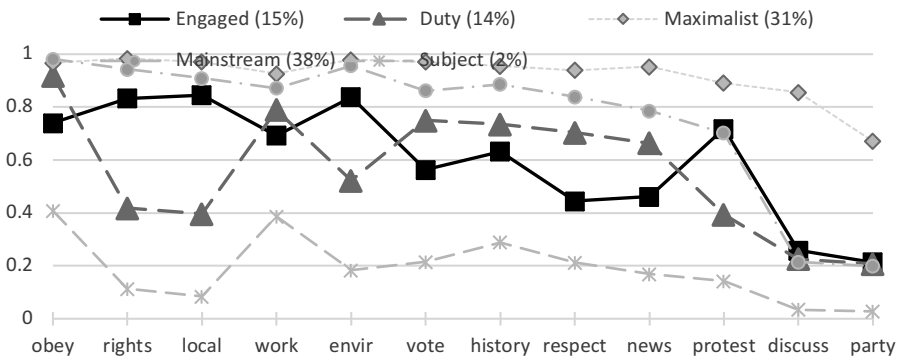
Our analyses focus on the 14 countries in which the IEA survey was conducted in all three observation periods[1]. Due to the heterogeneity of the included countries, it is imperative to test for measurement equivalence [5] to allow for a valid regression analysis that accounts for the multi-level structure of the data of individuals (level 1) nested in countries (level 2).

Our main theoretical focus is on the change of citizenship over time, controlling for standard socio-demographic covariates. Analyses were conducted using R 4.0.3, in combination with Latent Gold 5.1.

Figure 1 plots the relative emphases of the five LCs identified in the optimal model. The LC analysis specification included covariates for country and for year, and measurement equivalence tests (see Appendix A1). The findings identify a group of respondents (14%) who adhere to a duty-based norm, to an engaged group (15%) but also additional groups: a "maximalist" group (31%) that places high importance on all indicators; a "mainstream" group (38%) that reflects mean levels of importance; and a "subject" group (2%) that consistently attributes low importance to all items.

The step two multinomial logistic regression results in Table 1 use the mainstream norm as the reference category. The engaged and duty-based norms have remained stable in size over the three observation periods. However, the other norms identified in the analysis do change significantly in prevalence over time: the mainstream norm becomes less prevalent, while the maximalist and subject norms increase in relative size. There is relatively little distinction in the socio-economic status of adolescents in the different groups - these were therefore dropped form the table.

Figure 1. Distinctive citizenship norms: Latent class analysis results in 14 countries[2]



[1] Bulgaria, Chile, Colombia, Denmark, Estonia, Finland, Hong Kong, Italy, Latvia, Lithuania, Norway, Russia, Slovenia, and Sweden

[2] Citizenship norm items: Obeying the law (obey), taking part in activities promoting human rights (rights), participating in activities to benefit people in the community (local), working hard (work), taking part in activities to protect the environment (envir), voting in every election (vote), learning about the country's history (history), showing respect for government

Table 1: Citizenship norms: Multinomial logistic regression, mainstream group as reference

	Engaged (SE)	**Duty (SE)**	**Maximalist (SE)**	**Subject (SE)**
Intercept	-0.16(0.46)	-0.96(0.56)	-0.20 (0.18)	-2.26 (0.40) **
Year 2009	0.03 (0.33)	0.10 (0.26)	0.55 (0.11) **	1.88(0.45) **
2016	0.13 (0.33)	0.33 (0.27)	0.97(0.11) **	3.18(0.53) **
Female	-0.34 (0.11) **	-0.37 (0.09) **	-0.47 (0.06) **	-1.39 (0.21) **

Notes: IEA data in 14 countries pooled over 1999, 2009, and 2016; n=137,499. Control variables (gender, socio-economic status indicators (i.e., SES proxy of books at home, educational expectation, and parent's education) dropped from presentation for brevity

4 Discussion

The current study draws on high-quality longitudinal data to identify trends over time regarding changing citizenship norms. We proposed the use of the two-step LC estimator [1] as a highly appropriate modelling approach to answer the research questions that motivated our work. The methodology allows testing for measurement equivalence across 14 countries, and over three observation periods. Furthermore, by separating the measurement and structural model, we can separately focus on our two main research questions: the prevalence of the duty based and engaged citizenship norms in the 14 studied societies, and the change over time in this prevalence. While testing and modelling for measurement equivalence is very common for factor analytic models, these tests are less developed for LCA. As such, future research is needed to better understand the effect of ignoring measurement non-equivalence on the structural parameters, and on the power of detecting MI with local fit statistics.

Appendix A1. Latent class measurement equivalence tests

Models	LL	BIC(LL)	Npar	L^2	df	Class.Err.
Homogeneous model	-734275	1470016	124	142117	137375	0.19
Heterogeneous model	-710098	1450768	2584	93763	134915	0.23
Partial equivalence	**-715140**	**1437567**	**616**	**103847**	**136883**	**0.24**

Source: IEA data in 14 countries for 1999, 2009, and 2016; n=137,499.

References

representatives (respect), following the political issues in the newspaper on the radio or on tv (news), participating in a peaceful protest against a law believed to be unjust (protest), engaging in political discussions (discuss), and joining a political party (party).

[1] Bakk, Z., and J. Kuha. 2018. "Two-Step Estimation of Models between Latent Classes and External Variables." *Psychometrika* 83 (4): 871-892.

[2] Dalton, R., and C. Welzel, eds. 2014. *The Civic Culture Transformed: From Allegiant to Assertive Citizens*. Cambridge: Cambridge University Press.

[3] Inglehart, R., and P. Norris. 2017. "Trump and the Populist Authoritarian Parties: The Silent Revolution in Reverse." *Perspectives on Politics* 15 (2): 443-454.

[4] Kankaraš, M., G. Moors, and J. K. Vermunt. 2011. "Testing for Measurement Equivalence with Latent Class Analysis." In *Cross-Cultural Analysis: Methods and Applications*, edited by Eldad Davidov, Peter Schmidt and Jaak Billiet, 359-384. Routledge.

[5] Schulz, W., J. Ainley, and J. Fraillon, eds. 2011. *ICCS 2009 Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement

Group-wise penalized estimation schemes in model-based clustering

Strategie di stima penalizzata a livello di gruppo nel clustering basato su modello

Alessandro Casa, Andrea Cappozzo and Michael Fop

Abstract Gaussian mixture models provide a probabilistically sound clustering approach. However, their tendency to be over-parameterized endangers their utility in high dimensions. To induce sparsity, penalized model-based clustering strategies have been explored. Some of these approaches, exploiting the link between Gaussian graphical models and mixtures, allow to handle large precision matrices, encoding variables relationships. By assuming sparsity levels similar across components, these methods fall short when the dependence structures are group-dependent. Our proposal, by penalizing group-specific transformations of the precision matrices, automatically handles situations where under or over-connectivity between variables is present. The performance of the method is shown via a real data experiment.

Abstract *La sovra-parametrizzazione dei modelli di mistura Gaussiani, che rappresentano un approccio probabilistico al clustering, mette a rischio la loro utilità in dimensioni elevate. Per questo motivo sono state proposte strategie di stima penalizzate che permettono di gestire matrici di precisioni di grandi dimensioni, sfruttando il legame tra modelli grafici Gaussiani e modelli mistura. Questi metodi, assumendo sparsità simile tra tutte le componenti, falliscono quando la struttura di dipendenza varia di gruppo in gruppo. La nostra proposta, penalizzando una trasformazione delle matrici di precisione differente per ogni componente, gestisce situazioni in cui il numero di connessioni tra le variabili è diverso tra i gruppi. La validità del metodo è evidenziata grazie ad un'applicazione a dati reali.*

Alessandro Casa
Faculty of Economics and Management, Free University of Bozen-Bolzano
e-mail: alessandro.casa@unibz.it

Andrea Cappozzo
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano
e-mail: andrea.cappozzo@polimi.it

Michael Fop
School of Mathematics and Statistics, University College Dublin
e-mail: michael.fop@ucd.ie

Key words: Model-based clustering, Graphical lasso, EM algorithm, Gaussian graphical models

1 Introduction

Model-based clustering [2] represents a widely known and probabilistic-based strategy to cluster analysis. Here, the data generative mechanism is assumed to be adequately described by means of finite mixture models, with the Gaussian distribution being commonly considered as the component density when dealing with continuous data. Partitions are then practically obtained by drawing a one-to-one correspondence between mixture components and groups.

While being fruitfully adopted in a lot of different applications, one of the major shortcomings of this approach lies in its tendency to be over-parameterized in high-dimensional spaces. In fact, the number of parameters to estimate scales quadratically with the number of the observed variables, endangering the practical applicability of the method in some scenarios. To overcome this issue, several different approaches have been proposed in the literature (see [1] for a review on the topic).

Here, we focus specifically on a class of strategies that aims to induce parsimony by adopting penalized estimation schemes. More specifically, in [6] the number of association parameters to be estimated is drastically reduced by penalizing the component precision matrices via a graphical lasso penalty. Conveniently, zero entries in these matrices imply conditional independence between the corresponding variables, and the dependence structure can be visually represented by means of Gaussian graphical models. The adoption of a common shrinkage factor for all the component implies that the number of non-zero entries is similar across precision matrices for different components. This assumption can hinder group discrimination as it can be quite restrictive in those settings where the associations between the variables show cluster-dependent patterns.

To overcome this drawback, in this work we propose a generalization of the approach by [6], where we penalize component-specific transformations of the precision matrices rather than the matrices themselves. As a result, our method turns out to be more flexible and adaptive, without requiring the specification of additional hyper-parameters, as it automatically encompasses those situations where under or over-connectivity is witnessed in the class-specific graphical models. The rest of the paper is structured as follows. In Section 2 we outline the proposal, while in Section 3 we show the validity of the approach by applying it on a real data example. Lastly, in Section 4 we conclude with some remarks and highlighting possible future research directions.

2 Proposed methodology

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathbb{R}^p$, be the set of the observed data. Coherently with the model-based formulation, to cluster the data into K different groups, we consider Gaussian mixture models. Given the considerations in the previous section, the parameters of the model are estimated by maximizing a penalized log-likelihood function which reads as:

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} ||\mathbf{P}_k * \Omega_k|| \tag{1}$$

where π_k's denote the mixing proportions, with $\pi_k > 0, \forall k$ and $\sum_k \pi_k = 1$, and $\phi(\cdot; \mu_k, \Omega_k)$ is the density of a multivariate Gaussian distribution with mean vector $\mu_k \in \mathbb{R}^p$ and $p \times p$ precision matrix Ω_k. Therefore, the first term in (1) represents the log-likelihood of a Gaussian mixture model, while the second one introduces the graphical lasso penalty, with shrinkage hyper-parameter λ. More specifically, with $||\cdot||$ we denote the element-wise L_1 norm, with $*$ the Hadamard product, and $\mathbf{P}_k$s are the matrices that drive the component-specific transformation of Ω_k. By penalizing the elements of $\mathbf{P}_k * \Omega_k$, instead of the ones in Ω_k as conversely done in [6], we scale the effect of λ and we uncover group-wise conditional dependence structures among the variables. As a consequence, our approach automatically encompasses those settings where the number of non-zero entries in Ω_ks is dissimilar.

Since the $\mathbf{P}_k$s encode information about class-specific sparsity patterns, they play a pivotal role in our proposal, therefore the focus is shifted towards their specification. We adopt a data-driven procedure, relying on estimated sample precision matrices $\hat{\Omega}_1^{(0)}, \ldots, \hat{\Omega}_K^{(0)}$, obtained conditionally on carefully initialized partitions. The weight matrices are then defined as $\mathbf{P}_k = f(\hat{\Omega}_k^{(0)})$, with $f : \mathbb{S}_+^p \to \mathbb{S}^p$ a function from the space of positive semi-definite matrices to the space of p-dimensional symmetric matrices.

Hereafter we describe two viable options to define $f(\cdot)$. Nonetheless, we are aware that different routes can be taken when specifying $f(\cdot)$, with subjectivity and prior information potentially playing a relevant role in the process.

- According to the first proposal, which can be seen as a multiclass generalization of the approach by [4], $\mathbf{P}_k$ is defined as

$$P_{k,ij} = 1/|\hat{\Omega}_{k,ij}^{(0)}| \tag{2}$$

  with $P_{k,ij}$ and $\hat{\Omega}_{k,ij}^{(0)}$ denoting the (i,j)-th elements of the matrices $\mathbf{P}_k$ and $\hat{\Omega}_k^{(0)}$ respectively. This specification allows to inflate/deflate the penalty terms on the elements of Ω_k according to the element-wise magnitude of $\hat{\Omega}_k^{(0)}$. In fact, when $|\hat{\Omega}_{k,ij}^{(0)}|$ is close to 0, $P_{k,ij}$ would impose an extra shrinkage on $\Omega_{k,ij}$.
- The second proposal sets the elements of $\mathbf{P}_k$ proportional to the distance between $\hat{\Omega}_k^{(0)}$ and $\text{diag}(\hat{\Omega}_k^{(0)})$, where $\text{diag}(\hat{\Omega}_k^{(0)})$ is a diagonal matrix whose elements are

Table 1 ARI, number of estimated parameters d_Ω and Median Frobenius Distance, for different penalized model-based clustering methods.

	ARI	d_Ω	MFD
Zhou et al.(2009)	0.6724	320	830
$\mathbf{P}_k$ as in (2)	0.7199	242	421
$\mathbf{P}_k$ as in (3), Frobenius	0.6875	312	701
$\mathbf{P}_k$ as in (3), Riemannian	0.6812	314	798

equal to the ones in $\hat{\Omega}_k^{(0)}$. The entries of $\mathbf{P}_k$ are computed as

$$P_{k,ij} = \frac{1}{\mathscr{D}\left(\hat{\Omega}_k^{(0)}, \text{diag}\left(\hat{\Omega}_{\text{k}}^{(0)}\right)\right)}, \tag{3}$$

for $i, j = 1, \ldots, p$. With $\mathscr{D}(\cdot,\cdot)$ we denote a suitable measure of distance between positive semi-definite matrices. Since $\mathbb{S}_+^p$ is a non-Euclidean space, we employ Frobenius and Riemannian distances (see [3] for a detailed discussion).

These two strategies share the same rationale, as they aim to penalize more strongly those entries corresponding to weaker sample conditional dependencies. Nonetheless, while for the second approach $\mathbf{P}_k$s depend on a group specific constant, in the first one the induced penalty is entry-wise different, thus possibly more accurate when the sample estimates $\hat{\Omega}_k^{(0)}$s are regarded as reliable. Lastly note that in [6] $\mathbf{P}_k$ is assumed to be a matrix of ones for all $k = 1, \ldots, K$.

Once $\mathbf{P}_1, \ldots, \mathbf{P}_K$ are specified, the model is estimated employing an EM-algorithm with the graphical lasso embedded in the M-step, when estimating sparse precision matrices.

3 Application

Our proposal is here employed on the Olive Oil dataset, which is publicly available in the `R` package `pgmm` [5]. The data report the percentage composition of $p = 8$ fatty acids for $n = 572$ samples of olive oil, coming from $K = 9$ different regions in Italy. The aim of the analyses consists in recovering the group structure, given by the geographical partition, of the oils by using their lipidic characteristics.

In the analyses we compare our method, considering different specifications of $\mathbf{P}_k$ as outlined in the previous section, with the strategy proposed by [6]. Different competitors are compared in terms of clustering performances via the Adjusted Rand Index (ARI). Moreover, we evaluate also the number of non-zero parameters d_Ω, as a proxy of model complexity, and the Median Frobenius Distance (MFD) computed as:

$$\text{median}_{k=1,\ldots,K}\left(||\hat{\Omega}_k - \bar{\Omega}_k||_F\right)$$

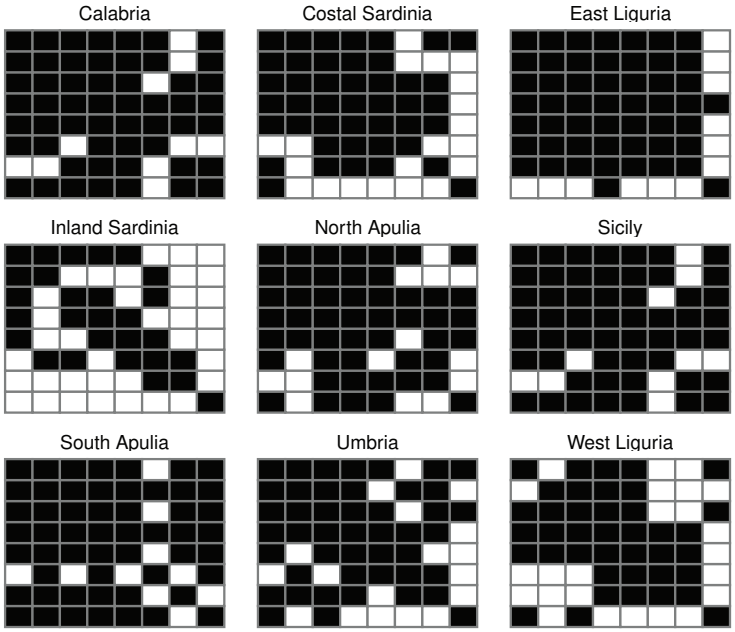


Fig. 1 Estimated precision matrices, with $\mathbf{P}_k$ defined as in (2). Black squares denote the presence of an edge between the two variables.

where $||\cdot||_F$ denotes the Frobenius norm, while $\bar{\Omega}_k$ is the k-th component empirical precision matrix, computed using the true labels, which allows to evaluate how the model identifies the conditional association structure among the variables. The results are reported in Table 1.

We immediately note that, including a data-driven specification for $\mathbf{P}_k$ slightly improves the clustering performance with respect to the all-one matrix as in [6]. Furthermore, our proposals are able to obtain a reduction in the total number of non-zero parameters d_Ω, especially when defining the weight matrices as in (2). This latter approach appears to be the best one also when considering the Median Frobenius Distance, thus when evaluating how good the method is in recovering the true conditional dependencies. Figure 1 displays the component precision matrices estimated using this method; from here we see that the association structure varies appreciably across regions, with our proposal exploiting this behaviour in the estimation step.

4 Conclusion and discussion

In this work we showed how, in the penalized clustering framework, partitions retrieval can be jeopardized when imposing a single penalty on the component preci-

sion matrices. In fact, automatically enforcing similarities in the estimated graphical models across groups, this can be harmful when it comes to groups discrimination.

More specifically, we have proposed a generalization of the approach outlined in [6]. Here the authors, by considering a single penalization parameter, implicitly assume that all the groups present a similar degree of sparsity. Therefore, this method does not account for those situations where one or more components shows under or over-connectivity with respect to the others. For this reason, we have devised a procedure which penalizes a group-specific transformation of the component precision matrices. The proposal automatically encompasses situations where the groups are characterized by a different amount of non-zero entries in the corresponding precision matrices. In our work, we proposed several different ways to define the transformed precision matrices to be penalized. Numerical explorations on real data have confirmed the validity of the method.

Lastly note that, while outlined for Gaussian mixtures parameterized in terms of precision matrices, this penalized approach can be fruitfully generalized to component covariance matrices. Moreover, if paired with a carefully chosen penalization term on the component means, this methodology can be used to perform variable selection in the model-based clustering context.

References

1. Bouveyron, C., Brunet-Saumard, C. Model-based clustering of high-dimensional data: A review. Comput Stat Data An, **71**, 52-78 (2014)
2. Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E. Model-based clustering and classification for data science: with applications in R. Cambridge University Press (2019)
3. Dryden, I.L., Koloydenko, A., Zhou, D. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann Appl Stat, **3(3)**, 1102-1123 (2009)
4. Fan, J., Feng, Y., Wu, Y. Network exploration via the adaptive LASSO and SCAD penalties. Ann Appl Stat, **3(2)**, 521-541 (2009)
5. McNicholas, P.D., ElSherbiny, A., McDaid, A.F., Murphy, T.B. pgmm: Parsimonious Gaussian Mixture Models. R package version 1.2.4 (2019)
6. Zhou, H., Pan, W., Shen, X. Penalized model-based clustering with unconstrained covariance matrices. Electron J Stat, **3**, 1473-1496 (2009)

Extending finite mixtures of latent trait analyzers for bipartite networks

Estensione di un modello a mistura finita per reti bipartite

Dalila Failli, Maria Francesca Marino, and Francesca Martella

Abstract The paper extends the Mixture of Latent Trait Analyzers (MLTA) for clustering bipartite networks to account for nodal attributes. Bipartite networks are particularly useful to represent relations between disjoint sets of nodes, called sending and receiving nodes. The MLTA model is able not only to cluster the sending nodes of a bipartite network, but also capture the latent variability of network connections within each group. We extend this approach by including nodal attributes to study how nodes' characteristics affect the group membership probability. A simulation study is conducted to evaluate the proposed approach.

Abstract *L'articolo estende il modello Mixture of Latent Trait Analyzers (MLTA) per il clustering di reti bipartite, tenendo conto degli attributi nodali. Le reti bipartite sono utili per rappresentare le relazioni tra due insiemi disgiunti di nodi, chiamati sending e receiving nodes. Il modello MLTA è in grado non solo di raggruppare i sending nodes di una rete bipartita, ma anche di catturare la variabilità latente delle connessioni tra sending e receiving nodes. Questo approccio viene esteso includendo gli attributi nodali nella parte latente del modello con l'obiettivo di valutare se e come le caratteristiche dei nodi influenzano la probabilità di appartenenza al gruppo. Uno studio di simulazione è stato condotto con l'obiettivo di valutare la bontà dell'approccio proposto.*

Key words: Model-based clustering, Network data, Nodal attributes, EM algorithm, Variational inference

---

Dalila Failli
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze, e-mail: dalila.failli@unifi.it

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze, e-mail: mariafrancesca.marino@unifi.it

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Roma, e-mail: francesca.martella@uniroma1.it

1 Introduction

Over the years, mathematical and computational tools have been developed to analyze networks. Networks are collections of interconnected units (nodes) that can capture interactions within a system. Several social, technological, and biological processes can be represented as networks. Consequently, network data analysis is attractive in different research fields, both applied and theoretical. Bipartite networks are a particular type of networks, which represent the relations between two disjoint sets of nodes, called sending and receiving nodes.
A relevant aspect of network analysis concerns the identification of clusters of nodes characterized by similar behaviors. For this purpose, latent class models [2] and stochastic blockmodels [13] are frequently applied within a model-based clustering framework. A different modelling approach is that based on the so-called Mixture of Latent Trait Analyzers (MLTA) [9], which can be effectively employed for clustering bipartite networks [8]. Such a model is obtained by combining features of latent trait and latent class analysis. In detail, it assumes that sending nodes can be grouped into homogeneous classes (or groups), as in the latent class model. Together with the group membership, the propensity of each sending node to be connected with the receiving nodes depends also on the presence of a multi-dimensional continuous latent variable, as in the latent trait framework. This latter variable allows us to capture the latent variability of network connections within each group. Therefore, compared to the latent trait model, the MLTA allows to consider latent classes (or groups) of nodes sharing some unobserved characteristics. Compared to the latent class model, the MLTA allows to capture the latent variability of the connections between sending and receiving nodes within each group. In addition, when dealing with large networks, the assumption of local independence upon which the latent class model is based may lead to the identification of too many groups, making the interpretation of the results difficult. Conversely, the MLTA model allows to overcome such an issue.
In this paper, we extend the MLTA approach by accounting for the effect that the characteristics of sending nodes, called nodal attributes, may have on the clustering formation. In this respect, a multinomial logit specification for the prior probabilities of the finite mixture is considered.
The paper is organized as follows: in Section 2, we extend the MLTA model to account for nodal attributes, also describing model assumptions, parameter estimation, and model selection. Section 3 shows the results of a simulation study conducted in order to verify the efficacy of the proposed approach. Section 4 contains concluding remarks and further extensions of the approach.

2 Model-based clustering for bipartite networks

Let $\mathscr{N} = \{n_1, n_2, \ldots, n_N\}$ denote the set of sending nodes and $\mathscr{R} = \{r_1, r_2, \ldots, r_R\}$ the set of receiving nodes. Note that the terms "sending" and "receiving" do not

refer to nodes that actually "send" a connection and to those who actually "receive" a connection, but are simply used to distinguish the two non-overlapping sets. The relationship structure of a bipartite network can be formally described by a random matrix $\mathbf{Y} = \{Y_{ik}\}$, called *incidence matrix*, whose generic element is given by

$$Y_{ik} = \begin{cases} 1 \text{ if sending node } n_i \text{ is connected with receiving node } r_k, \\ 0 \text{ otherwise.} \end{cases}$$

Model-based clustering can be used for grouping the sending nodes of a bipartite network. In particular, [8] propose to extend the Mixture of Latent Trait Analyzers (MLTA) [9] in the context of bipartite networks. The model broadens the latent class and latent trait analysis by assuming that a set of N sending nodes can be divided into G distinct classes (or groups), and that the propensity of each sending node to be connected with the R receiving nodes depends on both the group membership and on a multi-dimensional continuous latent variable. Our contribution is to further extend the MLTA for bipartite networks to account for nodal attributes in the latent model structure.

2.1 Model assumptions

The MLTA for bipartite networks treats the sending nodes as observations and the receiving nodes as observed variables. The model assumes that every sending node belongs to an unobserved group identified by the latent random variable $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})' \sim \text{Multinomial}(1, (\eta_1, \ldots, \eta_G))$, whose generic element is given by

$$z_{ig} = \begin{cases} 1 \text{ if sending node } n_i \text{ belongs to group } g, \\ 0 \text{ otherwise.} \end{cases}$$

The parameter η_g denotes the probability that a randomly selected sending node belongs to group g, with $g = 1, \ldots, G$, under the constraints that $\sum_{g=1}^{G} \eta_g = 1$ and $\eta_g \geq 0$.
Furthermore, the model assumes that the conditional distribution of the vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iR})$, given that node n_i belongs to the g-th group, is specified by a latent trait model with parameters b_{gk} and $\mathbf{w}_{gk}$, $g = 1, \ldots, G$, and $k = 1, \ldots, R$. In this sense, denoting with $\mathbf{u}_i$ a D-dimensional continuous latent variable, conditional on $\mathbf{z}_i$ and $\mathbf{u}_i$, response variables contained in the $\mathbf{y}_i$ vector are assumed to be independent Bernoulli random variables with parameters $\pi_{gk}(\mathbf{u}_i)$, $k = 1, \ldots, R$, modelled via the following logistic function:

$$\pi_{gk}(\mathbf{u}_i) = p(y_{ik} = 1 \mid \mathbf{u}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-(b_{gk} + \mathbf{w}'_{gk}\mathbf{u}_i)]}, \quad 0 \leq \pi_{gk}(\mathbf{u}_i) \leq 1. \tag{1}$$

Here, b_{gk} is the model intercept and represents the *attractiveness* of the receiving node r_k for the sending nodes belonging to the g-th group. On the other side, $\mathbf{w}_{gk}$ are the slopes associated with the latent variable $\mathbf{u}_i$ and are meant to capture the *heterogeneity* in the behavior of sending nodes belonging to the g-th group in the way they connect to the receiving node r_k. Larger values for these parameters indicate a greater difference for the sending nodes belonging to the g-th group in the probability of creating a connection with the receiving node r_k. On the other hand, the choice of a model with constant $\mathbf{w}_{gk}$ parameters across groups ($\mathbf{w}_{k1} = \mathbf{w}_{k2} = \ldots = \mathbf{w}_k$) suggests that the latent trait has the same effect in all groups.

Note that the conditional probability in equation (1) is an increasing function of $b_{gk} + \mathbf{w}'_{gk}\mathbf{u}_i$. If $\mathbf{w}_{gk} = \mathbf{0}$, response variables y_{ik}, $k = 1, \ldots, R$, do not depend on the latent trait, and the model identifies the simplest situation of independence between the response variables, conditional on the group membership only. Furthermore, as highlighted by [9], the slope parameters $\mathbf{w}_{gk}$ are only identifiable up to a rotation of the factors.

The model is completed by assuming that the continuous D-dimensional latent trait $\mathbf{u}_i$ is distributed according to a Gaussian density with null mean vector and identity covariance matrix, i.e. $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{I})$.

Therefore, the MLTA model allows to cluster the observations according to a categorical latent variable, as in the latent class model. In addition, the residual dependence between the response variables associated to a given sending node is fully explained by a continuous latent variable, as in the latent trait model. Consequently, the MLTA model represents an extension of the latent class and the latent trait models and includes such models as particular cases. In fact, when the dimension of the continuous latent variable is equal to zero ($D = 0$), the MLTA corresponds to a latent class model. On the other hand, when there are no groups, namely $G = 1$ and $D > 0$, the MLTA coincides with a latent trait model.

To account for the possible effect that nodal attributes may have on the clustering of nodes, we propose to relax the homogeneity assumption of the latent class prior probabilities η_g. In detail, we let them vary according to the observed nodal features by specifying a *latent class regression* model [6]. Let $\mathbf{x}_i$ denote the J-dimensional vector of nodal attributes for the sending node n_i, and let β_g denote the corresponding J-dimensional vector of coefficients for the g-th latent class. We assume that prior probabilities η_{ig} are related to the observed $\mathbf{x}_i$'s via a generalized (multinomial) logit link function [1]:

$$\eta_{ig} = \frac{\exp\{\mathbf{x}_i \beta_g\}}{\sum_{g'=1}^{G} \exp\{\mathbf{x}_i \beta_{g'}\}}, \quad g = 2, \ldots, G.$$

Note that the first latent class is the reference class, so that $\beta_1 = \mathbf{0}$.

2.2 Parameters estimation

Let $\theta = (\beta_2, \ldots, \beta_G, b_{11}, \ldots, b_{GR}, \mathbf{w}_{11}, \ldots, \mathbf{w}_{GR}))$ represent the vector of all model parameters. Starting from the model assumptions detailed in the previous section, the model log-likelihood function can be written as:

$$\ell(\theta) = \sum_{i=1}^{N} \log \Big(\sum_{g=1}^{G} \eta_{ig} \int \prod_{k=1}^{R} p(y_{ik} \mid \mathbf{u}_i, z_{ig} = 1) p(\mathbf{u}_i) d\mathbf{u}_i \Big), \tag{2}$$

where $p(y_{ik} \mid \mathbf{u}_i, z_{ig} = 1) = \prod_{k=1}^{R} (\pi_{gk}(\mathbf{u}_i))^{y_{ik}} (1 - \pi_{gk}(\mathbf{u}_i))^{1-y_{ik}}$. From equation (2), it is evident that the model corresponds to a finite mixture of latent trait models with node-specific proportions η_{ig}.
The integral to be solved in equation (2) in order to derive the log-likelihood function cannot be computed analytically. To overcome this issue, an indirect estimation approach based on the EM algorithm can be used. In detail, [9] propose to use a double EM algorithm with a *variational approach* [15] to approximate the likelihood function.
The objective of the variational approach is to maximize a lower bound of the likelihood function, which is a function also of auxiliary parameters (variational parameters) $\xi_{ig} = (\xi_{i1g}, \ldots, \xi_{iRg})$, with $\xi_{igk} \neq 0$, $\forall k = 1, \ldots, R$. As highlighted in [11], the logarithm of the component densities $p(\mathbf{y}_i \mid z_{ig} = 1)$ may be approximated by the following lower bound:

$$\begin{aligned} \mathscr{L}(\xi_{ig}) &= \log(\tilde{p}(\mathbf{y}_i \mid z_{ig} = 1, \xi_{ig})) \\ &= \log \Big(\int \prod_{k=1}^{R} \tilde{p}(y_{ik} \mid \mathbf{u}_i, z_{ig} = 1, \xi_{igk}) p(\mathbf{u}_i) d\mathbf{u}_i \Big), \end{aligned}$$

where

$$\tilde{p}(y_{ik} \mid \mathbf{u}_i, z_{ig} = 1, \xi_{igk}) = \sigma(\xi_{igk}) \exp\Big(\frac{A_{igk} - \xi_{igk}}{2} + \lambda(\xi_{igk})(A_{igk}^2 - \xi_{igk}^2) \Big),$$

$$\sigma(\xi_{igk}) = (1 + \exp(-\xi_{igk}))^{-1},$$

$$A_{igk} = (2y_{ik} - 1)(b_{gk} + \mathbf{w}'_{gk}\mathbf{u}_i),$$

$$\lambda(\xi_{igk}) = \Big(\frac{1}{2} - \sigma(\xi_{igk}) \Big) / 2\xi_{igk}.$$

To estimate model parameters, we proceed by iteratively alternating the steps described below.

1. The E-step consists of computing the expected value of the complete data log-likelihood function, given the observed data and the current value of the parameter estimates. This is equivalent to computing the posterior probabilities of the latent variables z_{ig} as follows:

$$\hat{z}_{ig}^{(t+1)} = \frac{\hat{\eta}_{ig}^{(t)} \exp(\mathscr{L}(\hat{\xi}_{ig}^{(t)}))}{\sum_{g'=1}^{G} \hat{\eta}_{ig'}^{(t)} \exp(\mathscr{L}(\hat{\xi}_{ig'}^{(t)}))}.$$

2. In the M-step, the multinomial logit coefficients β_g and the prior probabilities η_{ig} of each component of the finite mixture are estimated [4]. The coefficients $\hat{\beta}_g$ are obtained by maximizing the likelihood of a multinomial logit model, with weights provided by the posterior group membership probabilities $\hat{z}_{ig}$ derived at the previous step, via a Newton-Raphson step. Prior probabilities are updated accordingly.
3. The likelihood function is approximated using a second EM algorithm nested within the first.

   - In the E-step, we identify the sufficient statistics for the approximate posterior distribution of the continuous latent variables $\mathbf{u}_i$, given the observations $\mathbf{y}_i$, the posterior probabilities $\hat{z}_{ig}$ computed at step $(t+1)$, and the variational parameters $\hat{\xi}_{ig}$ estimated at the t-th step of the algorithm:

   $$\tilde{p}(\mathbf{u}_i \mid \mathbf{y}_i, \hat{z}_{ig}^{(t+1)} = 1, \hat{\xi}_{ig}^{(t)}).$$

   This corresponds to a Gaussian distribution, with covariance matrix and mean vector given by

   $$\hat{\mathbf{C}}_{ig}^{(t+1)} = \Big[\mathbf{I} - 2\sum_{k=1}^{R} \lambda(\hat{\xi}_{igk}^{(t)}) \hat{\mathbf{w}}_{gk}^{(t)} \hat{\mathbf{w}}_{gk}^{(t)\prime}\Big]^{-1},$$

   $$\hat{\mu}_{ig}^{(t+1)} = \hat{\mathbf{C}}_{ig}^{(t+1)} \Big[\sum_{k=1}^{R} \Big(y_{ik} - \frac{1}{2} + 2\lambda(\hat{\xi}_{igk}^{(t)}) \hat{b}_{gk}^{(t)}\Big) \hat{\mathbf{w}}_{gk}^{(t)}\Big].$$

   - In the M-step, the variational parameters ξ_{ig} are estimated by ensuring that $\tilde{p}(\mathbf{y}_i \mid \hat{z}_{ig}^{(t+1)} = 1, \hat{\xi}_{ig}^{(t+1)})$ is as close as possible to the true $p(\mathbf{y}_i \mid \hat{z}_{ig}^{(t+1)} = 1), \forall \mathbf{y}_i$. To this end, the expected value of the complete data log-likelihood function is maximized with respect to each $\xi_{igk} \in \xi_{ig}, k = 1, \ldots, R$:

   $$Q(\hat{\xi}_{igk} \mid \hat{\xi}_{igk}^{(t)}) = E[\log \tilde{p}(y_{ik} \mid \mathbf{u}_i, \hat{z}_{ig} = 1, \hat{\xi}_{igk}) p(\mathbf{u}_i)]. \tag{3}$$

   The maximum is achieved for:

   $$\begin{aligned} \hat{\xi}_{igk}^{(t+1)2} &= E[(b_{gk} + \mathbf{w}'_{gk}\mathbf{u}_i)^2] \\ &= \hat{\mathbf{w}}_{gk}^{(t)\prime} (\hat{\mathbf{C}}_{ig}^{(t+1)} + \hat{\mu}_{ig}^{(t+1)} \hat{\mu}_{ig}^{(t+1)\prime}) \hat{\mathbf{w}}_{gk}^{(t)} + 2\hat{b}_{gk}^{(t)} \hat{\mathbf{w}}_{gk}^{(t)\prime} \hat{\mu}_{ig}^{(t+1)} + \hat{b}_{gk}^{(t)2}. \end{aligned}$$

   - Let $\hat{\zeta}_{gk}^{(t+1)} = (\mathbf{w}_{gk}^{(t+1)\prime}, b_{gk}^{(t+1)})'$ and $\hat{\mu}_{ig}^{(t+1)} = (\mu_{ig}^{(t+1)\prime}, 1)'$; updates for $\mathbf{w}_{gk}$ e b_{gk} are obtained as:

$$\hat{\zeta}_{gk}^{(t+1)} = -\Big[2\sum_{i=1}^{N}\hat{z}_{ig}^{(t+1)}\lambda(\hat{\xi}_{igk}^{(t+1)})E[\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i']_g^{(t+1)}\Big]^{-1}\Big[\sum_{i=1}^{N}\hat{z}_{ig}^{(t+1)}\Big(y_{ik}-\frac{1}{2}\Big)\hat{\mu}_{ig}^{(t+1)}\Big],$$

where

$$E[\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i']_g^{(t+1)} = \begin{bmatrix} \hat{\mathbf{C}}_{ig}^{(t+1)} + \hat{\mu}_{ig}^{(t+1)}\hat{\mu}_{ig}^{(t+1)\prime} & \hat{\mu}_{ig}^{(t+1)} \\ \hat{\mu}_{ig}^{(t+1)\prime} & 1 \end{bmatrix}.$$

- The lower bound of the component densities and the log-likelihood function are approximated as follows:

$$\begin{aligned}
\mathscr{L}(\hat{\xi}_{ig}^{(t+1)}) = \sum_{k=1}^{R} & \Big[\log(\sigma(\hat{\xi}_{igk}^{(t+1)})) - \frac{\hat{\xi}_{igk}^{(t+1)}}{2} - \lambda(\hat{\xi}_{igk}^{(t+1)})\hat{\xi}_{igk}^{(t+1)2} + \Big(y_{ik}-\frac{1}{2}\Big)\hat{b}_{gk}^{(t+1)} \\
& + \lambda(\hat{\xi}_{igk}^{(t+1)})\hat{b}_{gk}^{(t+1)2}\Big] + \frac{\log|\hat{\mathbf{C}}_{ig}^{(t+1)}|}{2} + \frac{\hat{\mu}_{ig}^{(t+1)\prime}[\hat{\mathbf{C}}_{ig}^{(t+1)}]^{-1}\hat{\mu}_{ig}^{(t+1)}}{2}, \\
\ell(\hat{\theta}^{(t+1)}) & \approx \sum_{i=1}^{N}\log\Big(\sum_{g=1}^{G}\hat{\eta}_{ig}^{(t+1)}\exp(\mathscr{L}(\hat{\xi}_{ig}^{(t+1)}))\Big).
\end{aligned}$$

The procedure is repeated until convergence.
Since the approximation of the log-likelihood function obtained via the variational approach is always less than or equal to the true log-likelihood function, it may be useful to derive a more accurate approximation at the last step of the algorithm via a Gauss-Hermite quadrature.
Furthermore, the estimates obtained at the convergence of the algorithm may coincide with a local maximum rather than the global one. Thus, it is recommended to run the algorithm several times with different initial values of the parameters, and choose the optimal solution as the one corresponding to the maximum value of the likelihood function.
After estimating model parameters, each observation can be assigned to one of the G groups on the basis of the estimated posterior probability $\hat{z}_{ig}$ via a Maximum a Posteriori (MAP) rule.

2.3 Standard errors and model selection

To evaluate the uncertainty associated with the estimates obtained from the EM algorithm, a *non-parametric bootstrap* [7] is proposed. Given an incidence matrix $\mathbf{Y}$ with N sending nodes and R receiving nodes, this method consists in extracting with repetition S samples from the incidence matrix $\mathbf{Y}$, where each sample has the same size of $\mathbf{Y}$. In detail, for each bootstrap sample, N rows of the incidence matrix are drawn with repetition, so that each sending node can appear several times.
Let $\hat{\theta}_{(s)}$ denote the vector of estimators obtained from the s-th bootstrap sample. By paying attention to the ordering of the parameters at each iteration, bootstrap standard errors correspond to the square root of the diagonal elements of the following

matrix:

$$\mathrm{V}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\theta}_{(s)} - \hat{\theta}_{(.)}\right)\left(\hat{\theta}_{(s)} - \hat{\theta}_{(.)}\right)',$$

where $\hat{\theta}_{(.)}$ is the empirical mean vector $\hat{\theta}_{(.)} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_{(s)}$.
The number of latent classes G, as well as the size D of the continuous latent variable, are not considered as model parameters, but rather as quantities to be fixed a priori. To identify the optimal model, the MLTA model is estimated for several values of G and D. The model corresponding to the smallest value of the chosen information criterion, such as the *Bayesian information criterion* (BIC) [14] or the *Akaike's information criterion* (AIC) [3] is selected as the optimal one.

3 Simulation study

To evaluate the ability of the proposal in terms of correctly identifying the latent model structure (estimating the β_g parameters) and classifying the sending nodes, we conducted a large scale simulation study as described below.

3.1 Simulation setup

Six different scenarios are considered; these are based on a variable number of groups ($G = 3$, $G = 4$) and sending nodes ($N = 200$, $N = 500$, $N = 1000$), while the number of receiving nodes R is kept constant and equal to 14. Furthermore, a univariate continuous latent trait variable is considered ($D = 1$). Last, as regards the latent class variable, block membership is defined via a single nodal attribute x_i which is drawn from a Gaussian distribution with mean and variance equal to 1, so that class membership is defined by the following model specification:

$$\mathrm{logit}(\eta_{ig}) = \beta_{0g} + \beta_{1g}x_i, \quad g = 2,\ldots,G.$$

In each scenario, the number of random starting values for the model parameters is set to 10. Simulation results are based on 500 samples.

3.2 Simulation results: latent class parameters

We report in Figures 1 and 2 the distributions of the parameters β_g across samples, for different values of N and G. By looking at the figures, it is evident that the proposal works well when the number of latent classes is small ($G = 3$), besides the size of the network. On the other hand, when the number of groups increases ($G = 4$), in order to obtain good performances in terms of parameter recovery, a

larger amount of information is needed. Simulation results show that as the size of the network increases (the number of sending nodes increases), we are more and more able to identify the true values of model parameters.

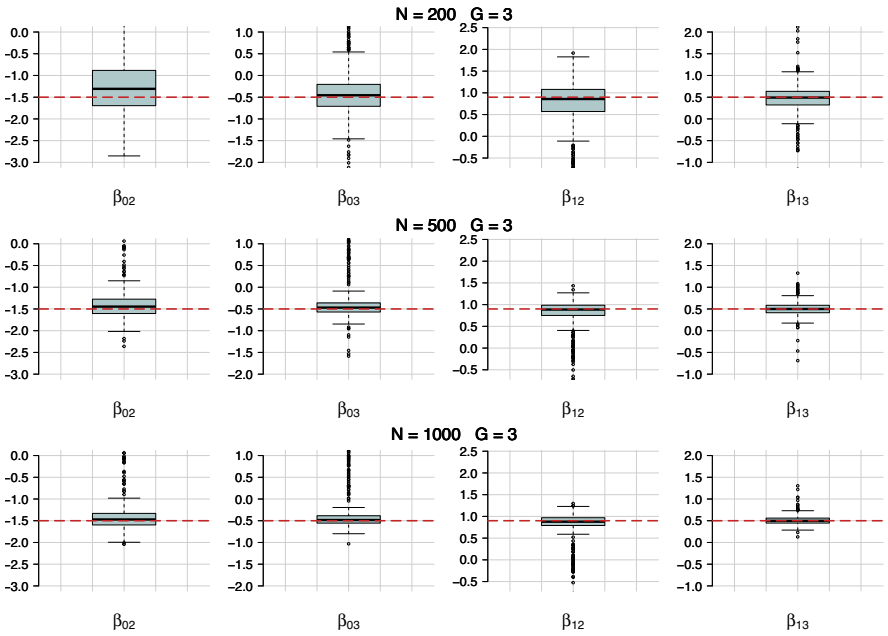


Fig. 1 Distribution across samples of the parameter estimates $\hat{\beta}_g$ for varying N and $G = 3$. The red lines correspond to the true values of the parameters.

3.3 Adjusted Rand Index

The ability of the proposal in correctly classifying the sending nodes is evaluated via the Adjusted Rand Index (ARI) [10]. Results are shown in Table 1. When looking at these results, we notice that when the number of groups increases, the ARI worsens. However, as expected, the classification improves if the number of sending nodes increases. Specifically, when $G = 4$, the average ARI for $N = 200$ and $N = 500$ is 0.6865 and 0.7561, respectively. For $N = 1000$, the classification improves and the average ARI reaches the value of 0.8162, thus suggesting a good performance of the proposed method.

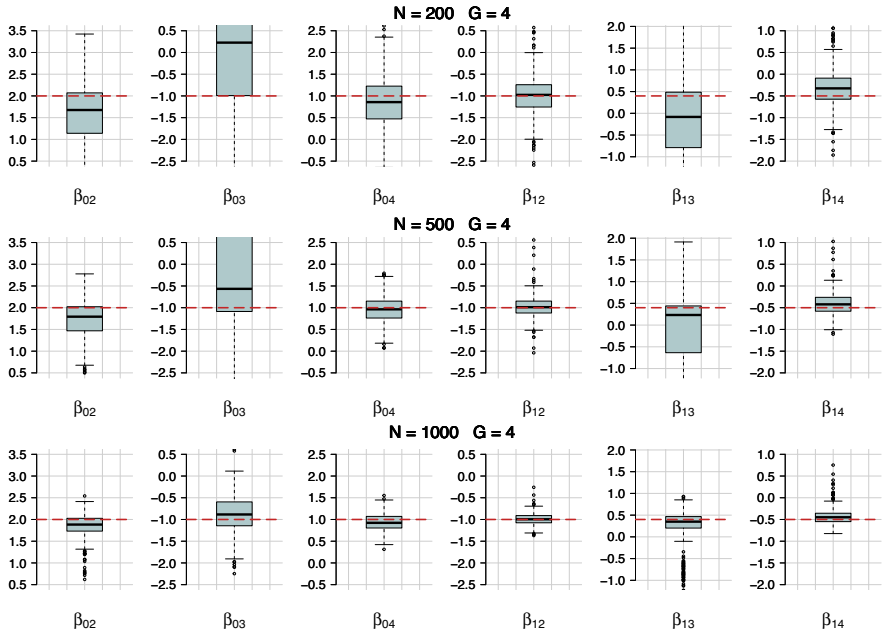


Fig. 2 Distribution across samples of the parameter estimates $\hat{\beta}_g$ for varying N and $G = 4$. The red lines correspond to the true values of the parameters.

Table 1 Distribution across samples of the Adjusted Rand Index for varying N and G.

		Adjusted Rand Index					
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
N=200	G=3	0.1534	0.6303	0.8023	0.7304	0.8681	0.9569
	G=4	0.1173	0.5517	0.7077	0.6865	0.8145	0.9293
N=500	G=3	0.2391	0.8382	0.8663	0.8222	0.8899	0.9578
	G=4	0.2092	0.6001	0.8281	0.7561	0.8579	0.9249
N=1000	G=3	0.3751	0.8568	0.8733	0.8305	0.8876	0.9264
	G=4	0.5059	0.8303	0.8502	0.8162	0.8667	0.9007

4 Conclusions

The paper proposes an extension of the Mixture of Latent Trait Analayzers for clustering bipartite networks. In particular, we extend the model to account for nodal attributes on the definition of the latent model structure. The aim is that of identifying how nodal features affect the clustering. In detail, the sending nodes' attributes are exploited to model the prior probability that a sending node belongs to a specific group by considering a multinomial logit specification. A simulation study is con-

ducted to assess the performance of the proposal in terms of parameter recovery and clustering. The results of the simulation study show that, if the number of sending nodes and the number of groups are small, both the model's estimation and the classification are good. As the number of groups increases, a higher number of sending nodes ensures good performance of the proposed method.
A further development consists in the application of the proposal for the analysis of the bipartite network entailing the relation between COVID-19 patients (the sending nodes) and the behaviors they adopted to prevent infection (receiving nodes). The aim is that of identifying groups of patients with similar behaviors in terms of preventive measures, also taking into account individual characteristics. In addition, two further lines of research may entail the extension of the model to the case of response variables with more than two categories, as well as the analysis of longitudinal bipartite networks.

References

1. Agresti, A.: Categorical Data Analysis. John Wiley & Sons, Hoboken (2002)
2. Aitkin, M., Vu, D., Francis, B.: Statistical modelling of the group structure of social networks. Soc. Netw. **38**, 74–87 (2014)
3. Akaike, H.: A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. **19**, 716–23 (1974)
4. Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., Rathouz, P.J.: Latent Variable Regression for Multiple Discrete Outcomes. J. Am. Stat. Assoc. **92**, 1375–1386 (1997)
5. Bartholomew, D. J., Knott, M., Moustaki, I.: Latent Variable Models and Factor Analysis: A Unified Approach. 3rd ed. Wiley, Hoboken (2011)
6. Dayton, C. M., Macready, G. B.: Concomitant-Variable Latent-Class Models. J. Am. Stat. Assoc. **83**, 173–178 (1988)
7. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. **7**, 1–26 (1979)
8. Gollini, I.: A mixture model approach for clustering bipartite networks. In: Ragozini, G., Vitale, M.P. (eds.) Challenges in Social Network Research: Methods and Applications, pp. 79–91. Springer International Publishing (2020)
9. Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**, 569–588 (2014)
10. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
11. Jaakkola, T. S., Jordan, M. I.: Bayesian logistic regression: A variational approach. In: Madigan, D., Smyth, P. (eds.) Proceedings of the 1997 Conference on Artificial Intelligence and Statistics. Ft. Lauderdale, FL (1997)
12. Jones, S. P.: Imperial College London Big Data Analytical Unit and You-Gov Plc., Imperial College London YouGov Covid Data Hub, v1.0, YouGov Plc. (2020)
13. Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. J. Am. Stat. Assoc. **96**, 1077–1087 (2001)
14. Schwarz, G.: Estimating the Dimension of a Model. Ann. Stat. **6**, 461–464 (1978).
15. Tipping, M. E.: Probabilistic Visualisation of High-Dimensional Binary Data. In: Kearns, M., Solla, S., Cohn, D. (eds.) Advances in Neural Information Processing Systems 11, pp. 592–598. MIT Press (1998)
16. Tukey, J.W.: Bias and Confidence in Not-Quite Large Sample. Ann. Math. Stat. **29**, 614–623 (1958)

A Fast Majorization-Minimization Algorithm for Convex Clustering

Un Veloce Algoritmo di Maggiorazione-Minimizzazione per Clustering Convesso

D.J.W. Touw, P.J.F. Groenen, Y. Terada

Abstract Convex clustering is introduced as a clustering method which combines aspects of k-means and hierarchical clustering. Existing algorithms to minimize the loss function that corresponds to this model struggle with analyzing data sets that are larger than several thousands of objects. We propose an algorithm that leverages sparsity and cluster fusions in combination with a majorization-minimization algorithm that is at least 100 times faster than current state-of-the art implementations.
Abstract *Il clustering convesso e' un metodo di clustering che combina aspetti propri del k-means e del clustering gerarchico. Gli algoritmi esistenti incorrono in problemi quando devono minimizzare la funzione obiettivo di questo modello in presenza di grande quantita' di dati. Proponiamo un algoritmo che bilancia la sparsita' e le fusioni tra clusters tramite un algoritmo di maggiorazione-minimizzazione che e' almeno 100 volte piu' veloce delle attuali implementazioni.*

Key words: convex clustering, grouped-Lasso, majorization-minimization, unsupervised learning

D.J.W. Touw
Department of Econometrics, Erasmus University Rotterdam, Rotterdam, The Netherlands, e-mail: touw@ese.eur.nl

P.J.F. Groenen
Department of Econometrics, Erasmus University Rotterdam, Rotterdam, The Netherlands, e-mail: groenen@ese.eur.nl

Y. Terada
Graduate School of Engineering Science, Osaka University, Osaka, Japan, e-mail: terada@sigmath.es.osaka-u.ac.jp

1 Introduction

Clustering is a type of unsupervised machine learning which is used to extract useful information from large bodies of data. In recent publications, a new method called convex clustering has been developed which combines aspects of the two popular techniques k-means [9] and hierarchical clustering [3].

In the convex clustering framework [5, 8, 10], each p-dimensional object $\mathbf{x}_i$ in the data is represented by a p-dimensional vector $\mathbf{a}_i$. In the loss function, two forces act on $\mathbf{a}_i$. One penalizes its squared Euclidean distance to $\mathbf{x}_i$ and the other its Euclidean distances to the other $\mathbf{a}_j$, with $j \in \{1,\ldots,n\} \setminus \{i\}$, which is regulated by a parameter λ. The second force (the Euclidean distance between the $\mathbf{a}_i$) acts as a grouped-Lasso: when minimizing the loss for $\lambda > 0$, some $\mathbf{a}_i$ may become identical. In case that $\mathbf{a}_i = \mathbf{a}_j$, we say that objects i and j are clustered together. In contrast to previous publications, we scale the loss function in order to ensure that the minimizer is scale invariant. We define the convex clustering loss function as

$$L(\mathbf{A}) = \frac{\|\mathbf{A}-\mathbf{X}\|_2^2}{2\|\mathbf{X}\|_2^2} + \lambda \frac{\sum_{i<j} w_{ij}\|\mathbf{a}_i - \mathbf{a}_j\|_2}{\|\mathbf{X}\|_2 \sum_{i<j} w_{ij}}, \tag{1}$$

where the $n \times p$ matrix $\mathbf{A}$ has rows $\mathbf{a}_i$, the $n \times p$ matrix $\mathbf{X}$ has rows $\mathbf{x}_i$, and w_{ij} is a user-defined weight that reflects the importance of clustering objects i and j. In general, a larger value for λ corresponds to a smaller number of clusters. Minimizing the convex clustering loss function for a sequence of values for λ yields a sequence of solutions for $\mathbf{A}$ which is referred to as the *clusterpath* (see Fig. 1).

In our research, we develop a new algorithm called *convex clustering through majorization-minimization* (CMM) to minimize (1). This algorithm is more efficient than current state-of-the art implementations like the *alternating minimization algorithm* (AMA) [2] and the *semismooth Newton-GC augmented Lagrangian method* (SSNAL) [11]. To achieve this, we make use of sparsity in the user-defined weights, fusions of the rows in $\mathbf{A}$ that are identical, and a majorization-minimization (MM) algorithm.

2 Methodology

In this section, we discuss the three aspects of the CMM algorithm. First, we explain how sparsity is imposed on the user-defined weights in the model. Second, cluster fusions are discussed. Third, we show how these aspects are combined in an efficient MM algorithm.

2.1 Sparsity

In existing literature, it has been shown that the user-defined weights in (1) have a large effect on the ability to correctly identify clusters [2, 5]. The data can be used to introduce sparsity in the weights by setting w_{ij} to zero if objects i and j are not among the k nearest neighbors of each other. Hence, the elements of the weight matrix $\mathbf{W}$ are computed as

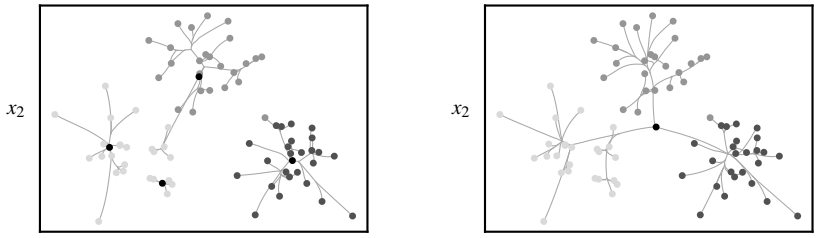
$$w_{ij} = w_{ji} = \begin{cases} \exp\left(-\phi \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\text{mean}_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2}\right) & \text{if } (i,j) \in \mathscr{S} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathscr{S} = \{(i,j) : i \in \mathscr{N}_j^k \vee j \in \mathscr{N}_i^k\}$ and $\mathscr{N}_i^k$ is the set of the k nearest neighbors of object i. However, even though this approach has shown to perform better than a dense weight matrix, it comes with a drawback. If k is small with respect to n, there may be groups of objects that are not connected via nonzero weights. In that case, the clusterpath does not terminate in a single cluster. An example of such a situation is provided in Fig. 1a, where the nonzero weights are determined by the $k = 3$ nearest neighbors. The data in this figure are generated from three distinct clusters, but the minimum number of clusters that the clusterpath can attain is four. In order to guarantee a fully connected weight structure, the value for k must be chosen as a large fraction of n, affecting the scalability of any algorithm used to minimize the loss.

To avoid incurring a large and unnecessary computational burden, we propose to add nonzero weights by using the structure of a symmetric circulant matrix $\mathbf{S}$ [4]. In $\mathbf{S}$, $s_{ij} = s_{ji}$ and each row is the same, but shifted one position to the right with respect to the previous row. An example of the most sparse symmetric circulant that is not the identity matrix for $n = 6$ is

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

If we interpret $\mathbf{S}$ as an adjacency matrix, it corresponds to a connected graph. Hence, we redefine the set of indices of nonzero weights in (2) as $\mathscr{S} = \{(i,j) : i \in \mathscr{N}_j^k \vee j \in \mathscr{N}_i^k \vee s_{ij} = 1\}$. For the clusterpath in Fig. 1b, again $k = 3$ is used, but additional weights are added according to $\mathbf{S}$. As a result, the clusterpath now has a complete cluster hierarchy which terminates at one cluster.



(a) Disconnected clusterpath (b) Connected clusterpath

Fig. 1 In Panel (a): A clusterpath that results from repeatedly minimizing the convex clustering loss function for increasing values for λ, where the final clusters are indicated by black dots. Weights are nonzero for pairs of objects that are among the $k = 3$ nearest neighbors of each other. In Panel (b): A clusterpath obtained in a procedure identical to the clusterpath in Panel (a), except for the construction of the weights. In addition to using the $k = 3$ nearest neighbors for nonzero weights, the nonzero entries in a symmetric circulant matrix are used to ensure that there are no groups that are disconnected from the rest.

2.2 Cluster Fusions

During the computation of the clusterpath, the number of clusters decreases for increasing values for λ. This means that for some $\lambda_1 > 0$ there are i and j for which $\mathbf{a}_i = \mathbf{a}_j$. Theoretically, there is no guarantee that this remains the case for $\lambda_2 > \lambda_1$. In fringe cases, it has been shown that clusters can split again for larger values for λ, but it has been conjectured that weights decreasing in $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ do not cause cluster splits [2, 5, 12]. However, to guarantee cluster hierarchy, we take the approach suggested by [5] and combine the rows of $\mathbf{A}$ that are identical. This results in a $c \times p$ matrix $\mathbf{M}$, where c is the number of clusters, that holds the unique rows of $\mathbf{A}$. Furthermore, we define the $n \times c$ cluster membership matrix $\mathbf{U}$ as

$$u_{ik} = \begin{cases} 1 & \text{if observation } i \text{ belongs to cluster } k \\ 0 & \text{otherwise,} \end{cases}$$

such that $\mathbf{A} = \mathbf{UM}$. Substituting $\mathbf{UM}$ for $\mathbf{A}$ in (1), we obtain the following loss function

$$L(\mathbf{M}) = \frac{\|\mathbf{UM} - \mathbf{X}\|_2^2}{2\|\mathbf{X}\|_2^2} + \lambda \frac{\sum_{k<l} (\mathbf{U}^\top \mathbf{WU})_{kl} \|\mathbf{m}_k - \mathbf{m}_l\|_2}{\|\mathbf{X}\|_2 \sum_{i<j} w_{ij}}, \tag{3}$$

which can be minimized over $\mathbf{M}$. In addition to enforcing cluster hierarchy, the loss is minimized over a $c \times p$ matrix instead of an $n \times p$ matrix, which allows for a significant reduction in the computational burden when $c \ll n$.

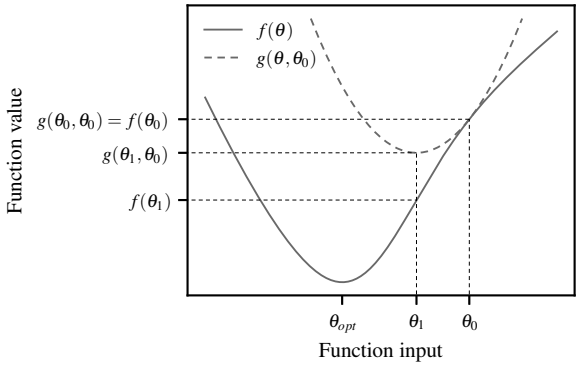


Fig. 2 Example of minimization by an MM algorithm. The majorization function $g(\theta,\theta_0)$ is greater than the target function $f(\theta)$ on the entirety of its domain, except for the supporting point θ_0 where they are equal. Minimizing $g(\theta,\theta_0)$ yields the new supporting point θ_1. Repeating this procedure until convergence results in a solution close to θ_{opt}.

2.3 Majorization-Minimization

In order to minimize (3), we make use of MM. In such an algorithm, the complicated target function is replaced by a simpler function of which the minimum can be computed analytically. If $f(\theta)$ is the target function, the majorization function $g(\theta,\theta_0)$ should satisfy

$$g(\theta,\theta_0) \geq f(\theta) \quad \text{and} \quad g(\theta_0,\theta_0) = f(\theta_0),$$

where θ_0 is called the supporting point, as it is the point where $g(\theta,\theta_0)$ "rests" on $f(\theta)$. If θ_1 is the minimizer of $g(\theta,\theta_0)$, the aforementioned criteria ensure that $f(\theta_1) \leq f(\theta_0)$ as

$$f(\theta_1) \leq g(\theta_1,\theta_0) \leq g(\theta_0,\theta_0) = f(\theta_0),$$

which is also illustrated by Fig. 2. Setting θ_0 to θ_1 and repeating these steps causes the supporting point to converge towards a local minimum. If the target function is convex and coercive, this is the global minimum [1]. For further reading on MM, which is also known as the *concave-convex procedure*, we refer to [6, 7, 13].

In CMM, the majorization function $g(\mathbf{M},\mathbf{M}_0)$ of $L(\mathbf{M})$ has a very important property. The time complexity of finding the minimum is $\mathcal{O}(cpk)$, where c and p are the number of rows and columns of $\mathbf{M}$, respectively, and k is the number of neighbors used to construct the weight matrix $\mathbf{W}$ in (2). The combination of this majorization function with sparsity in the weights and cluster fusions results in an efficient algorithm to perform convex clustering.

3 Numerical Experiments

In our experiments, we compare CMM with two other algorithms. To our knowledge, the SSNAL algorithm is currently the fastest way to perform convex cluster-

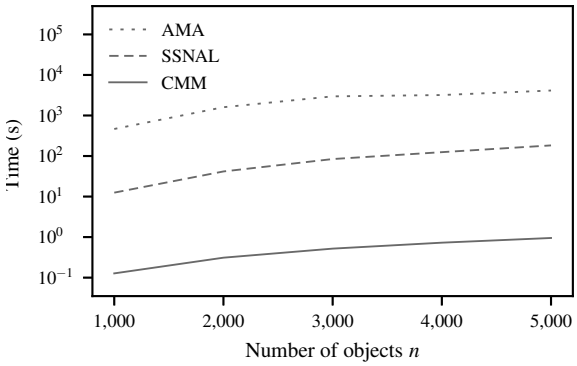


Fig. 3 The average time required to compute a clusterpath for $\lambda \in \{0.0, 0.2, \ldots, 110.0\}$ for three algorithms. The data sets are generated according the two interlocking half moon clusters in $\mathbb{R}^2$. For 1,000 objects, CMM is roughly 100 times faster than SSNAL and 3,650 times faster than AMA, this increases to 190 and 4,350 times faster for 5,000 objects.

ing. In [11], the authors compare it to AMA and find that in some cases, SSNAL is over 80 times faster.

To compare the three algorithms, we used the two interlocking half moons data generating process to generate several data sets. These range from 1,000 to 5,000 objects in $\mathbb{R}^2$, with ten realizations of the data for each value for n. To compute the weight matrix, we set $\phi = 2$ and $k = 15$. In these experiments, we did not use the symmetric circulant matrix to ensure a clusterpath that ends in a single cluster, as AMA and SSNAL do not include this option in their implementations. For the clusterpath, we used $\lambda \in \{0.0, 0.2, 110.0\}$. In Fig. 3, we report the average elapsed time per clusterpath for each of the algorithms. These results show that CMM is not only faster than the other algorithms, it also scales better for larger data sets. If we look at $n = 1{,}000$, CMM is roughly 100 times faster than SSNAL and 3,650 times faster than AMA. For $n = 5{,}000$, the speedup of CMM over the other two algorithms increases to 4,350 and 190, respectively. Furthermore, even though SSNAL attained, on average, the lowest value for the loss function, the result obtained by CMM deviated at most 0.01%.

4 Conclusion

In our research we developed a new algorithm to perform convex clustering. In CMM, we use a combination of cluster fusions and a symmetric circulant matrix to guarantee a complete cluster hierarchy. Furthermore, our minimization algorithm is more efficient than state-of-the-art alternatives like AMA and SSNAL. The introduction of CMM allows convex clustering to be applied to larger data sets than before, allowing for more research on the clustering method itself.

References

1. Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
2. Chi, E. C., Lange, K.: Splitting Methods for Convex Clustering. J. Comput. Graph. Stat. **24(4)**, 994–1013 (2005)
3. Gan, G., Ma, C. Wu, J.: Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia, PA (2007)
4. Gower, J.C., Groenen, P.J.F.: Applications of the Modified Leverrier-Faddeev Algorithm for the Construction of Explicit Matrix Spectral Decompositions and Inverses. Util. Math. **40**, 51–64 (1991)
5. Hocking, T.D., Joulin, A., Bach, F, Vert, J.-P.: Clusterpath: An Algorithm for Clustering Using Convex Fusion Penalties. In: 28th international conference on machine learning, Bellevue, WA (2011)
6. Hunter, D.R., Lange, K.: A Tutorial on MM Algorithms. Am. Stat. **5(1)**, 30–37 (2004)
7. Lange, K., Hunter, D.R., Yang, I.: Optimization Transfer Using Surrogate Objective Functions. J. Comput. Graph. Stat. **9(1)**, 1–20 (2000)
8. Lindsten, F., Ohlsson, H., Ljung, L.: Just Relax and Come Clustering!: A Convexification of K-Means Clustering. Technical report, Department of Electrical Engineering, Linköping University, Linköping (2011)
9. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Le Cam, L.M., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley, CA (1967)
10. Pelckmans, K, De Brabanter, J, Suykens, J.A.K., De Moor, B.: Convex Clustering Shrinkage. In: PASCAL Workshop on Statistics and Optimization of Clustering Workshop, London (2005)
11. Sun, D., Toh, K.-C., Yuan, Y.: Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm. J. Mach. Learn. Res. **25(3)**, 243–251 (2021)
12. Weylandt, M., Nagorski, J., Allen, G.I.: Dynamic Visualization and Fast Computation for Convex Clustering via Algorithmic Regularization. J. Comput. Graph. Stat. **29(1)**, 87–96 (2020)
13. Yuille, A.L., Rangarajan, A.: The Concave-Convex Procedure. Neural Comput. **15(4)**, 915–936 (2003)

Statistical Methods for Complex Evolutionary Data

A FANOVA model with repeated measures for detecting patterns in biomechanical data

Un modello FANOVA con misure ripetute per rilevare modelli nei dati biomeccanici

Ana M. Aguilera and Christian Acal and Manuel Escabias

Abstract Biomechanics data are usually curves that represent the human movement when subjects are submitted to multiple conditions. The main objective of this paper is to detect possible differences in gait patterns when a group of children between 8 and 11 years old go to school with different book-bags (walking without bag, carrying a backpack and pulling a trolley). Given the stochastic nature of available data, a functional data analysis is carried out. In particular, we conduct a novel approach for functional analysis of variance (FANOVA) with repeated measures. This methodology, which has never been used in the field of biomechanics, is based on a basis expansion of sample curves. The obtained results reveal significant differences gender and in the kind of book-bags.

Abstract *I dati biomeccanici sono solitamente curve che rappresentano il movimento umano quando i soggetti sono sottoposti a condizioni multiple. "L'obiettivo principale di questo lavoro consiste nel rilevare possibili differenze nei modelli di andatura per un gruppo di bambini tra gli 8 e gli 11 anni che si reca a scuola portando i libri in diversi modi (camminando senza borsa, portando uno zaino e tirando un trolley). Data la natura stocastica dei dati disponibili, viene eseguita un'analisi dei dati funzionali, considerando un nuovo approccio per l'analisi funzionale della varianza con misure ripetute. Questa nuova metodologia, mai utilizzata in campo biomeccanico, si basa sull'espansione in basi delle curve campionarie. I risultati rivelano differenze significative nel genere e nella tipologia di trasporto dei libri.*

A.M. Aguilera
Department of Statistics and O.R. and Math Institute (IMAG), University of Granada, Spain.
e-mail: aaguiler@ugr.es
C. Acal
Department of Statistics and O.R. and Math Institute (IMAG), University of Granada, Spain.
e-mail: chracal@ugr.es
M. Escabias
Department of Statistics and O.R. and Math Institute (IMAG), University of Granada, Spain.
e-mail: escabias@ugr.es

Key words: Functional Data Analysis, Analysis of Variance, Repeated Measures, Biomechanics

1 Introduction

Biomechanical data usually reveal the linear acceleration or the position of some joint (angle formed with some axis) in terms of some continuous argument such as time or percentage of gait cycle, among others. Historically, the analysis of movement curves has been carried out by means of multivariate methods from discrete observations of the curves or even by summary measures of them. However, given the stochastic nature of biomechanical data, it makes more sense to apply a functional approach that leverages all information of curves and therefore, enable to draw more accurate conclusions. In this sense, the first step would be to reconstruct the functional form of curves. For this purpose, two methodologies are available in the literature: the first option consists of using the non-parametric techniques proposed by [6]. The second one is based on the projection of sample curves in a finite-dimension space generated by a basis [8, 9]. The latter approach is assumed in the current manuscript. Since biomechanical curves tend to be smooth, B-splines basis is a suitable candidate in practice. A comparison of different types of penalized smoothing with B-splines basis was performed in [3].

One of the main objectives in the field of biomechanics, is to detect possible differences in gait patterns when the subjects perform multiple activities (repeated measures design). Here, we focus on evaluating how the rotation of joints is affected when children carry different book-bags to the school. This problem can be worked out through a functional analysis of variance with repeated measures (FANOVA-RM). The first testing approach for this theoretical framework was proposed by [7]. However, this procedure does not take the within group variability into account but only the between group variability. In order to solve this drawback, [10] introduced two new statistics which were generalized by considering the basis expansion of sample curves in [2]. For the case of two-way FANOVA problem, where one factor represents the repeated measures effect and the second one denotes the independent group contribution, a novel approach has been recently introduced in [1]. This methodology is considered in the current work to detect if the type of book-bags used by the children influences in the gait pattern (repeated measures), as well as to analyse if there are differences depending on gender (independent groups).

In addition to this introduction, the rest of the paper is organized as follows: Sect. 2 briefly describes the theoretical framework of FANOVA-RM. The application with the biomechanical data can be seen in Sect. 3. Finally, some concluding remarks are made in Sect. 4.

2 Functional analysis of variance with repeated measures

In a repeated measures design in which there are g independent samples of curves (one per group), the response functional variable X is repeatedly measured on each subject at m different time periods or treatment conditions. Hence, let $\{x_{ijk}(t) : i = 1,2,\ldots,m; j = 1,2,\ldots,g; k = 1,2,\ldots,n_j; t \in T\}$ denote g independent samples of curves defined on a continuous interval T. That is, $x_{ijk}(t)$ is the response of the kth subject in the jth group under the ith treatment. It is assumed that the design is balanced (each treatment is applied to all subjects) and sample curves belong to the Hilbert space $L^2[T]$ of squared integrable functions with the usual inner product $< f|g >= \int_T f(t)g(t)dt,\ \forall f,g \in L^2[T]$.

In FANOVA-RM model, sample curves verify the following functional linear model

$$x_{ijk}(t) = \mu(t) + \alpha_i(t) + \beta_j(t) + \theta_{ij}(t) + \varepsilon_{ijk}(t),\ \forall t \in T,$$

with $\mu(t)$ being the overall mean function; $\alpha_i(t)$ and $\beta_j(t)$ the ith and jth main-effect functions of treatments and groups, respectively; $\theta_{ij}(t)$ is the (i,j)th interaction-effect between treatments and groups; and $\varepsilon_{ijk}(t)$ are i.i.d. errors with distribution $SP(0,\gamma(s,t))\ \forall i = 1,2,\ldots,m; j = 1,2,\ldots,g; k = 1,2,\ldots,n_j$.

Unlike the framework with independent measures where the only objective is to study the between-subject variability, in the repeated measures layout it is essential to analyse the intra-subject variability as well. In order to take this issue into account, we follow the approach proposed by [1]. If the basis expansion of sample curves is considered, i.e., sample curves belong to a finite-dimension space generated by a basis $\{\phi_1(t),\ldots,\phi_p(t)\}$, so that they can be expressed as

$$x_{ijk}(t) = \sum_{h=1}^{p} y_{ijkh}\phi_h(t),$$

where p must be sufficiently large to get an accurate representation of curves, the FANOVA-RM problem is reduced to perform a multivariate analysis of variance with repeated measures on the multivariate response defined by the basis coefficients of sample curves, y_{ijkh}.

In this point, Doubly Multivariate Model (DMM) or Mixed Multivariate Model (MMM) can be applied to solve the problem [4, 5, 11]. The multivariate normality hypothesis and homogeneity of covariance matrices must be verified in both approaches. Besides, MMM requires the multivariate sphericity condition (barely fulfilled in practice), whereas DMM does not impose an assumption as restrictive as the mixed model; the covariance matrix must be only positive definite. Nevertheless, MMM is more powerful than DMM so it should be the first option when sphericity condition is satisfied. For those situations where sphericity is not verified, [4] proposed disrupted the degrees of freedom of F-statistics. Likewise, given that normality and homogeneity hypotheses are also rarely verified in functional data analysis, a permutation testing procedure was adapted in [1].

Through these procedures, we can test the main hypothesis tests associated with FANOVA-RM model:

- Are there significant differences among treatments?

$$H_0 : \alpha_1(t) = \alpha_2(t) = \ldots = \alpha_m(t) = 0,\ \forall t \in T; \tag{1}$$

- Are there significant differences among independent groups?

$$H_0 : \beta_1(t) = \beta_2(t) = \ldots = \beta_g(t) = 0,\ \forall t \in T; \tag{2}$$

- Are there significant interaction-effects between groups and treatments?

$$H_0 : \theta_{ij}(t) = 0,\ \forall i, j;\ \forall t \in T; \tag{3}$$

against the alternative, in each case, that its negation holds.

3 Application in Biomechanics

The application is focused on a wide experimental study carried out in the biomechanics laboratories of the Sport and Health University Research Institute (iMUDS) of the University of Granada (Spain). This study aims to detect possible differences in gait patterns when children (25 boys and 28 girls between 8 and 11 years old) go to school with different book-bags and weights. Recorded biomechanical data consists of curves of the gait cycle measured in 101 equidistant points, over a specific platform under three conditions (walking, carrying a backpack and pulling a trolley) on the 3-axial angular rotation in multiple joints. In the current application, we only use part of this study. In particular, we consider the thorax angular rotation (radians) measured on axis Z for the conditions of walking, carrying a backpack that weighs 10% of the subject's weight and pulling a trolley that weighs 10% of the subject's weight. A child was removed for having an anomalous behaviour in comparison with the rest (outlier). For the functional reconstruction of sample curves, a cubic B-spline basis of dimension 20 was considered. The sample mean functions for each condition depending on gender can be seen in Fig. 1.

In order to check the hypothesis tests defined in (1), (2) and (3), FANOVA-RM methodology based on basis expansion of sample curves is applied. Given the conditions of this study ($g = 2$, $m = 3$ and $p = 20$) only the MMM can be conducted. Besides, since normality is not verified for this dataset, we perform the MMM model through the permutation testing procedure developed in [1].

On the basis of the results obtained during the analysis, we can conclude that the kind of book-bag plays a fundamental role in the thorax angular rotation on axis Z (p-value associated with the Pillai's trace statistic is 0.002), as well as there are also significant differences according to gender (p-value is equals to 0.012). Finally, we do not find interaction-effect between gender and conditions (p-value is 0.641).

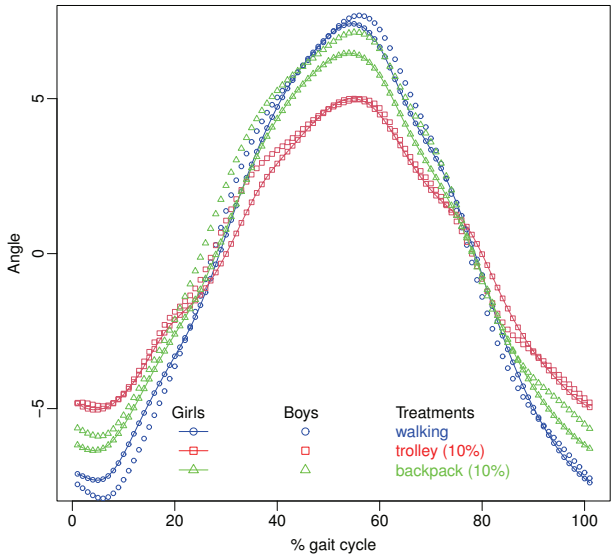


Fig. 1 Sample group mean functions depending on gender and condition.

4 Conclusions

In this work, gait patterns when children go to the school with different book-bags have been evaluated by means of a functional analysis of variance with repeated measures based on the basis expansion of sample curves. The results have proven that the kind of book-bags have a significant influence, at least, in the thorax angular rotation on axis Z. While it is true that this research study is found in an initial phase yet, these first results are really important and should set out many questions for parents, Governments and teachers about the preventive measures to be adopted that guarantee the health of children.

Acknowledgements. This work is partially supported by the project FQM-307 of the Government of Andalusia (Spain) and the project PID2020-113961GB-I00 of the Spanish Ministry of Science and Innovation (also supported by the FEDER programme). Authors also acknowledge the financial support of the Consejería de Conocimiento, Investigación y Universidad, Junta de Andalucía (Spain) and the FEDER programme for project A-FQM-66-UGR20. Additionally, authors would like to express the financial support by the IMAG–María de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033.

References

1. Acal, C., Aguilera, A. M.: Basis expansion approaches for Functional Analysis of Variance with repeated measures. Adv. Data Anal. Classi. in press (2022)
2. Acal, C., Aguilera, A.M., Sarra, A., Evangelista, A., Di-Battista, T., Palermi, S.: Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. Stoch. Environ. Res. Risk Assess. **36**, 1083–1101 (2022)
3. Aguilera, A.M., Aguilera-Morillo, M.C.: Penalized PCA approaches for B-spline expansions of smooth functional data, Appl. Math. Comput. **219***(14)*, 7805–7819 (2013)
4. Boik, R. J.: The mixed model for multivariate repeated measures: Validity conditions and an approximate test. Psychometrika **53***(4)*, 469–486 (1988)
5. Boik, R. J.: Scheffes mixed model for multivariate repeated measures: a relative efficiency evaluation. Commun. Stat. Theor. M. **20***(4)*, 1233–1255 (1991)
6. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and practice. Springer-Verlag (2006)
7. Martinez-Camblor, P., Corral, N.: Repeated measures analysis for functional data. Comput. Stat. Data. Anal. **55***(12)*, 3244–3256 (2011)
8. Ramsay, J. O., Silverman, B. W.: Applied functional data analysis: Methods and case studies. Springer-Verlag (2002)
9. Ramsay, J. O., Silverman, B. W.: Functional data analysis (Second Edition). Springer-Verlag, (2005)
10. Smaga, L.: A note on repeated measures analysis for functional data. AStA Adv. Stat. Anal. **104***(1)*, 117–139 (2020)
11. Timm, N. H.: Multivariate analysis of variance of repeated measurements. In: Krishnaiah, P.R. Analysis of Variance, vol. 1 of Handbook of Statistics, pp. 41–87, Elsevier (1980)

Modes of variation for Lorenz Curves

Modi di variazione per curve di Lorenz

Enea G. Bongiorno and Aldo Goia

Abstract This work illustrates how to perform functional principal component analysis and to compute the modes of variations for a sample of Lorenz curves. In particular, to coherently implement functional principal component analysis in a proper manner, Lorenz curves are suitably transformed. The procedure is applied at the income Lorenz curves for the Italian regions in the years 2000, 2006 and 2010.
Abstract *Questo lavoro illustra come implementare l'analisi delle componenti principali funzionali e come calcolare i modi di variazione per un campione di curve di Lorenz. In particolare, al fine di implementare in maniera coerente l'analisi delle componenti principali funzionali, le curve di Lorenz sono trasformate opportunamente. La procedura è applicata alle curve di Lorenz del reddito per le regioni italiane negli anni 2000, 2006 e 2010.*

Key words: Lorenz curves, Modes of variation, income distributions

1 Introduction

In some applications, ranging from Economics to Biology, from Chemistry to Environmetrics, it is interesting to consider the notion of concentration, that is the attitude of a non–negative r.v. X to redistribute its total mass over the individuals within the population. This concept allows to represent and distinguish situations ranging from the maximum concentration setting (when one individual holds the total mass) to the equidistribution one (when each individual hold the same mass).

Enea G. Bongiorno
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: enea.bongiorno@uniupo.it

Aldo Goia
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: aldo.goia@uniupo.it

A formal way to depict the concentration of a probability law is given by the Lorenz Curve (LC) [5] that is defined by

$$\begin{aligned} L:\,[0,1] &\to [0,1] \\ p &\mapsto L(p)=\textstyle\int_0^p Q(t)dt/\mu, \end{aligned}$$

where $\mu=\mathbb{E}[X]$, $Q(p)=\inf\{x:F(x)\geq p\}$ is the quantile function of X defined for any $p\in(0,1)$ and with F being the cdf of X. For a LC one has $L(0)=0$, $L(1)=1$, $L(p)\leq p$ and L is continuous, increasing and convex on $[0,1]$. As an instance, consider the empirical LCs (i.e. based on the empirical versions of mean and quantile function) of household income of the 20 regions of Italy for the years 2000, 2006, 2010 estimated from the Bank of Italy Survey on Household Income and Wealth, see Fig. 1. Since $L(p)$ is the percentage of the income X held by the $p100\%$ "poorest" part of the population, each curve represents how the income concentrates within a region population in a given year.

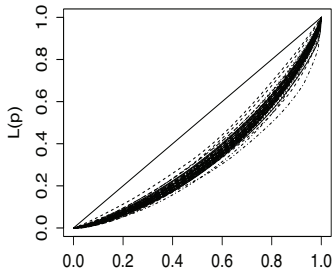


Fig. 1 Each curve illustrates the concentration of family income in a given year (2000, 2006, 2010) and region for a total of 60 empirical LCs.

These curves can be seen as a sample of a random element taking values in $\mathscr{L}or$, the family of continuous, increasing and convex functions from $[0,1]$ to itself passing through the origin and $(1,1)$. In this view, one can explore data by borrowing techniques from functional data analysis (FDA): a recent branch of statistics that studies those phenomena whose observations are (discretized) curves; see e.g. [3, 4, 6]. Altough a standard FDA approach for LCs is possible, in general, it is not advisable. In fact, LCs are special functional data not directly observed but estimated from a sample of a real random variable: this leads to a double stochasticity issue that could impact over usual FDA techniques. Moreover, given the constrained nature of the Lorenz curve process, $\mathscr{L}or$ is not a structured space (for instance Hilbert) and then classical methods should be used with caution.

The aim of this work is to explore the variability of the described data by means of the "modes of variation". For a given functional process, its j-th mode of variation is the mean function perturbed by $\pm k\sqrt{\eta_j}v_j$ where, $k>0$ and $\{\eta_j,v_j\}$ are the j-th eigenelements of the covariance operator of the process. As a consequence, modes of variation are usually computed after the functional principal component analysis

(FPCA), but, given the above remarks on LCs, a naive application of FPCA leads to modes of variation not belonging to $\mathscr{L}or$ and then to incoherent interpretations. To tackle such issue, a preliminary transformation of data is necessary.
The remain part of this work is divided in two sections: Sect. 2 describes the embedding proposed by [1] and the procedure to compute the modes of variations whereas Sect. 3 illustrates some shortcomings arising with a naive FPCA and applies the method presented in Sect. 2 to the Bank of Italy dataset (see Fig. 1).

2 Embedding and FPCA

Consider

$$\mathscr{L}or = \{L \in C^2_{[0,1]} : L(0) = 0, L(1) = 1, L' > 0, L'' > 0\},$$

where L' and L'' denote the first and second derivative of L respectively.
The following map

$$\psi(L) = -\ln\left(L''\right) + \int_0^1 \ln(L''(p))dp, \qquad \forall L \in \mathscr{L}or$$

is a bijection from $\mathscr{L}or$ into the separable Hilbert space $\mathscr{L}^2_c = \{g \in \mathscr{L}^2_{[0,1]} : \int g = 0\}$ and its inverse, for any $g \in \mathscr{L}^2_c$, is given by

$$\psi^{-1}(g)[p] = p + (p-1)\int_0^p z\exp\left(-g(z)\right)/\kappa_g dz + p\int_p^1 (z-1)\exp\left(-g(z)\right)/\kappa_g dz$$

where $\kappa_g = \int_0^1 \int_0^p \exp\{-g(z)\}dzdp$ is a scale technical factor. Hence, thanks to ψ, $\mathscr{L}or$ can be endowed with a Hilbert structure inherited by $\mathscr{L}^2_c$. This allows to properly perform FPCA and to compute modes of variations in $\mathscr{L}^2_c$ as usual. Moreover, ψ^{-1} can be used to map the obtained results back in $\mathscr{L}or$.
In particular, given a sample of empirical LCs $\{\widehat{L}_i(p), i = 1,\ldots,n\}$ each one estimated from a sample drawn from a random variable X_i, the following procedure can be implemented.

An embedding approach for Lorenz FPCA

1. Get $\widetilde{L}''_i(p)$ from $\widehat{L}_i(p)$ by using a suitable smoother (e.g. local polynomial).
2. Embed the LC in the Hilbert space $\mathscr{L}^2_c$ by means of ψ:

$$\psi(\widehat{L}) = -\ln(\widetilde{L}''_i) + \int_0^1 \ln(\widetilde{L}''_i(p))dp.$$

3. Implement the FPCA in $\mathscr{L}^2_c$ by computing the empirical
   - mean $\widehat{\mu}$, covariance operator $\widehat{\Sigma}$ and its eigenelements $\{\widehat{\lambda}_j, \widehat{\xi}_j\}$;

- j-th mode of variation of $\psi(\widehat{L})$ that is

$$\widehat{m}_{j,k} = \widehat{\mu} \pm k\sqrt{\widehat{\lambda}_j}\widehat{\xi}_j,$$

for any $k > 0$ and $j \in \{1, \ldots, n\}$.

4. Pull $\widehat{m}_{j,k}$ back into $\mathscr{L}or$ by using ψ^{-1}, to get $\widehat{M}_{j,k} = \psi^{-1}(\widehat{m}_{j,k})$ the j-th mode of variation in $\mathscr{L}or$.

The described procedure is statistically consistent since, under mild regularity conditions on the cdf F and as $n \to \infty$, $\widehat{M}_j(k)$ converges in probability to $M_j(k)$ the theoretical j–th modes of variation when LCs are integrally observed.

3 Application

In this section the proposed approach is applied to the Bank of Italy dataset (see Fig. 1). To better understand why an embedding approach is advantageous to study the modes of variation instead of a direct one approach, the FPCA is firstly performed on the original dataset of empirical LCs: the corresponding first three modes of variations for different k are plotted in Fig. 2. From the latter, it emerges that the direct approach provides coherent interpretations only for small values of k since for large values of k the modes of variations are no longer LCs. Fig. 3 depicts the modes of variations computed via the embedding approach for different k. As expected, since they are elements of $\mathscr{L}or$, it is possible to understand how the first three PCs impacts on the mean and how they explain the variability of LCs.

Another interesting point is the analysis of the information brought by the factor plane. Since the phenomenon under study is rather complex, some synthetic indexes, such as the Gini one, are often used to help the researchers. The PCs allow to explain the basic dynamics that regulate the composition of the LCs and therefore to go beyond the analysis of a single index. To do this, consider the track-plots that allow to appreciate the dynamics over time of the LCs with respect to the first two PCs; see Fig. 4. Note that, even if the Gini index for one specific region can assume similar values in distinct years, it can be placed in different quadrants of the factorial plane over the time suggesting the presence of latent structures that can not be detected by the synthetic index alone.

Acknowledgements The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The financial support of Università del Piemonte Orientale is acknowledged by the authors.

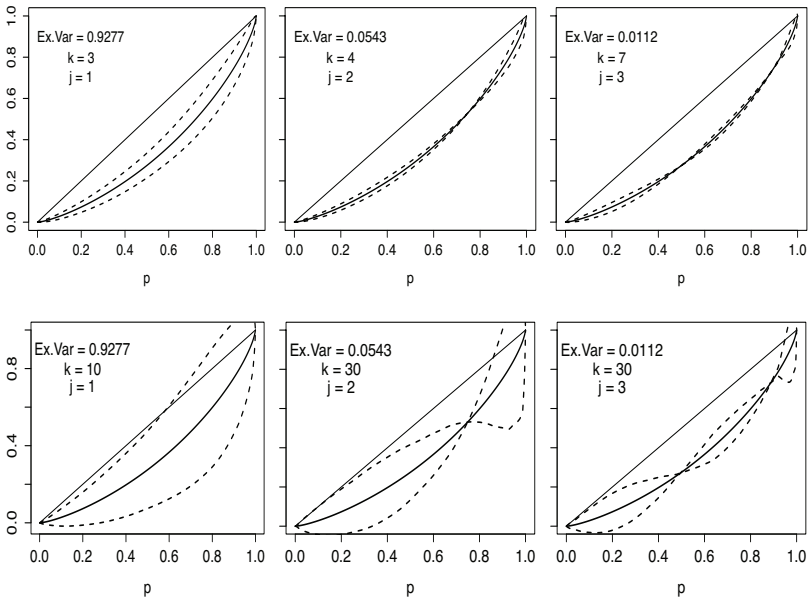


Fig. 2 Fraction of explained variance of the j-th PC, mean curve (solid line) and modes of variation for $j = 1,2,3$ and different k (dashed lines) for the sample of original LCs.

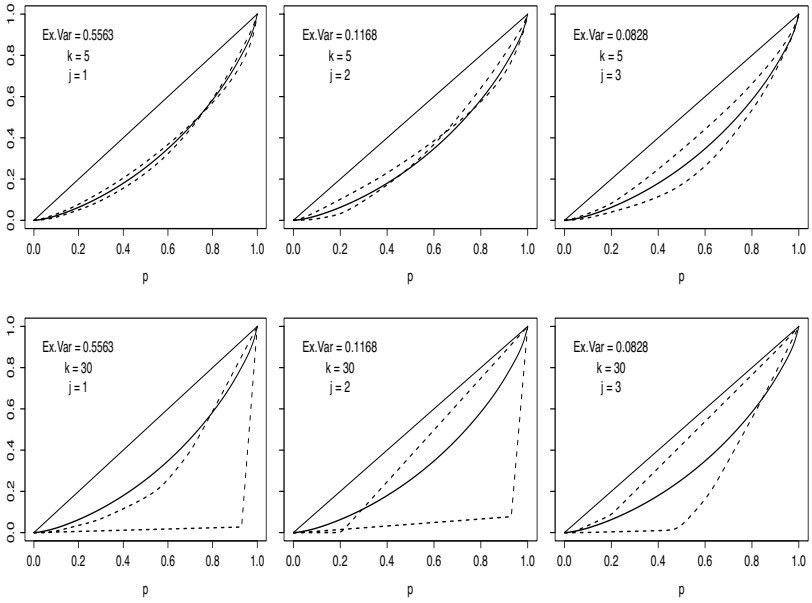


Fig. 3 Fraction of explained variance of the j-th PC, mean curve $\widehat{M}_j(0)$ (solid line) and modes of variation $\widehat{M}_j(k)$ for different values of j and k (dashed lines) for the sample of LCs in Fig. 1.

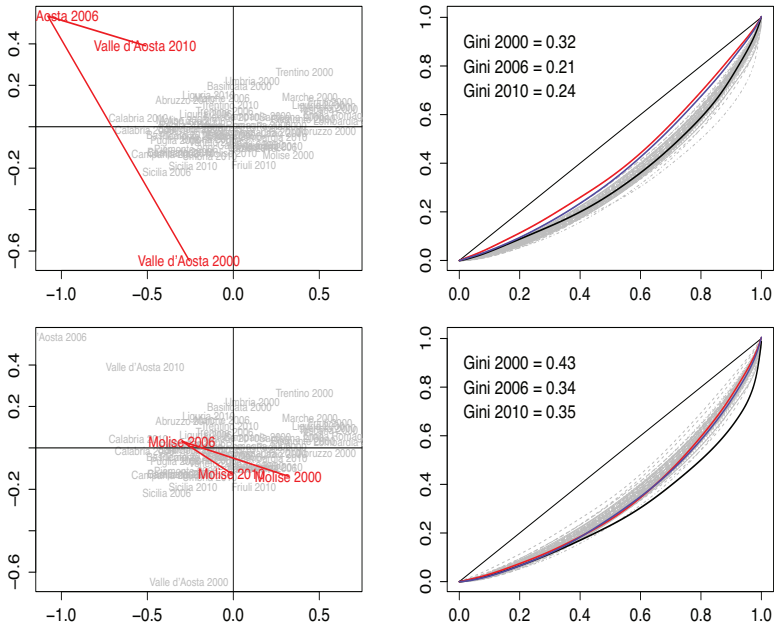


Fig. 4 (Left) Track-plots in the factorial plane of the first two PCs. (Right) Assosiated LCs and Gini indexes.

References

1. Bongiorno, E.G., Goia, A.: Describing the Concentration of Income Populations by Functional Principal Component Analysis on Lorenz curves. J. Multivariate Anal., **170**, 10–24 (2019)
2. Bosq, D.: Linear Processes in Function Spaces: Theory and Applications. Lectures Notes in Statistics, 149, Springer–Verlag, Berlin (2000)
3. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and practice. Springer Series Stat. (2006)
4. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. Chapman and Hall/CRC (2017)
5. Lorenz, M.O.: Methods of measuring the concentration of wealth. Amer. Statistical Assn. J., **9** (70), 209–219, (1905)
6. Ramsay, J.O., Silverman, B.W.: Functional data analysis, 2nd ed., New York: Springer (2005)

Analyzing textual data through Word Embedding: experiences in Istat

Analizzare dati testuali attraverso il Word Embedding: esperienze in Istat

Mauro Bruno, Elena Catanese, Massimo De Cubellis, Fabrizio De Fausti, Francesco Pugliese, Monica Scannapieco, Luca Valentino

Abstract *In recent years, language modelling and embedding spaces have attracted a huge attention within the community of researchers and official statisticians, the latter having started dedicated investments in the area. On the basis of the experiences that we have been making following this trend, in this study, we propose an advanced methodology to extract meaningful information from an unstructured corpus of textual data such as tweets. We present WordEmBox, a tool based on popular word embedding algorithms, namely: Word2Vec and Bterm for Topic Modeling (BTM) for short texts. We trained and tuned these algorithms to extract topic-oriented clusters of words. In the present work, we focus on a case study on Twitter data highlighting the findings of the approach.*

Abstract *Negli ultimi anni la modellazione del linguaggio naturale e degli spazi di embedding hanno attirato un'enorme attenzione all'interno della comunità di ricercatori e statistici ufficiali. In questo studio proponiamo una metodologia avanzata per estrarre informazioni significative da un corpus non strutturato di dati testuali come i tweet. Presentiamo la WordEmBox, uno strumento software basato su popolari algoritmi di Word Embedding: Word2Vec e Bterm per Topic Modeling (BTM) usato per testi brevi. Abbiamo addestrato e ottimizzato questi algoritmi per estrarre cluster di parole orientate ad argomenti significativi per le statistiche ufficiali. Nel presente lavoro ci concentriamo su un'analisi di dati di Twitter, illustrando i risultati ottenuti con il nostro approccio.*

[1] Mauro Bruno, Istat; mbruno@istat.it;
Elena Catanese, Istat; catanese@istat.it
Massimo De Cubellis, Istat; decubell@istat.it
Fabrizio De Fausti, Istat; defausti@istat.it
Francesco Pugliese, Istat; pugliese@istat.it
Monica Scannapieco, Istat; scannapi@istat.it
Luca Valentino, Istat; valentin@istat.it

Key words: word embedding, word2vec, semantic analysis, natural language processing, tweets

1 Introduction

In the field of unsupervised Machine Learning, Words Vector Spaces are arising as promising tools to extract word representations from wholly unstructured textual big data in an unsupervised way [1]. In general, these selected word representations are astoundingly good at capturing syntactic and semantic regularities within language patterns. Indeed, every relationship appears as a relation-specific vector offset enabling vector-oriented reasoning. The most important underlying insight of words' vector representations is the "distributional hypothesis": "You shall know a word by the company it keeps" [2]. Words Vector Spaces methods allow to train big corpora and in general perform better as size of the corpora increases.

The most popular methods in the Words Vector Spaces ecosystem can be divided into two families: (i) word embeddings that are context independent, i.e. these models produce as output just one vector (embedding) for each word, combining all the different senses of the word into one vector; (ii) methods that can generate different word embeddings for a word, thus capturing the context of a word - that is its position in a sentence (for instance Bidirectional Encoder Representations from Transformers (BERT) [3]).

Methods of the first class include:

- Word2Vec created in 2013 by the Google team [4], a toolkit that can train vector space models faster than the previous approaches.
- Global Vectors for Word Representations (GloVe) that can learn context and word vectors by factorizing a global word-word co-occurrence matrix [5].
- FastText, presented in 2017 by Facebook's AI team, which allows to train quickly models on large corpora and to compute word representations for words that do not appear in the training data [6].

Another very popular unsupervised learning task is Probabilistic Topic Modelling, which is used to extract latent semantic structures, usually related to topics, in an extended text body. Most popular methods are probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA). While word embeddings are prediction-based models, i.e the model, given the vector of a word, predicts the context word vectors, these methods are count-based models where similar terms have same counts for different documents. For both, word embeddings and PLSA, LDA, the similarity can be calculated using similarity's metric systems.

In Sect. 2 we provide an overview of the models used both in WordEmBox and in our simulations, namely Word2Vec, LDA and bi-term. In Sect.3 we provide a description WordEmBox's main functionalities. In Sect. 4 the main results of the application of our approach to a specific case study are presented.

2 Related Works

2.1 Word Embeddings: Word2Vec

The greatest novelty introduced by Word Embeddings is the "Embeddings Algebra", namely the resulting embedding space seems to have directions of semantic and syntactic meaning that can be exposed through simple operations on word vectors [7]. The word vectors capture syntactic and semantic regularities that are found by analysing a huge corpus. The similarity between vectors is usually calculated by dot-products or other vectorial operations.

Word2Vec is the most popular algorithm for learning word embeddings by harnessing the power of training a "shallow" neural network to learn word vectors [3]. According to the prediction target, Word2Vec adopts two different algorithms: Skip-gram predicts the context of a given word and continuous bag-of-words (CBOW) predicts the central word given the context. Both in Skip-gram and CBOW words are one-hot encoded and eventually, at the end of the training process, what it is taken in consideration for this task is not the predictive output but its internal structure. Basically, the outcome of the algorithm are the internal synaptic weights from the input to the hidden neural network, which, for each word, represent the coordinates of the word within the embedding space, namely the embedding vector.

Word2Vec provides several hyper-parameters to tune in order to enhance the quality of the learned language model. The main hyper-parameters are:

1) **Embedding space dimension:** the vector space size to which the words of the corpus are mapped
2) **Window size:** the width of the sliding window defining context's size
3) **Iterations:** the number of times the weights of the neural network are updated during training

2.2 Topic Modelling: Latent Dirichlet Allocation (LDA) and Biterm

The principal methods for Topic Modelling are: Probabilistic Latent Semantic Analysis (PLSA) [8] and Latent Dirichlet Allocation (LDA) [9].

PLSA is a technique to model co-occurrence information under a probabilistic framework in order to discover the underlying semantic structure of the data. Latent means that all the topics are treated as latent or hidden variables and are found by reducing the dimensions of a count matrix, namely a document-word matrix $N*M$, where N is the number of documents and M is the size of the vocabulary. While standard latent semantic analysis downsizes the occurrence tables usually via a singular value decomposition, probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model.

LDA is a generative probabilistic model that describes each document as a mixture of topics and each topic as a distribution of words. LDA generalizes PLSA with matrix factorization and works by decomposing the corpus document word matrix

(the larger matrix) into two parts (smaller matrices): the Document Topic Matrix and the Topic Word. LDA assumes that each document is generated by a statistical generative process, namely each document is a mix of topics, and each topic is a mix of words.
A weak point of LDA and PLSA is that the application of these models on short texts will suffer from the data sparsity problem [10], namely the sparse word co-occurrence patterns in individual document. To solve these issues Cheng et al. [1] have implemented a new model (Biterm), able to learn topics over short texts by directly modelling the generation of biterms in the whole corpus. Biterm assumes that two words in a biterm share the same topic extracted from a mixture of topics over the whole corpus. In this model a topic is also represented as a word distribution as conventional topic model.

3 WordEmBox

WordEmBox is a software tool developed by Istat [12], with the aim of providing a set of functionalities that allow the user to interact with Word Embedding (WE) models. The main purpose is to allow the exploration of WE models both through the native features of the model (i.e. affinity test and analogy test), and through their graphical representation based on the use of graphs.

As you can see in the figure below, the current version of WordEmBox offers a set of functionalities, i.e., Affinity, Analogy and Graphs, described in the following subsections.

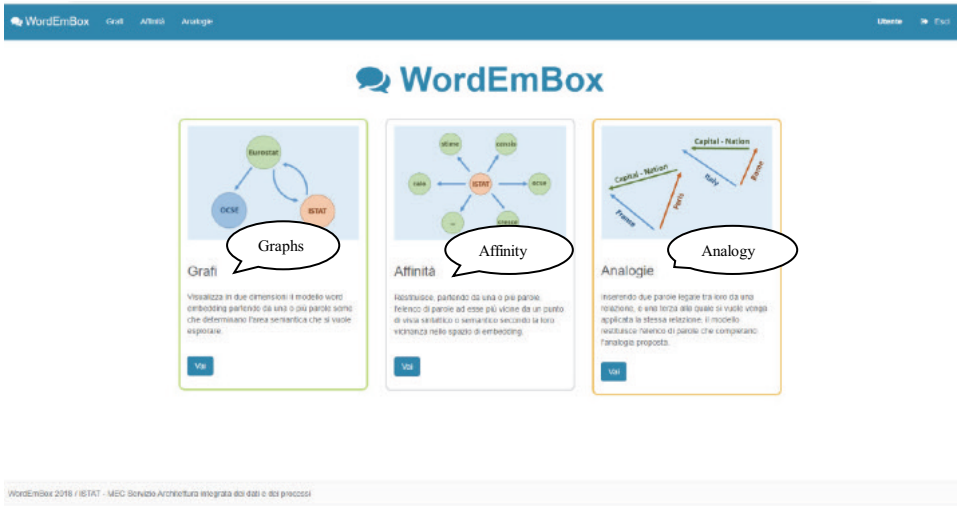


Figure 1: WordEmBox Homepage

3.1 Affinity functionality

This functionality, starting from one or more seed words, returns the list of closest words from a syntactic or semantic point of view, according to their proximity in the embedding space. The functionality requires two input parameters:

- One or more seed word(s): indicating the word(s)/vector(s) with respect to which the *n* closest words/vectors are searched in the embedding space, according to the cosine distance metric. It is also possible to insert more words from which to start the exploration; in this case the search for related words will take as reference (starting point) the vector sum of the words/vectors inserted. This feature is useful to solve disambiguation, that is the cases where words may have different semantic meanings (i.e., the word 'Rome' can be referred to a historical, geographical or sporting context).
- number of words: indicating the number of related words to be obtained as output.

3.2 Analogy functionality

The resolution of analogies is one of the most amazing features of word embedding models. It is based on vector calculation, which allows to calculate a semantic relationship as a vector-difference between couple of words. Indeed, by inserting two words linked by a relationship, and a third one to which the same relationship is to be applied (i.e., *man* is to *king* as *woman* is to *x*), the model returns the list of words that complete the proposed analogy (*x* equal to *queen*). This is possible thanks to the WE model representation of words/vectors.

The analogy functionality needs the following parameters:

- number of words: indicating the number of words to be returned as possible solutions of the analogy
- word 1, word 2, word 3: the three words that build the analogy

The output displays the list of words, found as possible solution to the analogy proposed, and the relative distances expressed according to the metric of the cosine distance.

3.3 Graph functionality

This functionality allows to visualize the WE model in a two-dimension canvas, starting from one or more seed-words that determine the semantic area to be explored.

As we know, the graphic representation of the WE model is quite difficult, due to the width of the embedding space. In addition to traditional statistical methods to reduce space dimensionality, maintaining the relationships between vectors (i.e., Principal Component Analysis and t- distributed Stochastic Neighbour), we have thought to use an alternative way to represent the Word Embeddings, i.e., graphs (mathematical structure made of nodes and edges). Graphs proved to be a very efficient tool in the visualization and manual exploration of WE model. Moreover, graphs are capable to represent such models by bringing out the clusters of words close to each other and their syntactic and semantic relations.

Graphs and all their expressive power have been embedded in the WordEmBox software application, as one of the main features. We have devised the following three different methods for combining basic graphs according to different exploration strategies: (i) geometric; (ii) linear; (iii) geometric oriented.
All three types of graphs require the following input parameters: the *width*, the *number of iterations* and one or more *seed-words* to define the starting point for exploring the embedding model. The *width* parameter determines how many words/nodes close to the seed word (or from the second iteration onwards, close to the words found) must be displayed in the graph at each iteration. The *iteration* parameter determines the number of desired iterations and finally the *word* parameter indicates the seed word(s) from which to start the model's exploration. Moreover, in the WordEmBox there are two other additional parameters: *mode* and *layout,* which respectively determine the graphical appearance of the graph and the type of graph (geometric, linear, geometric oriented).
The differences between the three types of graphs are that the *geometric graph* tends to expand the range of exploration very quickly, rapidly losing the initial semantic focus provided by the seed words; the *linear graph* remains much more focused, but only explores a narrow sub-model; the *geometric-oriented graph* often provides a satisfactory compromise between the previous two.

4 Results

In this paragraph we investigate a sample of one month period tweets collected by using the Istat[1] economic filter consisting of a list of relevant keywords. The dataset is composed by 855,865 tweets. The aim of the following analysis is to understand the impact of the Ukrainian crisis on the economic mood of Italian Twitter users as observed through our filter. For this reason, the time window was set from 20th February (few days before the start of the conflict) to 20th March. Let us notice that the filter has not been specifically designed to sample tweets related to the war, and indeed only 10 percent of tweets in the sample contain the word "guerra" (war).

In the first section we compare the word embedding WordEmBox model trained on this dataset with a model trained with tweets from June 2016 to June 2017 (SMEI17). For both trainings we used a CBOW model and windows=8, dimension=200.
In the second section we perform a topic analysis, by analyzing two different time windows and compare two topic modelling approaches, traditional LDA and bi-term with a k-means clustering built from the WordEmBox model.

4.1 WordEmBox analysis of the Russia-Ukraine conflict

In this case study, we use the functionalities of WordEmBox to explore Italian Twitter users' discussion about the Russian-Ukraine conflict and their concerns related to economy. We show the WordEmBox functionalities namely the graph (geo-oriented) and affinity.
Fig. 2 shows the graph analysis, and the affinity words list of the word "Guerra" (War). The graph shows three areas: the first area (on the top of Fig. 2) concerns the consequences of the war on the Italian economy, i.e. "crisi" "recessione" and "inflazione" (crisis, recession and inflation); the second area (in the middle) concerns the conflict, i.e. the words "conflitto", "invasione", "sanzioni"(conflict, invasion and sanctions); finally the third area (in the bottom) concerns politics, and is closely related to the second , i.e. the words "Ucraina", "Putin", "Russia", "Europe" and "USA" .

[1]More detailed information about the Italian Social Mood on Economy Index (SMEI) can be found at https://www.istat.it/it/files//2018/07/Methodological_Note_social-mood.pdf

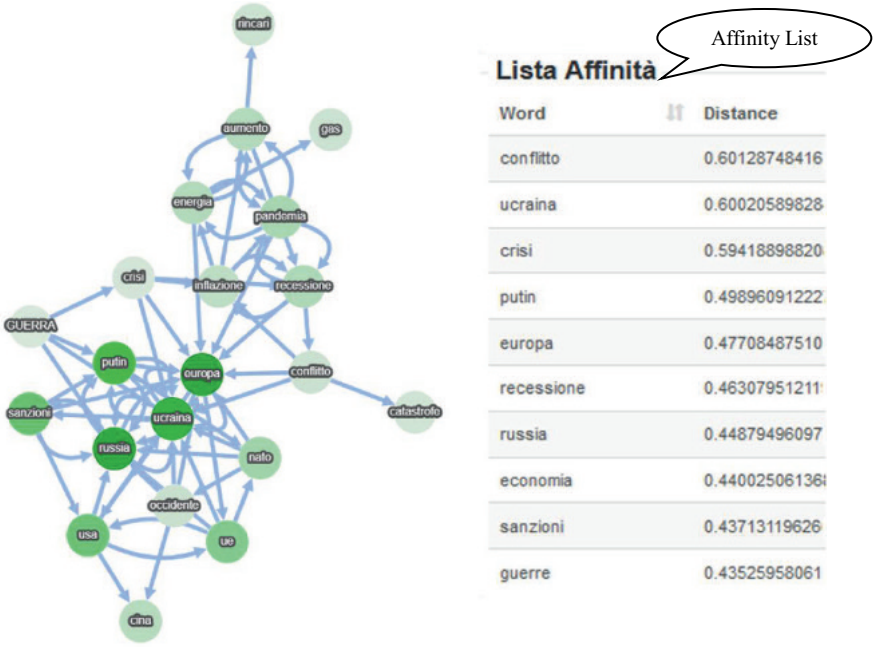


Figure 2: Graph analysis and affinity list for word "GUERRA" with the WordEmBox

In Fig. 3 and Fig. 4 we compare the graph analysis of the word "prezzi" (prices) and "banche" (banks) of the RUC model with the SMEI17 model.
In the RUC model the Twitter users express their concerns on prices related to the increase of costs of fuel and gas, while the conversations previously concentrated to the purchasing power of salaries in buying and in consumption goods.

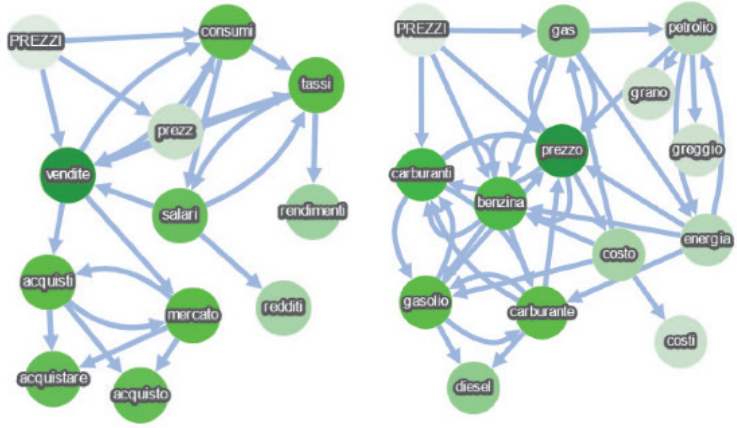


Figure 3: Graph analysis for word "PREZZI" in the two models 2017SME and UkraineWar

Looking at Fig. 4, it can be observed that while in RUC the debate on banks focuses on sanctions for Russians account holders (SWIFT), in the SMEI 2017 the focus was on the crisis witnessed by the Italian banks Monte dei Paschi (MPS) and Banche Venete.

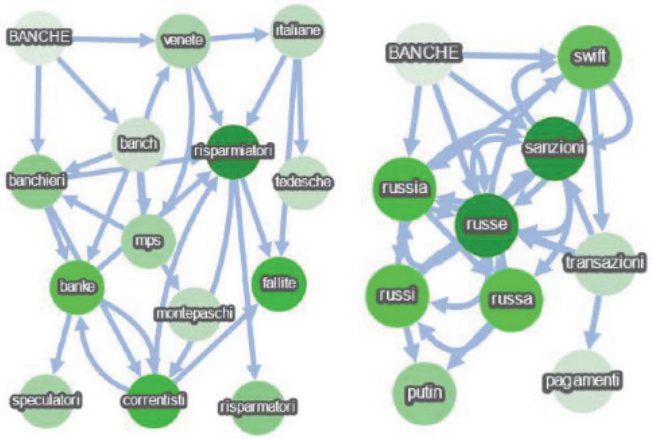


Figure 4: Graph analysis for word "BANCHE" in the two models 2017SME and UkraineWar

4.2 Topic modelling

In the previous section we analyzed word embedding graphs and affinities for one month period. In this section, we characterize groups of words related to two contiguous time periods relating to the first 4 weeks of the war in Ukraine: 20 February - 6 March (P1) and 7 March - 20 March (P2). In particular, a challenging analysis that we will show spots the dynamics of the terms from one temporal period to another.

Starting from a Word Embedding, it is possible to figure out partially automatic way of identifying relevant topics that are latent in the WE representation. To such a purpose, we designed and implemented a Word2Vec model for each time window (same tuning as in the previous sub-section). For each resulting Word2Vec model, we select a sub-model (WE1 and WE2) consisting of the vectors related to the 1000 most frequent terms in the corpus. On each sub-model, we ran a cluster analysis via k-means [13] with the usual similarity distance. The optimal number of clusters, k=11, was identified via silhouette metric analysis [14] and turned out to be the same in WE1 and WE2. Each cluster does actually represent a topic to be analyzed. A preliminary descriptive analysis (see Fig. 5) shows the number of words emerging in each topic in the periods P1 and P2. In Period 2 we observe a predominant topic containing 191 words, while in Period 1 there is less variability in the number of words per topic. Our aim is to identify relevant topics. For this reason, for each

period P1 and P2, we pruned some topics either because containing only terms with a "syntactical" relevance (e.g. pronounce, adverbs, etc,) or because not informative.

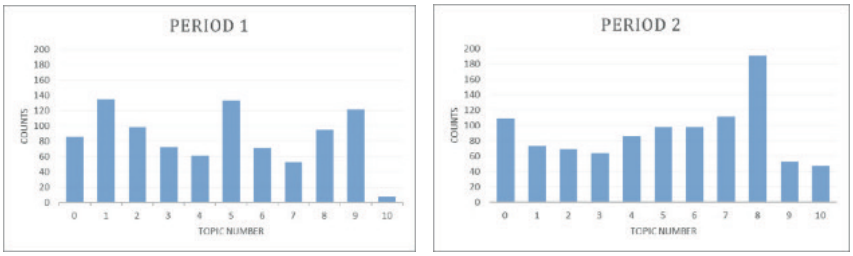


Figure 5: Counts of words in each topic in the two periods

We then analyzed the dynamics of terms in relevant topics. By looking at Fig. 6, it is possible to figure out term flows between topics from one period to another. In Fig. 6 only the most relevant flows are shown. More precisely: in the first period we observe two topics related to the Russian Ukrainian conflict: one related to war and economy, the other to energy and gas. In the second period, the topic containing the word war splits into two more specialized topics in one focused banks and the other containing the word war focuses on poverty. The topic related to gas and energy is similar in the two periods. The other two topics are: one concerning Italian political issues and the other one related to work and salaries. We observe that in the second period the concern of Italian Government is on military expenses.

Figure 6: Term flow analysis of the relevant topics. In English (left side): 0=gas, inflation, crisis; 1=economy, Russia, war, Ukraine, 10=Draghi, land registry, minimum, reform, 5=without work, Italy, Italians. In English (right side): 0=gas, euro, fuel, increase; 2=Russia, Ukraine, Putin, sanctions 4=war, country, poverty, occupation; 5=expense, government, military, millions; 7=taxes, labour, without, Italians

We then compared these results with the topic modelling obtained by bi-term. For the sake of comparability, we set the same number of topics. In this case we focus on the clusters containing the word war in the two periods P1 and P2. This topic modelling, shown in Fig. 7, displays the word cluster as circles whose diameter is proportional to the ratio between the count of the words contained in the cluster and the total words in the tweets' set considered. In this case the topics related to war automatically emerge, because the method allows clusters to have same words. We

observe analogously as above that in the second period there are more topics related to the conflict. Indeed, there are two topics in the first period and four in the second one. If we compare with the previous analysis, we see that in the first period the predominant topic of bi-term contains the same key-words (Economy, Russia, Ukraine) of the topic containing the term war of the previous model. If we compare the topic models in period P2, we observe the topic related to energy and gas, is associated to two topics (T1, T3) of bi-term and the relationship with the conflict increases with respect to P1. The topic on military expenses in this case is observed in both periods P1 and P2.

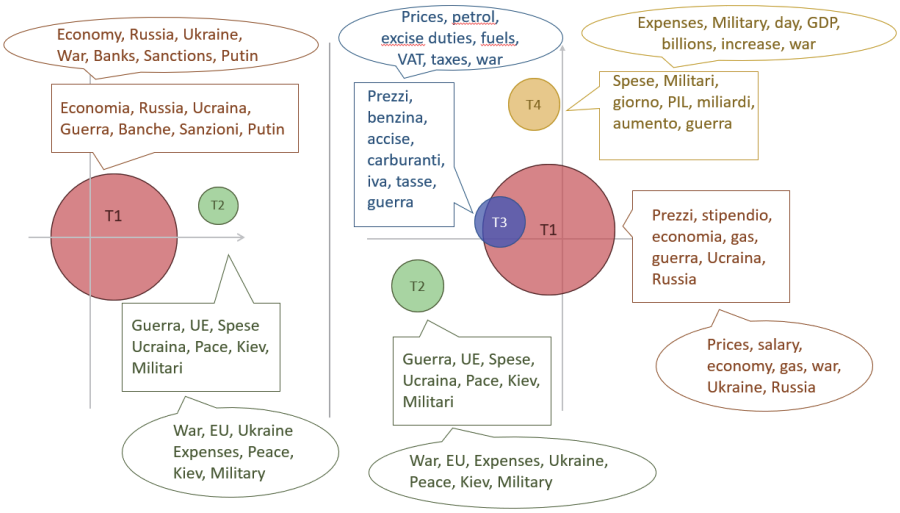


Figure 7: biTerm Topic Analysis over P1 (left) and P2 (right), clusters containing the term war.

5 Conclusions and Plans for the Future

The WordEmBox approach has shown the feasibility and usefulness of a human-driven exploration of word embedding spaces. Additionally, the topic modelling approach has proven very effective with a higher degree of automation. Overall, both the approaches have shown coherent and significant results. Indeed, the WordEmBox, with its graph functionalities, allows for a given word to identify a topic, similarly the topic modelling approach "captures" clusters of words related to certain topics. They also allow to evaluate the dynamics of topics between different time periods. Concerning the Italian Official Statistics needs, we are planning to update for the WordEmBox, to make it a more versatile analysis tool for WE models. More specifically WordEmBox application will be enriched with the following tools and facilities:

- The possibility to skip from one WE model to another in automatic manner, without the need for IT support.
- The k-means clustering based on WE model, described in Sect. 4.2, and a visualization of Word Clouds.
- The visualization of WE models through rotatable 3D graph structures.
- The introduction of an information on how close the words are to each other, by updating the current graph visualization with lines of different thickness.

6 References

1. Mikolov, T., Yih, W. T., & Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 746-751) (2013)
2. Firth, J.R. "A synopsis of linguistic theory 1930–1955". In: Studies in Linguistic Analysis: 1–32. (1957)
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. In: Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, pages 4171–4186. (2019)
4. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. In: CoRR abs/1301.3781 (2013b).
5. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, (2014)
6. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. Enriching word vectors with subword information. In: Transactions of the association for computational linguistics, 5, 135-1 (2017)
7. Levy O., Goldberg Y., Dagan I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. In: Trans. of the Association for Computational Linguistics, vol.(3): 211-225 (2015).
8. Thomas Hofmann. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, (1999)
9. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In: J. Mach. Learn. Res., 3:993–1022, March (2003).
10. L. Hong and B. Davison, "Empirical study of topic modeling in Twitter," In: Proceedings of the First Workshop on Social Media Analytics. ACM, pp. 80–88 (2010)
11. Cheng, X., Yan, X., Lan, Y., & Guo, J. Btm: Topic modeling over short texts.In: IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928-2941 (2014).
12. De Fausti F., De Cubellis M., Zardetto D.: Word Embeddings: a Powerful Tool for Innovative Statistics at Istat. In proceedings of JADT 2018, pages 174-182 (2018)
13. Jin X., Han J. *K*-Means Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425 (2011)
14. Wang, F., Franco-Penya, HH., Kelleher, J.D., Pugh, J., Ross, R. An Analysis of the Application of Simplified Silhouette to the Evaluation of *k*-means Clustering Validity. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2017. Lecture Notes in Computer Science, vol 10358. Springer, Cham. https://doi.org/10.1007/978-3-319-62416-7_21 (2017)

Functional Horvitz-Thompson estimator for convex curves

Stimatore di Horvitz-Thompson di curve convesse

Adelia Evangelista, Francesca Fortuna, Stefano Antonio Gattone, Tonio Di Battista

Abstract The work presents the Horvitz-Thompson estimator of convex curves in a functional setting. In many application fields, such as the analysis of Lorenz curves, or diversity profiles, practitioners work with convex curves. In such cases, a naive application of the Horvitz-Thompson functional estimator can lead to non-convex estimates. Thus, a constrained Horvitz-Thompson estimator for convex curves is introduced by defining these functions as a solution of a differential equation. A suitable sampling distribution of the proposed mean estimator has been derived, allowing to build simultaneous confidence bands, whose performance has been assessed by means of a simulation study.

Abstract *Il lavoro presenta lo stimatore di Horvitz-Thompson di curve convesse in un contesto funzionale. In molti ambiti applicativi, come nel caso delle curve di Lorenz o dei profili di diversità, si considerano curve convesse. Un'applicazione diretta dello stimatore funzionale di Horvitz-Thompson può condurre a stime non convesse. A tal fine, viene proposto uno stimatore di Horvitz-Thompson vincolato per curve convesse, definendo queste ultime come una soluzione di un'equazione differenziale. È stata ricavata un'opportuna distribuzione campionaria dello stimatore proposto, che consente di costruire bande di confidenza simultanee, valutate mediante uno studio di simulazione.*

Key words: Functional Horvitz-Thompson, Confidence bands, Convex curves

---

Adelia Evangelista
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: adelia.evangelista@unich.it

Francesca Fortuna
University of Roma Tre, Via Silvio d'Amico, 77, e-mail: francesca.fortuna@uniroma3.it

Stefano Antonio Gattone
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: gattone@unich.it

Tonio Di Battista
University of Chieti-Pescara, Viale Pindaro, 42,e-mail: tonio.dibattista@unich.it

1 Introduction

This work reports the main results of a recent paper [10]. The aim is to obtain a constrained functional Horvitz-Thompson estimator in a design-based inference context. The basic motivation was that in several application fields such as income inequality [17], biological diversity [6, 7], ecology [5], industrial concentration [11], reliability [4] and disease risks [13], researchers work with convex curves. Example of convex curves are the Lorenz curve [12] in economics and the diversity profile [9] in ecology.

In order to build an estimator of the mean of convex curves in the functional domain two main topics need to be considered. The first one is related to the application of common methodologies of functional data analysis (FDA) in the presence of shape constraints, such as convexity [8]. FDA techniques usually require that functions have value in the separable Hilbert space of square integrable functions, known as $L^2(\mathscr{X})$, with the inner product $< f,g >= \int_{\mathscr{X}} f(x)g(x)\,dt, \forall f,g \in L^2(\mathscr{X})$, and the L^2 norm $||f|| =< f,f >^{1/2} < \infty$. The second matter regards the use of FDA in a design-based inference setting. In 2010s, Cardot et al. [1] investigated this issue by studying the theoretical properties of functional principal components when the curves are collocated with survey sampling techniques. Specifically, Cardot and Josserand [2] presented a Horvitz-Thompson estimator for functional data, under the hypothesis of error-free measurements. Later, this assumption was removed, estimating the data with local polynomials [3].

Gattone et al. [10] expanded the Cardot's approach to convex functions. The work focuses on the estimation of the Lorenz curve in a designed-based approach. As it is well known, the family of Lorenz curves are constrained to be non-negative, bounded, increasing and convex. In this framework, a suitable transformation of the Lorenz curve is proposed, using the differential equation approach introduced by Ramsay in 1998 [15]. In this way, the estimation is performed in the unconstrained $L^2(\mathscr{X})$ space. The resulting constrained Horvitz-Thompson estimator is biased for finite samples with an unknown distribution. Therefore a delta method procedure is implemented for reducing the bias, and to get a consistent and asymptotically normal estimator. Thus, confidence bands could be constructed by means of simulations of functional Gaussian processes based on the estimated covariance function.

The reminder of the work proceeds as follows. In Sect. 2 we derive the functional Horvitz-Thompson estimator of Lorenz curve with its asymptotic confidence bands. Sect. 3 describes the results of the simulation study, and conclusions are presented in Sect. 4.

2 Functional Horvitz-Thompson estimation of Lorenz curve

As outlined in Sect. 1, the aim of this work is to provide a consistent Horvitz-Thompson estimator of the Lorenz curve in a design-based approach. The first step is to move the observed Lorenz curves into a suitable functional space. To this end,

the differential equation approach proposed by Ramsay in 1998 [15] and previously adopted by Fortuna et alv[9] has been applied.
The Lorenz curve can be expressed as follows:

$$f(p) = CD^{(m-1)} \exp\Big(D^{-1}w(p)\Big), \quad 0 \leq p \leq 1, \tag{1}$$

that is as the solution of the differential equation $D^m f(p) = w(p)D^{m-1}f(p)$, where D^m is the m-th derivative of $f(p)$, $D^{-1}w(p) = \int_0^p w(s)ds$ represents the partial integration operator and $w(p)$ indicates an unconstrained function [10]. We set $C = \frac{1}{D^{-2}\exp(D^{-1}w_1)}$, in order to ensure that $f(1) = 1$. It is easy to prove that $f(p)$ in Eq. (1) is a convex curve when $m = 3$ [15]. The unconstrained function $w(p)$ can be written as a linear combination of basis functions:

$$w(p) = \sum_{b=1}^{B} c_b \phi_b(p), \tag{2}$$

where c_b stands for the $b-th$ coefficient and $\phi_b(p)$ is the $b-th$ basis function and $w(p)$ lies in the Hilbert space, $L^2(\mathscr{X})$. Then, the Lorenz curve can be rewritten in the constrained functional formm by putting $w(p)$ in Eq. (1).

The procedure follows with the definition of the Horvitz-Thompson (HT) estimator for the mean of convex curves. Starting from the HT estimator introduced in [2] and [3], the unconstrained functional HT estimator of the mean is defined as follows:

$$\overline{w}_{HT}(p) = \frac{1}{N}\sum_{i\in s}\frac{w_i(p)}{\pi_i} = \frac{1}{N}\sum_{i\in s}\frac{\sum_{b=1}^{B} c_{bi}\phi_b(p)}{\pi_i}, \tag{3}$$

where N is the size of the finite population U, $\{w_i(p)\}_{p\in[0,1]}$ is the unique function associated to each $i-th$ unit $\in U$, π_i are the first order inclusion probabilities and $s \in U$ is the selected sample of size n.

Including in Eq. (1) the estimator defined in (3), it is possible to construct the constrained HT estimator of the mean of convex curves as follows:

$$\overline{f}_{HT}(p) = \bar{C}D^{-2}\exp\Big(D^{-1}\overline{w}_{HT}(p)\Big). \tag{4}$$

with $\bar{C} = \frac{1}{D^{-2}\exp D^{-1}\overline{w}_{HT}(1)}$.

Finally, the application of the delta method allows the evaluation of the bias as follows:

$$\widehat{bias}\Big(\overline{f}_{HT}(p)\Big) = \frac{1}{2n^2}\Big[g''\Big(\overline{w}(p)\Big)\Big]\hat{\gamma}(p,p'), \tag{5}$$

where g stands for a non-liner function, $p \in [0,1]$, $g''\Big(\overline{w}(p)\Big)$ defines the second derivative and $\hat{\gamma}(p,p')$ is the covariance function of the unconstrained estimator.

Moreover from [14], it follows:

$$\sqrt{n}\left\{ g\left(\overline{w}_{HT}(p)\right) - g\left(\overline{w}(p)\right) \right\} \overrightarrow{d}\, \mathcal{N}\left(0, \left[g'\left(\overline{w}(p)\right)\right]^2 \gamma(p,p')\right), \quad \forall p \in [0,1]. \quad (6)$$

From this result it is possible to derive the pointwise confidence interval as follows:

$$P\left(\overline{f}_{HT}(p) \in \left[\overline{f}^*_{HT}(p) \pm q_\alpha \sqrt{Var\left(\overline{f}^*_{HT}(p)\right)}\right]\right) = 1-\alpha, \quad \forall p \in [0,1], \quad (7)$$

where q_α is the quantile of order $1-\alpha/2$ of the standard Normal distribution, and $Var\left(\overline{f}^*_{HT}(p)\right) = \left[g'\left(\overline{w}(p)\right)\right]^2 \hat{\gamma}(p,p')$.

Simultaneous confidence bands for the HT estimator are derived for the entire domain at once by:

$$P\left(\overline{f}_{HT}(p) \in \left[\overline{f}^*_{HT}(p) \pm c_\alpha \sqrt{Var\left(\overline{f}^*_{HT}(p)\right)}\right], \forall p \in [0,1]\right) = 1-\alpha, \quad (8)$$

defining c_α as an approximation of the supremum of a Gaussian process [3].

3 Simulation study

A simulation study is implemented to assess the behaviour of the Functional Horvitz-Thompson estimator of Lorenz curve proposed. All the computational aspects have been implemented through the R software [16].

Using a generalized distribution of second kind (GB2) the income vector of a finite population of 50000 units have been generated. Than, we consider three scenarios with different level of income inequality, which are respectively Low, Medium and High concentration (see [10] for more details).

In Fig.1 the pointwise empirical coverage are shown. The results highligth a good behaviour in the low concentration scenario, with the pointwise empirical coverage always between 0.9 and 1. Considering the Medium and High scenario, for $p > 0.8$ the coverage declines, indicating a worse behaviour in the final part of the curve.

The simultaneous empirical coverage is reported in Table 1. The values obtained confirm a conservative behaviour in the case of the Low concentration; in the Medium concentration scenario the empirical level equalizes the nominal coverage and for the High scenario there is a moderate under-coverage.

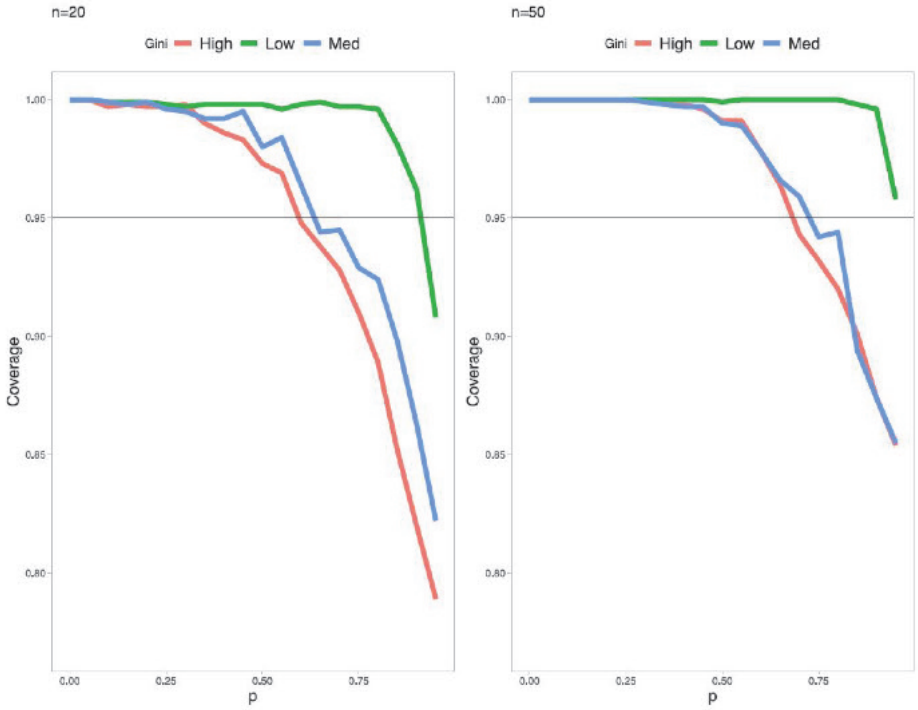


Fig. 1 Pointwise empirical coverage. Nominal level $1-\alpha = 0.95$.

Table 1 Empirical coverage of the *SCB* simultaneous confidence bands under different scenarios and sample size. Nominal level $1-\alpha = 0.95$.

Scenario	n	*SCB*
Low	20	0.99
concentration	50	0.99
Medium	20	0.95
concentration	50	0.96
High	20	0.91
concentration	50	0.92

4 Conclusion

The main results of a recent paper appeared on Econometrics and Statistics have been presented. In particular the functional Horvitz-Thompson estimator of Lorenz curve has been derived. Theoretical results have been derived and a simulation study has been implemented in order to evaluate the performance of the proposed procedure. Results show a good performance of the estimator. Future works will focus on to more complex constrained indicators both in social and ecological fields.

References

1. Cardot, H., Chaouch, M., Goga, C., Labruere, C.: Properties of design-based functional principal components analysis. J Stat Plan Inference **140**, 75–91 (2010)
2. Cardot, H., Josserand, E.: Horvitz-Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. Biometrika **98**, 107–118 (2011)
3. Cardot, H.,Degras, D., Josserand, E.: Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. Bernoulli **19**, 2067–2097 (2013)
4. Chandra, M., Singpurwalla, N.: Relationships between some notions which are common to reliability theory and economics. Math. Oper. Res. **6**, 113–121 (1981)
5. Damgaard, C., Weier, J.: Describing inequality in plant size or fecundity. Ecology **81**, Issue 4, 1139–1142 (2000)
6. Di Battista, T., Gattone, S.A.: Simultaneous inference on diversity of biological communities. Stat Methods Appt **13**, 129–136 (2004) doi: 0.1007/s10260-003-0076-9
7. Di Battista, T., Fortuna, F.: Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). Stat Anal Data Min **10**, Issue 1, 21–28 (2017)
8. De Sanctis, A., Di Battista, T.: Functional analysis for parametric families of functional data.Int. J. Bifurc. Chaos **22**, Issue 9, 1250226–1250232 (2012)
9. Fortuna, F., Gattone, S:A., Di Battista, T.: Functional estimation of diversity profiles. Environmetrics **31**, Issue 8, e2645 (2020) doi: 10.1002/env.2645
10. Gattone, S.A., Fortuna, F., Evangelista, A., Di Battista, T.: Simultaneous confidence bands for functional mean of convex curves. Econ. Stat. (2020) doi: https://doi.org/10.1016/j.ecosta.2021.10.019
11. Hart, P.E.: Entropy and other measures of concentration. J. R. Stat. Soc., A **134**, 73–89 (1971)
12. Lorenz, M.O.: Methods of measuring the concentration of wealth. J Am Stat Assoc **9**, 209–219 (1905)
13. Mauguen, A., Begg, C.B.:Using the Lorenz Curve to Characterize Risk Predictiveness and Etiologic Heterogeneity. Epidemiology **27**, Issue 4, 531–537 (2016)
14. Oehlert, G.W.: A Note on the Delta Method. Am. Stat. **46**, Issue 1, 27–29 (1992)
15. Ramsay, J.O.: Estimating smooth monotone functions. J. R. Stat. Soc., B **60**, 365–375 (1998)
16. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006) http://www.R-project.org
17. Kakwami, N.C.: Applications of Lorenz Curves in Economic Analysis. Econometrica **45**, Issue 3, 719–728 (1977)

Children, parents, grandparents: a look on changing relationships

Changes in social relationships of Italian older people. Evidence from FSS and SHARE Corona surveys

I cambiamenti nelle relazioni sociali degli anziani in Italia. Evidenze dalle indagini FSS e SHARE COVID-19

Elvira Pelle, Giulia Rivellini and Susanna Zaccarin

Abstract Using data from the 2016 Family and Social Subjects (FSS) survey and the first wave of the SHARE Corona Survey carried out in 2020, we aimed to depict the social network characteristics of older people in Italy, with a special focus on tracing the changes related to the outbreak of the SARS-CoV-2 virus.
Abstract *Sulla base dei dati relativi all'edizione 2016 dell'indagine Famiglie e Soggetti Sociali (FSS) e della recente indagine SHARE Corona condotta nel 2020, sono descritte le caratteristiche delle reti sociali della popolazione anziana in Italia, sottolineando i cambiamenti collegati alla diffusione del virus SARS-CoV-2.*

Key words: ego-centered networks, older people, Covid-19 pandemic-lockdown, SHARE Corona 1

1 The impact of Corona on social relationships

The implications of the COVID-19 pandemic and the associated containment measures on demography, society, and economy have been recently object of study for many researchers who focused on direct [8, 9] and indirect consequences [3, 4, 13]. Such implications also widely include changes in the network of social relations binding individuals to the people close to their everyday lives and in the avail-

Elvira Pelle
Department of Communication and Economics, University of Modena and Reggio Emilia, e-mail: elvira.pelle@unimore.it

Giulia Rivellini
Department of Statistical Science, Università Cattolica del Sacro Cuore, e-mail: giulia.rivellini@unicatt.it

Susanna Zaccarin
Department of Economics, Business, Mathematics and Statistics, University of Trieste, e-mail: susanna.zaccarin@deams.units.it

ability of tangible and intangible resources they exchange. As a result, the social space, which takes shape in the relations inside the (immediate or extended) family, and with friends, coworkers, or neighbors could have been strongly reduced by the health emergency period. Containment policies included a number of measures aimed at reducing physical contacts, considered an important factor in the SARS-CoV-2 transmission and these measures have been prolonged more than others, limiting personal networks of contacts and social gatherings. Later, during second and third waves, self-protective behaviours–mainly induced by the fear of being infected and by awareness of COVID-19 exposure among those in one's social network (*network-exposure* severity)– have changed habits and ways of spending time together, especially among older women aged 70 and more, as emerged in several European countries, such as Spain, Italy and Portugal [12].

The COVID-19 pandemic got also worsen the issue of social isolation and its consequences on mental health. Man is a social animal and the group - in its various forms - is the cornerstone for the development and progress, but also for the survival of himself. Due to the media effect, fear, and uncertainty about the situation, many people went under segregation circumstances, with an increase in the risk of mental disorders, level of anxiety and depression symptoms [18]. Among older adults in the 27 European countries –included both in regular and COVID-19 wave 8 of the "Survey of Health, Aging and Retirement in Europe" (SHARE, [5])– the lack of partner and children has exacerbated loneliness [3]. Southern European men and women had the largest reduction in all activities and health measures, with women experiencing the largest negative changes across all social activities and health measures consistently across European regions [16].

These consequences could be more severe for individuals living alone, to whom is associated a condition of relational vulnerability, especially for older people. The elderly are indeed generally more vulnerable than other population groups and need additional care and services in both pandemic and non-pandemic times. However, the aging process is highly diverse and context-dependent, with different levels of vulnerability regarding also the types of personal networks on which older people are embedded in their daily life [10].

In this contribution we aim to depict the ego-centered network characteristics of older people in Italy, with a special focus on tracing the changes related to the outbreak of the COVID-19 virus. More specifically, the analysis is carried out on two specific age groups: individuals aged 65 and over, living as or as a partner in a couple without cohabiting children, and with no other family (or non-family) members. The social networks of individuals in these two specific living arrangements are composed by others from outside, allowing a clear picture of the social contacts they can count, especially if pandemic restrictions are in force.

We use Family and Social Subjects (FSS) survey data to build the elderly's ego-centered network of contacts before the Corona outbreak as well as to explore the potential impact of the containment measures on social networks in Italy. Evidence from the first wave of the SHARE Corona Survey (hereafter SHARE Corona 1) targeted to the COVID-19 pandemic living situation of older people [17] will then

provide a basis ground to assess changes in social relationships effectively experienced by older Italian people.

2 A temporal line of observation through different surveys

Data on contacts and social relations that individuals entertain with others are often collected through large-scale national or international surveys. Focusing on Italy, approximately every five years since 1998, the Family and Social Subjects (FSS) survey carried out by the Istat (https://www.istat.it/it/archivio/256707) represents the primary statistical source collecting data on contacts interviewers entertained in person with not-cohabiting others (usually referred as *alters*), such as siblings, children and grandchildren, parents, grandparents, friends, and neighbors, as well as on support interviewers provided(received) to(from) outside the household. Moreover, the frequency of phone calls with not-cohabiting siblings, children and grandchildren, parents, and grandparents, is also collected, and in the last FSS edition, carried out in 2016, the frequency of video calls and messages (through sms, WhatsApp, email, social media) with the ”kin” as alters has been investigated. The frequency of face-to-face (f-t-f) contacts with friends has also been added, thus enriching previous information related only on the presence of this important alter category.

The FSS is based on a wide probability sample, allowing detailed network analyses in specific groups (by age, by living arrangements, etc.) of the population. The ego–centered network of contacts derived from the FSS can be regarded as the privileged group of alters with whom the respondent can potentially entertain or exchange relationships, although the content of the relationships is not specified for all alter types [1, p. 813].

Since the new FSS edition will not be planned before the next year, we have explored the potential impact of the containment measures on social relations of vulnerable groups of population in Italy, on the basis of 2016 FSS data [10]. Two ego-centered network definitions accounting for physical distance in light of the COVID-19 containment measures have been proposed, elaborating on the different meaning assumed for residential proximity and frequency of in-person contacts by the containment measures adopted in Italy [1], since March 2020. The first one named “easy-to-reach” network represents an accessible source of support including only the alters that live in the same municipality of the ego (the elder in our analysis), regardless of the frequency of f-t-f contacts; it can be activated in case of a new lockdown or a similar other emergency situation. The second network definition refers to the “accustomed-to-reach” network, which includes proximity as well the habit to meet in person since it is more likely that alters in the “accustomed-to-reach” network can be a real source of support in situations of reduced possibility to travel with respect to the “easy-to-reach” network. Besides these simulated personal networks, changes in relationships people indeed experimented in the first wave of the

[1] Istat (2020) confirms that, due to the lockdown, most Italians did not visit other people, and most people dedicated more time than usual to phone or video calls with relatives and friends.

COVID-19 pandemic (i.e., between March and June 2020) can be evaluated using the novel information from the SHARE Corona 1 [6], collected mainly between June and August 2020 and covering all EU Member States. As reported in [15], the SHARE-COVID19 project aims to understand pandemic "non-intended consequences and to devise improved health, economic and social policies." The questionnaire covers the most important life domains for the target population and asks specific questions about infections and life before and during the lockdown. More specifically, the main investigated topics are: a) health and health behavior: general health before and after the COVID-19 outbreak, health conditions that may impact the pathway of COVID-19, safety measures taken (e.g. social distancing, wearing a mask, using disinfection fluids), b) infections and healthcare: COVID-19 related symptoms, COVID-19 testing and hospitalization of the respondent and of family and friends, forgone medical treatment, satisfaction with treatments; c) changes in work and economic situation: unemployment, business closures, working from home, safety measures at the workplace, changes in working hours and income, financial support, financial hardship; d) social networks: changes in personal contacts with family and friends, help given and received, personal care given and received, and volunteering.

In the next Section, the characteristics of ego-centered networks of Italian older people are shown, underling the main changes before and after the COVID-19 pandemic first wave. Results on a set of activities usually done outside home are reported as well by network types.

3 Ego-centered networks of Italian older people before and after the outbreak of Corona

The two target groups we focused on represented the 75% and the 73% of the survey population aged 65 and over, respectively in the FSS 2016 and in the SHARE Corona 1. As shown in Table 1, the two surveys reported very similar socio-demographic characteristics, except for health condition and the number of children.

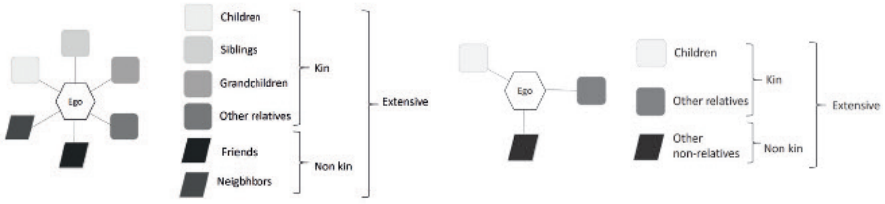
Disregarding contacts with grandparents (in FSS) and with parents (both in FSS and in SHARE Corona 1) because of the age of the target respondents, the potential alters in the ego–network of f-t-f contacts for each elder (ego) are represented by a maximum of six different alter's categories (alter roles) in FSS (children, siblings, grandchildren, relatives, neighbors, and friends) and of three categories in SHARE Corona 1 (children, other relatives, other non-relatives like neighbors, friends, or colleagues).

To facilitate comparison, three network types can be identified (Figure 1): *Kin* composed only of alters belonging to ego's kinship (at least one alter in children, siblings, grandchildren, and relatives categories in FSS or children, and other relatives in SHARE Corona 1); *Non–kin* composed only of neighbors and/or friends, and/or colleagues in SHARE Corona 1; *Extensive* composed of at least one alter in each group (*Kin* and *Non–kin*).

	FSS 2016		SHARE Corona 1	
	Single (n=1851)	In Couple (n=3234)	Single (n=275)	In Couple (n=836)
Gender				
Female	71.4	48.0	78.2	47.1
Male	28.6	52.0	21.8	52.9
Age				
65-74	33.3	55.3	33.8	52.6
75+	66.7	44.7	66.1	47.4
Health				
Good	30.5	39.4	34.2	49.7
Fair	42.5	42.1	49.1	39.5
Bad	27.0	18.5	16.7	10.8
Children				
Yes	74.2	93.4	79.3	87.3
No	25.8	6.6	10.7	12.7
Employment status				
Employed	2.3	3.2	2.1	1.5
Not employed	97.7	96.8	97.9	98.5

Table 1: Socio-demographic characteristics, FSS 2016 and SHARE COVID-19 (%, unweighted data)

Fig. 1: Network types in FSS 2016 and SHARE Corona 1 by alter roles



(a) FSS 2016 **(b)** SHARE Corona 1

Being embedded in interpersonal relations may derive a range of benefits [11], which could have been ever more important in the outbreak of Corona, when the risk of feeling lonely, sad or depressed can be higher than in other contingencies. The frequency of f-t-f contacts is then a fundamental prerequisite for considering an alter in the potential ego-network. Therefore, in both surveys an alter is included if he/she has f-t-f contacts with the ego at least once in a week.

Nevertheless, the residential proximity among people involved in the network could be another peculiarity which reinforce the value of the relational resources in a pandemic time. For this reason, we consider here the "accustomed-to-reach" network defined with FSS 2016 data, accounting both the habit to meet in person and the residence in the same municipality.

SHARE Corona 1 also collected data on testing and hospitalization of respondents and other people, from and/or outside their kinship, close to them. Only 8%

of singles and 8.5% of partners in a couple declared to have had anyone in their acquaintances with a positive result at a Corona virus test; 5.5% of singles and 6.5% of partners knew someone who has been hospitalized due to the Corona infection, while 4.6% and 6.9% respectively, knew someone who died. None of respondents have been affected or hospitalized by or due the Corona illness at the time of interviews.

Looking at the ego-networks, some changes between pre-pandemic and post-pandemic period emerged (Table 2).

	FSS 2016		SHARE Corona 1			
	accustomed-to-reach		f-t-f contacts		virtual contacts	
	Single	In Couple	Single	In Couple	Single	In Couple
No Alters	13.0	12.8	34.9	44.6	8.4	5.02
Extensive	36.6	39.4	14.3	9.9	53.8	55.6
Kin	30.8	30.9	44.5	40.4	37.4	38.7
Non-Kin	19.7	16.9	6.3	5.0	0.4	0.7

Table 2: Distribution of network types in FSS 2016 and SHARE Corona 1

Since the outbreak of Corona in 2020, around the 35% of the singles and 45% of the partners in a couple had no f-t-f contacts with others (*No Alters* network). For the "accustomed-to-reach" network the percentage were around the 13% for the same two groups. A *Kin* network type is even more widespread involving more than 40% of the elderly. On the contrary, *Non-kin* network as well *Extensive* network types appeared strongly reduced. These findings seem to be the results of two opposite behaviors related to the lockdown restrictions. First, frequent f-t-f contacts (at least about once a week for 50% of singles and for 47.4% of partners in a couple) were entertained with children that, under specific conditions, were allowed to visit their parents in case of need. Second, ties with non-kin alters and other relatives (around 80% of contacts with these categories happened less than a week or never) have been probably forcefully reduced.

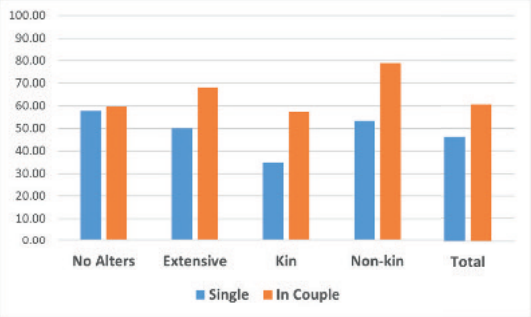
Ties with children have been strengthened also through virtual contacts with daily calls or digital means for more than one single out of two and for two partners in a couple out of three. This result is in line with studies referring to a pre-pandemic period, which highlighted the positive association between digital contacts and traditional forms of contact [7]. Moreover, in the ego-networks of virtual contacts appreciable lower percentages of *No Alters* can be noted.

As known, SARS-CoV-2 virus containment policies included measures aimed at reducing physical contacts limiting, in particular, activities not strictly necessary to people survival.

In general, 46.2% of singles and 60.6% of elderly in a couple have left their home in the three months preceding the interview (see Figure 2). Singles in a *Kin* network type stayed at home much more than singles in other network types.

Nevertheless some elderly left home, activities that imply the possibility to meet other people were drastically limited. For the group of elderly –singles or partners

Fig. 2: Frequency of elderly who left home since the outbreak of Corona by network type (source: SHARE Corona 1 survey, %)



in a couple– who declared to have left their house (see Figure 3) 86.4% of singles and 81.7% of partner in a couple declared to go shopping less often or not any more, and percentages grow up to 88.2 and 85.4, respectively, when referred to go out for a walk. Also the habits of meeting other acquaintances was very limited because of the fear of being exposed to COVID-19: more than 75% of singles and 72% of elderly in a couple did not meet more than 5 not-cohabiting people anymore, and 57.3% and 56%, respectively, did not go visit to other relatives any more.

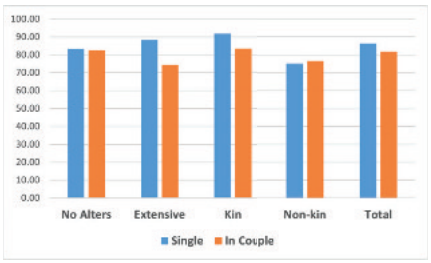
Considering the type of network in which older people were embedded some differences can be noted.

For singles, an impressive 91.9% and 94.1% in in a *Kin* network reduced or did not any more shopping or walk activities, respectively, while percentages are slightly less for other network types. For partners in a couple, shopping activities were reduced or ceased especially for those in a in a *Kin* network (83.5%), while going walk was reduced or ceased especially for those with a *No Alters* network (89.1%). As expected, activities like "Meeting with more than 5 people from outside your household" and "Visiting other family members" were not carried out any more, both for singles and partners in a couple, especially for those embedded in a *No Alters* network, while percentages decrease for other network typologies, down to the lowest value for the *Extensive* network. Surprisingly, visit other relatives showed a lower reduction with respect to the other kind of activities.

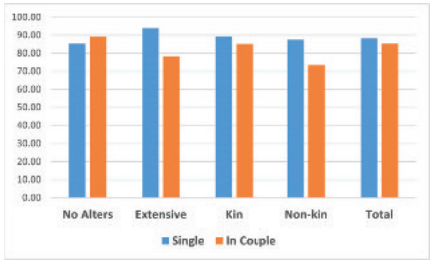
4 Concluding remarks

From the two different data sources clear changes in the social network characteristics of older Italian people between a pre-pandemic and a post-pandemic period emerged. A first evidence is related to the increase of people defined as *No Alters* because they declared to not having had any f-t-f contacts. This is observed especially among couples, where the presence of the partner is ensured even during the tightest

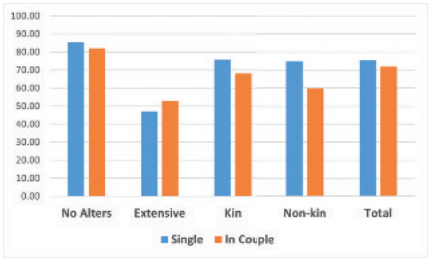
Fig. 3: Frequency –not any more/less often– of activities done since the outbreak of Corona as compared to before the outbreak for elderly who left home by network type (source: SHARE Corona 1 survey, %)



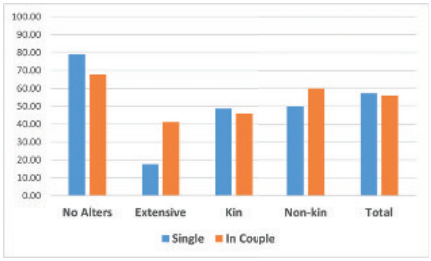
(a) Going Shopping



(b) Going Walk



(c) Meet more than 5 people



(d) Visit other relatives

lockdown. The opportunity to change the physical relation into a digital one seems then to lower the percentage of older people with a *No Alters* type network. The relationship between the egos and their children has been maintained more often than the relationship with other kin, while the relationships with friends and neighbors appear as the most compromised.

In the absence of other Italian large scale updated surveys, SHARE Corona 1 gives the opportunity to depict the social network characteristics of older people in Italy, with a special focus on the pandemic "non-intended" consequences as observed in a close period after the outbreak of Corona virus. In particular, in the first wave of the outbreak, doing activities like going shopping or going walk were carried out less often by the majority of elderly, either single or partner in a couple, as well as activities involving meeting other people (both form and outside the kinship). Nevertheless, being in a *Kin* network led older people to not leave their home for essential activities, like shopping, probably because the alters supported them in satisfying basic needs. SHARE COVID-19 project includes also the SHARE Corona 2 survey carried out in the mid of 2021, that will allow to verify if the reported changes are temporary or not. A deeper analysis of the two SHARE Corona waves will allow to better detail the association between ego-network characteristics and health related conditions (lonely, depression), and the kind of activities carried out during the lockdown.

References

1. Amati, V., Rivellini, G., Zaccarin, S.: Potential and effective support networks of young Italian adults. Social Indicators Research, 122(3), 807-831 (2015).
2. Amati, V., Meggiolaro, S., Rivellini, G., Zaccarin, S.: Relational Resources of Individuals Living in Couple: Evidence from an Italian Survey. Social Indicators Research, 134(2), 547-590 (2017).
3. Arpino, B., Mair, C., Quashie, N., Antczak, N.: "Loneliness Before and During the COVID-19 Pandemic: Are Unpartnered and Childless Older Adults at Higher Risk?." SocArXiv. September 28 (2021) doi:10.31235/osf.io/6v7bg.
4. Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A.,Quattrociocchi, W., Pammolli, F. Economic and social consequences of human mobility restrictionsunder COVID-19. Proceedings of the National Academy of Sciences, 117(27), 15530-15535 (2020).
5. Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber. Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology. DOI: 10.1093/ije/dyt088 (2013).
6. Börsch-Supan, A.: Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. COVID-19 Survey 1. Release version: 8.0.0. SHARE-ERIC. Data set (2022). DOI: 10.6103/SHARE.w8ca.800.
7. Danielsbacka, M., Tammisalo, K., Tanskanen, A.O.: Digital and traditional communication with kin: displacement or reinforcement?, Journal of Family Studies, (2022)
8. Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y., Mills, M. C. Demographic science aids in understanding the spread and fatality rates of Covid-19. Proceedings of the National Academy of Sciences, 117(18), 9696–9698 (2020).
9. Esteve, A., Permanyer, I., Boertien, D., Vaupel, J. W. National age and coresidence patterns shape COVID-19 vulnerability. Proceedings of the National Academy of Sciences 117.28: 16118-16120 (2020).
10. Furfaro, E., Rivellini, G., Pelle, E., Zaccarin, S.: Constructing personal networks in light of COVID-19 containment measures. Genus 77, 17 (2021).
11. Litwin, H. (Ed.): The social networks of older people: A cross-national analysis. Praeger Publishers (1996).
12. Litwin, H., Levinsky, M.: Network-exposure severity and self-protective behaviors: The case of COVID-19. Innovation in Aging, 5(2), igab015 (2021).
13. Luppi, F., Arpino, B., Rosina, A. The impact of covid-19 on fertility plans in Italy, Germany, France, Spain, and the United Kingdom. Demographic Research, 43, 1399–1412 (2020).
14. Pelle, E., Zaccarin, S., Furfaro, E., Rivellini, G.: Support provided by elderly in Italy: a hierarchical analysis of ego networks controlling for alter-overlapping. Stat Methods Appl 31, 133–158 (2022).
15. SHARE-ERIC: Results of the 1st SHARE Corona Survey; Project SHARE-COVID19 (Project Number 101015924, Report No. 1, March 2021). Munich: SHARE-ERIC (2021). DOI: 10.17617/2.3356927.
16. Scheel-Hincke, L.L., Ahrenfeldt, L.J., Andersen-Ranberg, K.: Sex differences in activity and health changes following COVID-19 in Europe—results from the SHARE COVID-19 survey. European journal of public health, 31(6), 1281–1284 (2021).
17. Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., Schuller, K., Stuck, S., Wagner, M., Börsch-Supan, A.: Collecting survey data among the 50+ population during the COVID-19 outbreak: The Survey of Health, Ageing and Retirement in Europe (SHARE). Survey Research Methods 14(2), 217–221 (2020).
18. Yordanov, V.: Covid-19 pandemic: a study on the relationship between social distancing and mental health status among people aged 50 and older in europe. Revista Inclusiones, 113–139 (2021).

Internet use and contacts with children among older Europeans

Uso di internet e contatti con i figli tra gli anziani europei

Bruno Arpino

Abstract Contacts with children represent one of the most important sources of support for older individuals. By using panel data from the Survey of Health, Ageing and Retirement in Europe (SHARE), I investigate to what extent use of the Internet is related to the frequency of contacts with children, especially those living at a distance. Asymmetric fixed effects panel models and ordered logistic regression models show a significant positive relationship between using the Internet and frequency of contacts both before and during the COVID-19 pandemic. These results point to the importance of digital contacts to strengthen intergenerational relationships and are especially relevant in the context of the pandemic and of increasing geographical mobility.

Abstract *I contatti con i figli rappresentano una delle più importanti forme di supporto per gli anziani. Utilizzando i dati del panel dell'indagine Survey of Health, Ageing and Retirement in Europe (SHARE) ho analizzato la relazione tra l'uso di Internet e la frequenza dei contatti con i figli, in particolare quelli che vivono a distanza. Modelli a effetti fissi asimmetrici e i modelli di regressione logistica ordinale mostrano una relazione positiva significativa tra l'utilizzo di Internet e la frequenza dei contatti sia prima che durante la pandemia di COVID-19. Questi risultati sottolineano l'importanza dei contatti digitali per rafforzare le relazioni intergenerazionali e sono particolarmente rilevanti nel contesto della pandemia e dei una crescente mobilità geografica.*

Key words: Internet use, intergenerational contacts, intergenerational relationships, older adults, COVID-19

[1] Bruno Arpino, Department of Statistics, Computer Science, Applications
University of Florence; email: bruno.arpino@unifi.it

1 Introduction

The demographic forces behind population ageing (i.e., fertility and mortality) have a strong impact on intergenerational relations: people are living longer with smaller family networks than in the past. For example, childlessness has increased and number of children has reduced over different cohorts. Intergenerational relations have been largely demonstrated to be crucial for older people physical and mental health (Dykstra 2007; Carr & Utz 2020) and shrinking kin networks imply that it will be more and more important to maintain social contacts with the available kin. It is yet to be understood to what extent the use of digital technologies helps older people in this respect.

Despite the process of digitalisation being not new, only recently we reached a stage that creates unprecedented opportunities and challenges for social relations. The World Wide Web became accessible to the public only in 1994 and Social Networking Sites (SNS) have been introduced only in the last 20 years (e.g., Facebook in 2004), offering unprecedented new forms of connecting with known and unknown persons, living wherever. Also, only recently the internet has become used by a large part of the older population in developed countries, although internet use still strongly varies with age ("age digital divide"), ranging in the EU-28 from above 95% for individuals younger than 44 to 79% and 61% in the age ranges 55-64 and 65-74, respectively. As expected, internet use is even lower among people aged 75+, and reaches a minimum of 5% among people aged 90+. So, the myth of older people excluded from the broadband society has to be refused, but a heterogeneous access and use of digital technologies persists.

In this paper I focus on the role of internet use for contacts with children among older Europeans. Intergenerational contacts between older individuals and their children constitute an important part of older people's social contacts (Tomassini et al. 2004). The Internet offers additional possibilities for maintaining intergenerational contacts (Peng et al. 2018), but it can also subtract time to offline relationships. Digital contacts can be particularly important for older people whose children live far away and in the context of a pandemic.

A few studies analysed whether or not online communication with relatives displace or reinforce more traditional forms of contact (e.g., face-to-face interactions), with mixed findings (e.g., Arpino et al. 2021a; Danielsbacka et al. 2021). Other studies focussed more explicitly on the role of Internet use on off-line social contacts. Kraut and colleagues (1998) reported a negative effect of internet use on the frequency of off-line interpersonal relations (the so-called ''Internet paradox''). Later evidence for positive, negative or null association between internet use and offline relations was found. Most studies used small non-representative samples and were often not focused on older adults and intergenerational relationships.

In this paper, I use large-scale nationally representative longitudinal data from SHARE and implement two sets of analyses. The main analyses use data from the pre-COVID period and attempt at estimating the effect of Internet use on frequency of contacts with children distinguishing those living geographically close from those living far away. By using asymmetric fixed effects, I separate the effect of starting from the effect of stopping using the Internet. In the second part of the analyses, I implement a sort of "case-study" focussed on the COVID-19 period. In this case, I analyse whether Internet use before the pandemic is associated with non-physical contacts (i.e. not face-to-face) with children.

2 Data and methods

I use data from the Survey of Health, Ageing and Retirement in Europe (SHARE), a panel survey representative of the non-institutionalized population aged 50+ administered every 2 years since 2004 in several European countries and Israel (Börsch-Supan et al. 2013). The main analyses use data from waves 5 to 8 because the Internet use variable is not available in previous waves. Wave 8 is the last pre-COVID wave; it started in October 2019 but was suspended in all countries in March 2020 due to the COVID-19 outbreak. All pre-COVID waves are based on computer-assisted personal interviewing (CAPI) (Börsch-Supan et al. 2013). A special dataset was added to wave 8. This survey has been administered with CATI (computer assisted telephone interviewing) between June and August 2020 to collect information on individuals' behaviors and conditions during the pandemic (SHARE Corona Survey 1; SCS1 hereinafter) (Börsch-Supan 2022). This data provides information collected after the onset of the pandemic. SHARE offer several advantages for the research question I address in this study. First, it is a longitudinal panel survey. Second, although measures of internet use were not collected in all waves and are not detailed, SHARE regular waves offers information on contacts with each child separately, together with additional relevant data, e.g. on geographical distance to each child.

In the main analyses, the outcome variable is based on a question about contacts with children asked separately for each one of the respondents' child: *During the past twelve months, how often this you have contact with CHILD either in person, by phone, mail, email or other electronic means?* Respondents report the approximate frequency among the following options: *Daily, Several times a week, About once a week, About every two weeks, About once a month, Less than once a month, Never*. I combined information about the contact frequency to each child into one single numerical variable. First, I attributed an equivalent number of contact days to each response category. For example, I used 365 for "Daily" and 54 for "About once a week". Then, I summed all resulting scores. I created two different outcome variables distinguishing children who live close (< 25 km) from the others by summing up the equivalised contact frequency for the children respecting the geographical criterion only. The information in the regular SHARE waves allows to

account for frequency of contacts to each child but it does not separate face-to-face from other forms of contact.

In the COVID-19 "case study" I use data from SCS1. In this case, the questionnaire distinguishing between face-to-face and other types of contact. Given that during the early phases of the pandemic face-to-face contacts were forbidden or strongly discouraged and that contacts at a distance have been important during the pandemic (Arpino et al. b,c) I focus on this type of contact. The outcome variable is based on the following question: *Since the outbreak of Corona, how often did you have contact by phone, email or any other electronic means with the following people from outside your home? (Was it daily, several times a week, about once a week, less often, or never?)*. I focus on the answers that refer to children, where differently from the pre-COVID information, all of them are considered together. I use the variable as an ordinal categorical variable. Differently from the pre-COVID analyses, here I cannot match contact frequency to each child's geographical proximity. So, I run analyses on the whole sample and distinguishing respondents with all children living close and respondents who have only children living far away.

The explanatory variable is a simple dichotomous item for Internet use: *During the past 7 days, have you used the Internet, for e-mailing, searching for information, making purchases, or for any other purpose at least once?*. I controls for several factors that may confound the relationship of interest: Age (5-year groups), gender, education, partnership status, working status, health (self-rated health, diagnosed illnesses, limitations with activities), living in a rural area, waves, country of residence.

The main analyses use longitudinal data from waves 5 to 8 and implement standard linear fixed-effects models and asymmetric linear fixed-effects models (ALFE). Fixed-effects models remove time invariant observed and unobserved confounders. Thus, observed time-invariant controls (e.g. gender and country of residence) are included in the COVID-19 analyses only. ALFE models allow distinguishing the effect of starting from that of stopping using the Internet. The COVID-19 analyses are based on ordered logistic regression models. I implemented a number of robustness checks, e.g. by dropping individuals with IADL (instrumental activities of daily living limitations) or individuals aged 85 and more for which Internet use probabilities are very low.

3 Results

I start presenting results from the main analyses based on pre-COVID waves. Table 1 reports estimates for the explanatory variable (Internet use) from standard linear fixed-effects model with all controls included. Two models are implemented that differ for the outcome variable that consider only geographically "close" children or

the others ("not close"). Results show that Internet use is positively associated with frequency of contact with children. However, the effect is statistically significant only when geographically distant children are considered (and in this case also the magnitude of the coefficient is considerably higher).

Table 1: Internet use and contacts with children before the COVID-19 pandemic. Estimated coefficients from standard linear fixed-effects models (standard errors in parentheses)

Explanatory variable	***Contacts with children***		
	close	***not close***	
Internet use	0.57	9.77	**
	(0.83)	(0.03)	

Notes: all controls are included. *** p<0.01; ** p<0.05; * p<0.1.

Table 2 reports results from the ALFE model. Interestingly, the finding highlights that the statistically significant association that emerged from the standard linear fixed-effect model above between Internet use and contact frequency was driven by changes in Internet use of those people who started using the Internet. In other words, starting using the Internet was associated with an increased frequency of contact with children far away, while stopping using the Internet was not associated with a statistically significant reduction in contact frequency.

Table 2: Starting and stopping using the Internet and contacts with children before the COVID-19 pandemic. Estimated coefficients from asymmetric linear fixed-effects models (standard errors in parentheses)

Explanatory variables	***Contacts with children***		
	close	***not close***	
Started using the Internet	0.45	14.59	***
	(0.181)	(0.002)	
Stopped using the Internet	-1.64	1.33	
	(0.361)	(0.36)	

Notes: all controls are included. *** p<0.01; ** p<0.05; * p<0.1.

I now turn to the analyses using data for the COVID-19 period. Table 3 reports the odds ratio for Internet use for three models estimated: 1) on the whole sample; 2) on the sub-sample of respondents with only "close" children; 3) the sub-sample of respondents with only "far" children. I found a positive association between internet use and frequency of non-physical contacts during the pandemic. This is slightly stronger when considering contacts with children living geographically close.

Table 3: Internet use and non-physical contacts during the COVID-19 pandemic. Estimated odds ratios from ordered logit models (standard errors in parentheses)

Explanatory variable	*Contacts with children*					
	All respondents		*Only Rs with all children close*		*Only Rs with all children far*	
Internet use	1.34	***	1.24	***	1.32	***
	(0.000)		(0.000)		(0.000)	

Notes: all controls are included. *** p<0.01; ** p<0.05; * p<0.1.

4 Concluding remarks

I found a significant positive relationship between starting using the Internet at older ages and increased frequency of contacts with children living far away before the onset of the COVID-19 pandemic. I also found that those older adults who were using the Internet before the pandemic were more likely to have non-physical contacts with children during the pandemic as compared to their counterparts who did not use the Internet.

These results point to the importance of digital tools to strengthen intergenerational relationships (i.e., "digital solidarity", Peng et al. 2018). Findings are especially relevant in the context of the COVID-19 pandemic and of increased geographical mobility of the younger generations which makes increasingly likely for older people to have at least some children living at a distance. The findings are important because previous studies have demonstrated that non-physical contacts, which are favoured by the use of digital technologies, positively influence older adults' mental health (Arpino et al. 2021c).

References

1. Arpino, B., Meli, E., Pasqualini, M., Tomassini, C., and Cisotto, E. (2021a). WhatsApp Grandpa? Determinants of grandparents-grandchildren digital contacts in Italy. SocArXiv. https://doi.org/10.31235/osf.io/yrvfz.
2. Arpino, B., Pasqualini, M., and Bordone, V. (2021b) Physically distant but socially close? Changes in non-physical intergenerational contacts at the onset of the COVID-19 pandemic among older people in France, Italy and Spain. European Journal of Ageing, 18, 185–194.
3. Arpino, B., Pasqualini, M., Bordone, V., and Solé-Auró, A. (2021c) Older people's non-physical contacts and depression during the COVID-19 lockdown. The Gerontologist, 61(2), 176–186.
4. Börsch-Supan, A. (2022). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. COVID-19 Survey 1. Release version: 8.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w8ca.800.
5. Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., ... & Zuber, S. (2013). Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). International journal of epidemiology, 42(4), 992-1001.

6. Carr, D., & Utz, R. L. (2020). Families in later life: A decade in review. Journal of Marriage and Family, 82(1), 346-363.
7. Danielsbacka, M., Tammisalo, K., & Tanskanen, A. O. (2022). Digital and traditional communication with kin: displacement or reinforcement?. Journal of Family Studies, 1-22.
8. Dykstra P (2007) Aging and social support. In: Ritzer G (ed) The Blackwell encyclopedia of sociology. Blackwell, Oxford, pp 88–93
9. Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being?. American psychologist, 53(9), 1017.
10. Peng S, Silverstein M, Suitor JJ, Gilligan M, Hwang W, Nam S, Routh B (2018) Use of communication technology to maintain intergenerational contact: toward an understanding of 'digital solidarity. Connect Families, pp 159–180.
11. Tomassini, C., Kalogirou, S., Grundy, E., Fokkema, T., Martikainen, P., Van Groenou, M. B., & Karisto, A. (2004). Contacts between elderly parents and their children in four European countries: current patterns and future prospects. European Journal of Ageing, 1(1), 54-63.

A time-based comparative approach to study the changing demography of grandparenthood in Italy

Un approccio temporale per studiare il cambiamento della demografia dei nonni in Italia

Cisotto Elisa, Meli Eleonora, Cavrini Giulia

Abstract This article analyses the most significant changes in the demography of grandparenthood in Italy over the last two decades, using data from the Family and Social Subjects Survey conducted by ISTAT (1998, 2003, 2009 and 2016). The median age at which half of the population over 35 is composed of grandparents has moved forward by at least five years during the two decades observed. The median age at grandparenthood increased by three years for both men and women, and this difference is more significant than that observed for the age at parenthood and equal to the advantage gained in life expectancy at age 60. Thus, despite the increase in life expectancy, due to the postponement of grandparenthood, the life span shared by grandparents and grandchildren has remained stable[2].

Abstract *In questo articolo vengono analizzati i cambiamenti più significativi nella demografia dei nonni in Italia negli ultimi due decenni, utilizzando i dati dell'Indagine Famiglia e Soggetti Sociali condotta dall'ISTAT (1998, 2003, 2009 e 2016). L'età mediana in cui la metà della popolazione sopra i 35 anni è composta da nonni, si è spostata in avanti di almeno 5 anni durante i due decenni osservati. L'età mediana alla nascita del primo nipote è aumentata di tre anni, sia per gli uomini che per le donne, e questa differenza è maggiore di quella osservata per l'età alla genitorialità e pari al vantaggio guadagnato in termini di aspettativa di vita a 60 anni. Nonostante l'aumento dell'aspettativa di vita, a causa del rinvio della nonnità, il periodo di vita condiviso da nonni e nipoti è quindi rimasto stabile.*

[1] Cisotto Elisa, Free University of Bozen-Bolzano; email: elisa.cisotto@unibz.it
Meli Eleonora, Italian National Institute of Statistics (ISTAT); email: elmeli@istat.it
Cavrini Giulia, Free University of Bozen-Bolzano; email: giulia.cavrini@unibz.it

[2] The extended version of the paper is published open access on Genus:
https://doi.org/10.1186/s41118-022-00153-x

Key words: Grandparenthood, Grandparents, Italy, Demography, Age at grandparenthood, Trends.

1 Introduction

During the twentieth century, all Western countries experienced a twofold demographic revolution, consisting, on the one hand, of an unprecedented increase in life expectancy and, on the other, a sharp decline in fertility. These demographic changes have led to the rapid ageing of the population, affecting the amount of time individuals spend in different family roles, marking the beginning and duration of intergenerational relationships. Among those, one of the most affected is grandparenthood.

This article examines the changes in grandparenthood in Italy over an 18-year period considering grandparents' demographic and socio-economic characteristics, the number of grandchildren, the prevalence of grandparenthood, and the context of life.

At the background level, Italy is a country that, since the second half of the 20th century, has experienced a more marked decline in mortality and fertility than other developed countries: life expectancy at birth has increased of 17 and 18 years from 1950 to1953 to 2019, while the total fertility rate (TFR) has decreased by 1 child per woman from 1952 to 2019. While increasing life expectancy could result in more life years spent as grandparents and a greater likelihood of sharing this role with other living grandparents, downward trends in fertility and time lags in parenting are likely to delay the onset of grandparenthood, shortening its duration, and increasing the share of people who never become grandparents.

Its family-centred social context and limited public childcare facilities make Italy an optimal country to study the transition to grandparenthood over time (Di Gessa et al., 2016; Glaser & Hank, 2018; Zamberletti et al., 2018). Grandparents in Italy are a crucial source of family support and play a central role in providing informal care for children. Moreover, the increasing participation of women in the labour market makes childcare by grandparents increasingly important (Arpino et al., 2014). In fact, in 2016, about 39% of grandchildren aged 0-13 were cared for by grandparents when their parents were at work, revealing an increase of about 10% over the last two decades (Pasqualini et al., 2021).

However, the demography of Italian grandparenthood has only been partially studied. Little is known about how the grandparental phase of life is changing in Italy, nor the positive or negative impact these changes may have on the Italian family and society. This paper analyses data from the Italian National Institute of Statistics (ISTAT) surveys on Italian households conducted in 1998, 2003, 2009 and 2016. After a brief introduction to the background, we consider grandparenthood as a critical event in the transition to old age and outline the possible changes that can be expected based on the demographic dynamics of the last two decades. The following sections introduce the data and methods and present the results.

2 Data and Methods

Data is drawn from two sources, the four waves of the Survey on Families, *Famiglia e Soggetti Sociali* (Family and Social Subjects - FSS), carried out by the Italian Institute of Statistics (ISTAT) in 1998, 2003, 2009 and 2016, and the Italian life tables by sex released by ISTAT for 1998 and 2016.

2.1 Study population and statistical methodology

In the FSS survey, grandparents are detected among people aged 35 and over. All interviewed were asked whether they were grandparents at the time of the survey (considering adopted, foster of biological grandchildren), and the age of up to three grandchildren.

Based on this sample, we first follow Margolis' approach (2016) to categorize the population into grandparents, non-grandparents because of childlessness, and non-grandparents because of children's childlessness, to estimate the relative contribution of each sub-population to the overall changing prevalence of grandparenthood over time.

In addition, to determine the timing of grandparenthood, the analytical sample is restricted to all parents aged 60 with at least one living child in 1998 and 2016, the two extreme years of those available. According to this selection, we adhere to the Leopold and Skopek's approach (2015a) and apply survival analysis to estimate: (a) the median age at grandparenthood and other four adult-life transitions (i.e., parenthood, empty-nest, end of active parenthood, retirement or inactivity), (b) the probability of becoming grandparents at different ages, (c) the expected length of grandparenthood. Individuals' life histories are reconstructed using the events sequences given by respondents to set the survival time axis to start at birth, and to end at the age at which each event occurs (Table 1). As the age of grandchildren can only be found for a maximum of three grandchildren per respondent, to estimate the birth of the older grandchild for those grandparents with four or more grandchildren (23% of grandparents in 1998, and 16% in 2016), we implement and refine the strategy suggested by Di Gessa et al. (2020). However, we proved the validity of the approach with a set of robustness tests.

All models are implemented by sex, in line with the different timing of major demographic events. All analysis and descriptive tables have been performed using normalized weights based on the population's marginal distribution coefficients provided by ISTAT. For ease of exposition, we present the results for the two extreme years of the time series, 1998 and 2016.

Table 1: Measures for the demography of grandparenthood

Transition	*Definition*
Grandparenthood	Age at birth of the first grandchild
Parenthood	Age at birth of the first child
End of active parenthood	Age at which the youngest child turns 16
Empty nest	Age at which all children live the parental home, or single parents live the household
Retirement or inactivity	Age at transition to retirement or economic inactivity for those ever in paid work in the past. Inactivity is considered to occur if the person becomes permanently sick or disable, homemaker or unemployed and no longer seeks work opportunities

3 Results

Overall, grandparents account for around 33% of our samples in 1998 (N=9.518), 2003 (N=10.274), 2009 (N=9.643), and fell to 31.5% (N=6.222) in 2016. In absolute terms, grandparents in Italy were around 10.9 million in 1998, and they turned to 12.3 million in 2016.

3.1 Prevalence of grandparenthood over time

Figure 1 shows the population prevalence among grandparents, non-grandparents because childless, and non-grandparents because of their children's childlessness. For both sexes, grandparents' prevalence is reduced over time, so that, on average, the age at which half population is made up of grandparents moved forward by at least five years from 1998 to 2016. Especially in middle age, we observe that parenthood is significantly reduced at any age (see also Table 2).

Conversely, the prevalence of childless and children-childless adults significantly increases over time, contributing to the lower ratio of grandparents over the two observed decades. Data in Table 2 break down numerically the reduction in the population prevalence of grandparents by exploring whether it is due to the general increase in childlessness, or to the increase in children's childlessness. To give an insight, between 1998 and 2016, increased grandmothers' childlessness explains the 73% of the total decline in grandmotherhood prevalence at age 50 to 54 (-9%). At age 60 to 64, figures are reversed, and the declined prevalence of grandmotherhood between the two years (-10%) is mainly driven by increased children's childlessness (88%). For men, the causal breakdown confirms a comparable trend to that of women, while statistically significant differences in the prevalence of grandfatherhood over time can be noted until older ages (i.e., 66 to 74).

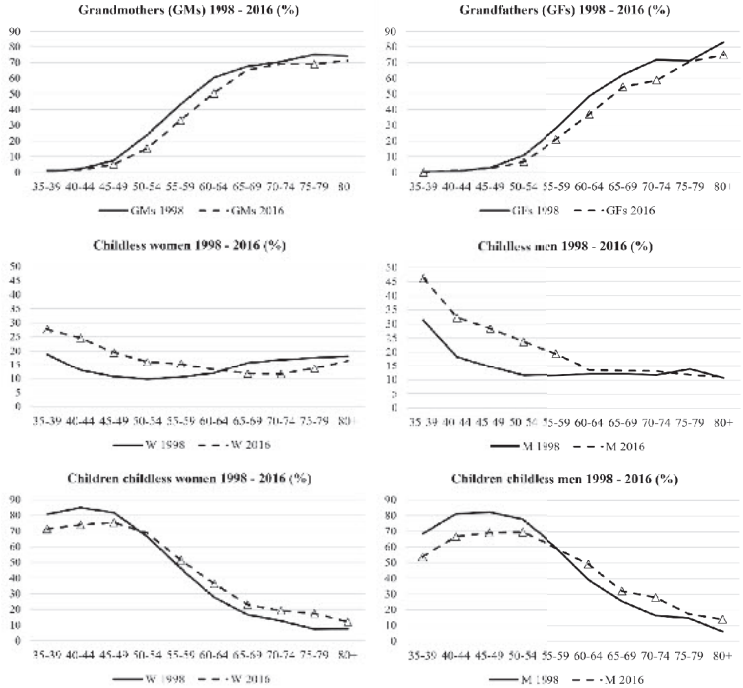


Figure 1: Prevalence of grandparents, childless and children-childless population by age, sex and year[1].

Table 2: Reasons for declining grandparenthood by sex and age class[2].

	Grandparents gap in % (2016-1998)		***Decline prevalence in grandparenthood due to:***		
			Increased childlessness	***Increased children childless***	***Total***
Grandfathers					
50-54	-4.2	*	293%	-193%	100%
55-59	-7.0	*	108%	-8%	100%
60-64	-11.6	*	13%	87%	100%
65-69	-7.7	*	15%	85%	100%
70-74	-13.0	*	11%	89%	100%
Grandmothers					
50-54	-8.6	*	73%	27%	100%
55-59	-10.2	*	48%	52%	100%
60-64	-9.9	*	12%	88%	100%
65-69	-2.4		-167%	267%	100%
70-74	-1.3		-395%	495%	100%

1 Δ = p< .05 test for differences 1998 – 2016

2 * = p< .05 test for differences 1998 – 2016

3.2 Timing of grandparenthood

Figure 2 reports and compares the median ages at different life-events, estimated by survival analysis for mothers and fathers in 1998 and 2016. Overall, the (median) age at which the first grandchild is born shifts forward by three years in the given time. Men become grandparents later than women, still, the relative change in the median age is uniform: from 54 to 57 for mothers, from 59 to 62 for fathers. The observation of estimated probabilities of being grandparents at different ages (Figure 3) clearly shows the magnitude of grandparenthood postponement. Less than 60% of mothers and 44% of fathers had become grandparents by the age of 60 in 2016, while this is valid for more than 71% of mothers and 56% of fathers in 1998.

Findings also indicate that increases in the age at first grandchild's birth are larger than those in age at parenthood, empty nest, end of active parenthood and retirement/inactivity (Figure 2). Moreover, the intersection of grandparenthood with the other analysed life events is mostly unchanged across the two surveys. Parenthood is anticipated by one year, in line with the U-shaped trend in fertility observed among the cohorts of grandparents under study. Consistently, age at the end of active parenthood is also anticipated and always precedes grandparenthood, while the empty nest phase of life comes after it. For what concerns transition to retirement or inactivity, in 2016 the sequences show a few years of anticipation to the first grandchild's birth compared to 1998.

Finally, by taking the difference between life expectancy at age 60 and the median age at grandparenthood, we obtain the expected length of grandparenthood for men and women in the two observed years. Estimates show that, even if grandparenthood has been delayed, the years of life shared by grandparents and grandchildren do not change over time. Indeed, the gains in life expectancy (+3 years for both men and women) equal the three-years time lag detected in entry into grandparenthood.

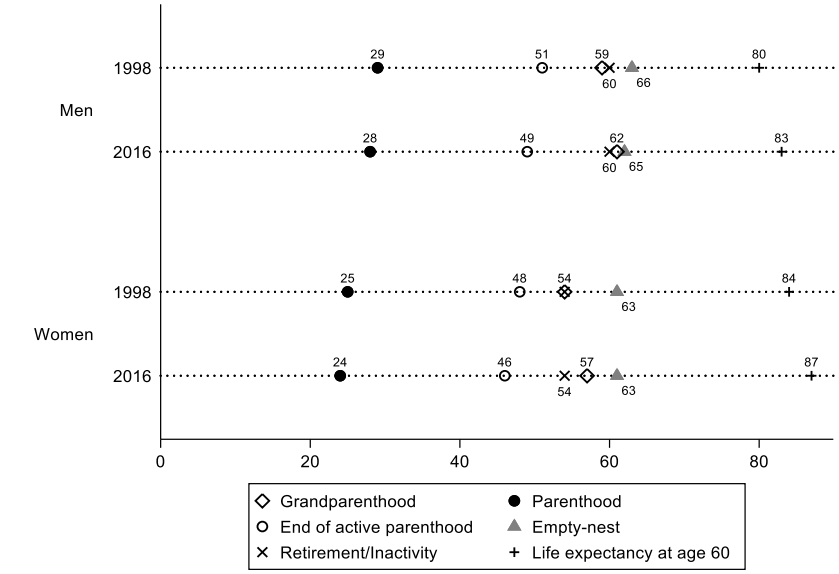


Figure 2: Timing of grandparenthood and other four life transitions, median age by sex and year.

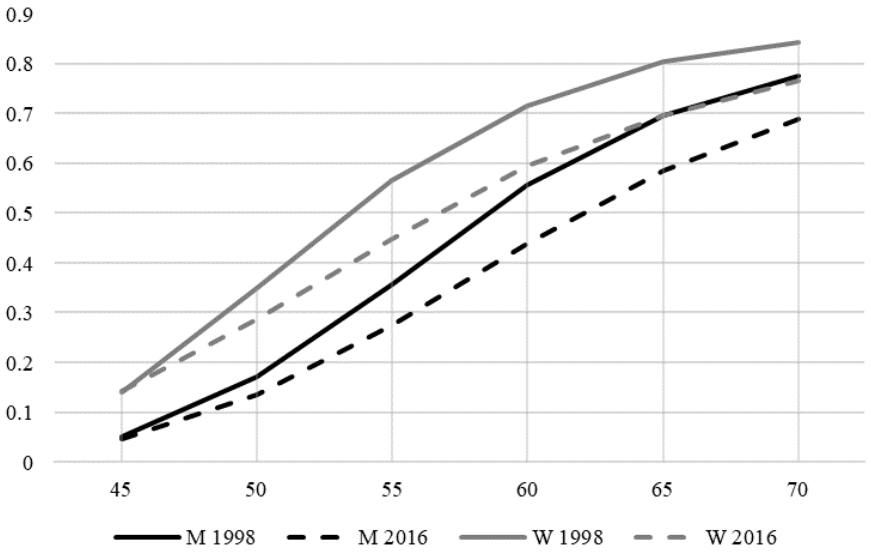


Figure 3: Probability of becoming grandparent at different ages by sex and survey year.

4 Conclusion and discussion

Exceptional increases in life expectancy and a sharp decline in fertility have led to a rapidly ageing population in Italy, involving the time-period that individuals spend in different family roles, such as grandparenthood. Besides, due to its limited welfare services, and the central role of grandparents in providing care to grandchildren, Italy is a country, which deserves particular attention in the international context. This study is the first applied to national representative survey data to assess changes in the demography of grandparenthood in Italy by using a set of measurements: grandparenthood prevalence, timing of grandparenthood in relation to other adult-life events, and estimated length of grandparenthood.

Overall, the results are consistent with previous international and national research. Grandparenthood is clearly delayed by age from 1998 to 2016 in Italy. Nevertheless, because of increasing life expectancy, the share of life potentially spent together by grandparents and grandchildren has not been reduced, but rather remained stable to 21 years for grandfathers and 30 for grandmothers. Still, on the one hand, grandparents are older today than in the past, so that the quality of intergenerational exchanges could be affected, for instance, by grandparents' worse health conditions. On the other hand, future generation of grandparents is expected to be healthier than the former, so that the expected quality of survival is a crucial research point for further studies on grandparenthood and intergenerational relations (Margolis & Wright, 2017). Accordingly, the study of the education gradient, as well as of cohort and regional dissimilarities is of prime interest (Leopold & Skopek, 2015b; Di Gessa et al. 2020).

Observing a set of life-events sequences, our study also confirms that the intersection of grandparenthood with other adult-life transitions has not changed much over the last 18 years. Yet, some considerations need to be made on retirement and inactivity. Indeed, even though minor changes can be noted for these transitions in relation to grandparenthood (for which retirement/inactivity slightly precedes the first grandchild's birth in 2016), this holds true for the current cohorts of grandparents. New retirement legislation leading to extending working life, together with rising female labour force participation, may affect the work-life transitions timing, increasing the share of grandparents still in employment when grandchildren are born or very young (Aassve et al., 2010; Arpino et al., 2014; Zanasi & Sieben, 2020).

This study draws strength from using nationally representative data of high quality and good response rates. Nevertheless, our results should be considered in view of some limitations. By survey design, it is impossible to know the precise age at grandparenthood for the entire grandparents' sample. Furthermore, our assessment of timing of grandparenthood is limited to older adults, as the very low prevalence of grandparents before the age of 60 does not allow to estimate median ages for younger cohorts. Finally, our cross-sectional period estimates offer the picture of two-point time, thus not controlling for potential selective mortality or survey attrition. Despite the mentioned limitations, our study raises awareness of the

evolution of grandparenthood in Italy in the first two decades of the twenty-first century, contributing to the benchmark for future comparisons and developments.

References

1. Aassve, A., Arpino, B., & Goisis, A.: Grandparenting and mothers' labour force participation: A comparative analysis using the Generations and Gender Survey. Demographic Research, 27(3), 53–84 (2010)
2. Arpino, B., Pronzato, C. D., & Tavares, L. P.: The effect of grandparental support on mothers' labour market participation: An instrumental variable approach. European Journal of Population/revue Européenne De Démographie, 30, 369–390 (2014)
3. Di Gessa, G., Bordone, V., & Arpino, B.: The role of fertility in the demography of grandparenthood: Evidence from Italy. Journal of Population Ageing. https://doi.org/10.1007/s12062-020-09310-6 (2020)
4. Di Gessa, G., Glaser, K., Price, D., Ribe, E., & Tinker, A.: What drives national differences in intensive grandparental childcare in Europe? Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 71(1), 141–153 (2016)
5. Glaser, K., & Hank, K.: Grandparenthood in Europe. European Journal of Ageing, 15(3), 221–223 (2018)
6. Leopold, T., & Skopek, J.: The delay of grandparenthood: A cohort comparison in East and West Germany. Journal of Marriage and Family, 77(2), 441–460 (2015a)
7. Leopold, T., & Skopek, J.: The demography of grandparenthood: An international profile. Social Forces, 94(2), 801–832 (2015b)
8. Margolis, R.: The changing demography of grandparenthood. Journal of Marriage and Family, 78(3), 610–622 (2016)
9. Margolis, R., & Wright, L.: Healthy grandparenthood: How long is it, and how has it changed? Demography, 54, 2073–2099 (2017)
10. Pasqualini, M., Di Gessa, G., & Tomassini, C.: A change is (not) gonna come: A 20-year overview of Italian grandparent–grandchild exchanges. Genus, 77(1), 33 (2021)
11. Zamberletti, J., Cavrini, G., & Tomassini, C.: Grandparents providing childcare in Italy. European Journal of Ageing, 15, 265–275 (2018)
12. Zanasi, F., & Sieben, I.: Grandmothers' transition to retirement: Evidence from Italy. Polis (italy), 34(2), 281–308 (2020)

Carry that weight: Parental separation and children's Body Mass Index from childhood to young adulthood

Separazione dei genitori e Indice di Massa Corporea dei figli dall'infanzia all'adolescenza

Marco Tosi

Abstract Drawing on data from the PSID Child Development Supplement and the Transition into Adulthood Supplement (1997-2017), I investigate whether and how parental union dissolution affects children's Body Mass Index (BMI) in the short and long run. The results from child-fixed effects linear regression models show that marital break-up is associated with increases in child BMI and an increased risk of becoming overweight/obese. The negative effect of marital break-up on children's weight status persists for at least ten years after parental separation. The findings indicate that unhealthy weight gains following parental separation are largely driven by female children, by children aged 18 or less at parental separation, and by children of low educated parents.

Abstract *Utilizzando i dati PSID su bambini e giovani adulti (1997-2017), l'autore affronta la domanda se le separazioni dei genitori influenzino l'indice di massa corporea dei figli nel breve e lungo periodo. I risultati dei modelli lineari ad effetti fissi mostrano che la separazione dei genitori è associata ad un aumento della massa corporea e del rischio di obesità. Questo effetto negativo persiste per almeno dieci anni dopo la separazione. L'aumento di peso dopo la separazione dei genitori è concentrato tra le bambine, tra i figli minorenni al momento della separazione, e tra i figli di genitori con un basso livello di istruzione.*

Key words: Body Mass Index, Obesity, Union Dissolution, Child Fixed Effects.

1 Introduction

[1] Marco Tosi, Department of Statistical Sciences, University of Padua, marco.tosi@unipd.it.

In the United States child obesity has dramatically increased over the last three decades, and approximately 18% of children and adolescents are now obese (Hales et al. 2017). Child obesity is associated with disruptive family events; and parental separation is one of the most disruptive event. The majority of previous studies find that parental union dissolution is associated with worse physical health and unhealthy BMI gains during childhood and young adulthood (Bzostek and Beck 2011; Goisis et al., 2020; Hernandez et al. 2014; Schmeer et al. 2012). Despite a bulk of studies have brought to light the association between parental union dissolution and children's obesity, longitudinal research examining the evolution of BMI from childhood to young adulthood is still scarce.

In this study, I use a long observation window (from age 5 to age 28) to distinguish between the short-term and long-term effects of parental separation on children's weight status. Additionally, the contribution of this study is to shed light on the mechanisms that may explain the BMI trajectories following parental union dissolution, as well as to examine how the consequences of parental separation are unevenly distributed across population subgroups (defined by parental education and children's sex).

2 Data and Methods

The analysis is conducted with data from the Child Development Supplement (CDS 1997, 2002, and 2007) and the Transition to Adulthood Supplement (TAS 2005, 2007, 2009, 2011, 2013, 2015 and 2017) of the Panel Study of Income Dynamics (PSID). In 1997, the CDS was introduced to collect detailed information on a random sample of children aged 0-12 who were re-interviewed in 2002 and 2007. As children grew-up and became adult, they transited from the CDS to TAS. 3,563 children born between 1988 and 1997 were eligible for the CDS-1997.

The study sample is restricted to children at least 5 years of age and young adults less than 29 years of age. The choice of the bottom age limit is driven by the fact that child BMI is measured among children aged 5 or older. I further select children whose parents remained in marital relationship or separated over time. The final sample includes 2,661 children corresponding to 12,425 child-year observations.

2.1 Variables

CDS data collect information on the weight and height of children aged 5 or over. The primary care givers (PCGs) or other care givers (OCG) reported the weight of the child, while the children's height was measured by the interviewer asking the children to take off their shoes and stand against a wall. In the TAS module information on weight and height was self-reported by young adults. I analyse both BMI score and the risk of being overweight/obese. Among children aged below 18, the BMI thresholds to identify the risk of overweight/obesity (by sex and age) are

based on its distribution (BMI exceeding the 95th percentile). Among young adults, BMI is derived from the arithmetic calculation of body weight and height ([Weight in pounds / Height in inches2] X 703), with a BMI equal to 25.0 or greater classified as "overweight/obesity".
Parental separation is measured through two variables, i.e. the marital status of the household heads and marital histories (i.e. number and timing of marriages and separations). Missing values (2.7%; N=432) in parental marital status are concentrated among children living with grandparents only. I create a dummy variable identifying parents who remain in partnership and those who divorce or separate between two consecutive waves. 794 children (2,804 observations) experience parental separation throughout the observation window, while 2,393 children (9,621 observations) live in intact families. I, then, calculate the number of years elapsed between the transition to parental separation and the date of each interview, by using information on the date of end of the first and last marriage.
The mediators included in the analysis regard economic resources, i.e. the total household income and the amount of money spent on food at the household level. I also consider father-child and mother-child closeness measured through a scale ranging from 1 (not close at all) to 4 (extremely close) in the CDS and a scale ranging from 1 (not close at all) to 7 (extremely close) in the TAS. I harmonized these answer categories on the basis of their distribution.
Moderators are considered as time-invariant indicators and are interacted with parental union dissolution in the following analyses. Child sex distinguishes between female children and male children. I distinguish between children who experience parental separation after age 18 (3,715 child-year observations) and those aged 18 or less at parental divorce (1,095 child-year observations). Age 18 is used as a threshold to capture child dependency/ independency on parents. Regarding parental education, I use the number of years that parents spent in education.

2.2 Analytical strategy

I use child-fixed effects linear regression models on changes in BMI score and the probability of becoming overweight/obese. The estimates are based on within-child changes in BMI, which has the advantage to account for time-constant characteristics. First, I examine within-child changes in BMI and overweight/obesity after parental separation. Second, I analyse the moderating effects of children's sex and age at separation, and parental education, by including interaction terms. Third, I consider the timing of parental separation by adding the child's age at parental separation and the time elapsed between marital break-up and measurements of the child's weight.

3 Results

Tables 1 and 2 presents findings from fixed effects linear regression models on child BMI and overweight/obesity respectively. Parental separation is associated with increases in children's BMI and an increased risk of becoming overweight/obese. After parental separation, children's BMI increases of 0.32 points and the risk of overweight/obesity increases of 4.3 percentage points. An increase in the amount of money spent on food is associated with a decrease in BMI and overweight/obesity among children and young adults. Children's BMI also decreases as mother-child relationships improve over time.

In the second sets of models, I include interaction terms between parental union dissolution and child sex, and between parental separation and parental education. The association between family instability and child BMI is largely driven by female children, for whom BMI score increases of 0.85 points after parental separation (main effect). Conversely, male children experience no changes in BMI after parental separation. Among female children, the risk of becoming overweight or obese is 9.1 percentage points higher after parental union dissolution than before the disruption, while it is approximately zero among male children and young adult men (0.091-0.102).

Table 1: Child-fixed effects linear regression models on changes in Body Mass Index.

	Coef.	***Coef.***	***Coef.***
Parental separation	0.321*	0.854**	1.742**
Parental remarriage	-0.110	-0.090	-0.132
Age	1.001**	1.003**	1.000**
Age^2	-0.017**	-0.017**	-0.017**
Total family income (log)	0.021	0.023	0.022
Money spent on food (log)	-0.036+	-0.035+	-0.037+
Closeness with father	0.018	0.020	0.018
Closeness with mother	-0.132*	-0.135*	-0.132*
Parental separation X Boy		-1.220**	
Parental separation X Education			-0.119*
R-squared	0.464	0.465	0.465
Child-year observations	12,425	12,425	12,425
N. of children	2,661	2,661	2,661

** *p<0.01,* * *p<0.05,* + *p<0.1*

The third sets of models introduce interactions between family instability and parental education. Among parents with 16 years of education, union dissolution is associated with a non-significant decrease of 0.16 points in child BMI (1.742-0.119*16). Conversely, among parents with 6 years of education, the BMI of children and young adults increases of 1.02 points after parental separation (1.742-0.119*6). The risk of child overweight/obesity increases of 0.7 percentage points after the separation of highly educated parents (0.167-0.010*16), while it increases of 10.7 percentage points after the separation of low educated parents (0.167-0.010*6). Parental education seems to protect against unhealthy BMI gain and the risk of overweight/obesity related to a family disruption. Robustness checks support this result, indicating that the association between parental union dissolution and

child BMI disappears among highly educated families. This finding is more robust in the analysis of BMI than in the analysis of overweight/obesity across different specifications of parental education.

Table 2: Child-fixed effects linear regression models on the probability of becoming overweight/obese.

	Coef.	***Coef.***	***Coef.***
Parental separation	0.043**	0.854**	1.742**
Parental remarriage	0.003	-0.090	-0.132
Age	-0.022**	1.003**	1.000**
Age^2	0.001**	-0.017**	-0.017**
Total family income (log)	-0.002	0.023	0.022
Money spent on food (log)	-0.004*	-0.035+	-0.037+
Closeness with father	-0.003	0.020	0.018
Closeness with mother	-0.004	-0.135*	-0.132*
Parental separation X Boy		-1.220**	
Parental separation X Education			-0.119*
R-squared	0.015	0.465	0.465
Child-year observations	12,425	12,425	12,425
N. of children	2,661	2,661	2,661

*** p<0.01, * p<0.05, + p<0.1.*

Table 3 shows that parental union dissolution is associated with an increase in BMI among children aged 18 or less at parental separation (interaction terms), but not among older children (main effects). After marital disruption, the risk of becoming overweight/obese increases of 2.8 percentage points for children aged 19 or over, while it increases of 10.5 percentage points for children aged 18 or less.

Table 3 shows no significant deviations in children's BMI from the baseline level before and upon marital break-up, indicating no anticipation effects in the build-up to parental separation nor immediate effect on BMI in the year of separation. Family conflicts and tensions occurring before the actual decision to separate seem to have no immediate influence on children's BMI. The BMI of children from dissolved families increases one year after parental separation and remains above the baseline for the following ten years. Similarly, the risk of becoming overweight/obese increases in the year after parental union dissolution and remains higher than the baseline. The association between parental divorce and unhealthy BMI gain persists for at least 10 years after parental separation. To check the robustness of these results, I perform sensitivity analyses by changing the reference category. I find similar patterns in child BMI and overweight when the reference category is five or seven years before parental separation.

Table 3: Child-fixed effects linear regression models on BMI and obesity risk.

	Coef.	***Coef.***	***Coef.***	***Coef.***
Parental separation	0.166		0.028*	
Separation X child age at sep. ≤ 18	0.806**		0.077**	
Time since/to parental separation				

(ref. -6 y. or more)				
-5/-3 y.		0.090		0.007
-2/-1 y.		0.375		0.029
0 (year of separation)		0.409		0.027
+1/+2 y.		0.625*		0.039+
+3/+5 y.		0.446*		0.049*
+6/+9 y.		0.473*		0.063**
+10 y. or more		0.536*		0.057*
R-squared	0.465	0.465	0.016	0.015
Child-year observations	12,425	12,425	12,425	12,425
N. of children	2,661	2,661	2,661	2,661

*** p<0.01, * p<0.05, + p<0.1. Control variables are those presented in Tables 1 and 2.*

4 Limitations

First, the main variable of interest, children's body mass index, is measured for children aged at least five. The reported changes in child BMI and the risk of child overweight/obesity could be underestimated. Second, the study results are generalizable to a specific cohort of children born 1988-1997. Third, this study provides little contribution on the mechanisms of why children's weight status is affected by parental separation. More detailed information on children's diets, activities, and time use, are needed to identify pathways.

References

1. Baltrus, P. T., Lynch, J. W., Everson-Rose, S., Raghunathan, T. E., & Kaplan, G. A. (2005). Race/ethnicity, life-course socioeconomic position, and body weight trajectories over 34 years: the Alameda County Study. Am J Public Health, 95(9), 1595-1601.
2. Bzostek, S. H., & Beck, A. N. (2011). Familial instability and young children's physical health. Soc Sci Med, 73, 282–292.
3. Goisis, A., Özcan, B., & Van Kerm, P. (2020). Do children carry the weight of divorce?. Demography, 56(3), 785-811.
4. Hales, C. M., Carroll, M. D., Fryar, C. D., & Ogden, C. L. (2017). Prevalence of obesity among adults and youth: United States, 2015–2016, NCHS data brief, 288.
5. Hernandez, D. C., Pressler, E., Dorius, C., & Mitchell, K. S. (2014). Does family instability make girls fat? Gender differences between instability and weight. J Marriage Fam, 76, 175–190.
6. Schmeer, K. K. (2012). Family structure and obesity in early childhood. Soc Sci Res, 41, 820–832.

Living conditions, well-being and poverty

Analyzing the impact of COVID-19 pandemic on elderly population well-being

Un'analisi dell'impatto della pandemia da COVID-19 sul benessere della popolazione anziana

Gloria Polinesi, Mariateresa Ciommi, Chiara Gigliarano

Abstract Aim of the paper is to analyse the effect of COVID-19 pandemic on multidimensional italian well-being of the population aged 50 or over by measuring the individual well-being changes before and after pandemic. To capture the multidimensional nature, we consider different dimensions: economic, health, social connections and work. Therefore, an individual well-being change index is constructed to measure non-directional, downward and upward movements. We use micro-data from waves 8 and 9 of the Survey of Health, Ageing and Retirement in Europe (SHARE). Findings suggest that employed and richer individuals suffer greater well-being losses highlighting the key role of health on well-being.

Abstract *Abstract in Italian* Scopo del documento è analizzare l'effetto della pandemia da COVID-19 sul benessere multidimensionale della popolazione italiana di età pari o superiore a 50 anni misurando i cambiamenti del benessere individuale prima e dopo la pandemia. Per cogliere la natura multidimensionale, prendiamo in considerazione diverse dimensioni: economica, sanitaria, relazioni sociali e status lavorativo. Pertanto, viene costruito un indice di cambiamento del benessere individuale per misurare i movimenti non direzionali, al ribasso e al rialzo. L'applicazione empirica utilizza i microdati 2019-2021 dell'Indagine SHARE. I risultati suggeriscono che gli occupati e gli individui più ricchi subiscono maggiori perdite di benessere, evidenziando il ruolo chiave della salute sul benessere.

Key words: Multidimensional well-being, SHARE data, COVID-19.

Gloria Polinesi
Dipartimento di Economia, Università degli Studi dell'Insubria, Varese, Italy e-mail: gloria.polinesi@uninsubria.it

Mariateresa Ciommi
Dipartimento di Scienze Economiche e Sociali, Università Politecnica delle Marche, Ancona, Italy e-mail: m.ciommi@univpm.it

Chiara Gigliarano
Dipartimento di Economia, Università degli Studi dell'Insubria, Varese, Italy e-mail: chiara.gigliarano@uninsubria.it

1 Introduction

Older adults, especially those with vulnerable health conditions, have been affected disproportionately by COVID-19 (see [7]). In fact, COVID-19 pandemic can cause social disruptions (e.g., job loss, social distancing, confinement), which, in turn, can affect the individual well-being. For this reason, it is important to assess the impact that the disruption of COVID-19 has on different dimensions of well-being of this vulnerable group.
Over the last decades approaches to measuring well-being have received much attention from both researchers and policy-makers, starting from the seminal work of [9] and [10].
Several initiatives have taken place proposing multidimensional indicators pointing out that income alone does not reflect the multi-faceted nature of the well-being. For example, many authors have proposed indicators to measure different aspects of the well-being's distribution (see [5] and [2]). Other contributions are, among others, [6] and [4]. For a deep overview we refer to [1] and to [3].
As well-being is a multidimensional concept, its measures should be able to capture not only the economic effects of the COVID-19 crisis but also the social ones.
Aim of the paper is, therefore, to understand the consequences of the COVID-19 outbreak on the dimensions (economic, social connections, work status and health) of well-being of elderly European population with an analysis based on the Survey of Health, Ageing and Retirement in Europe (SHARE) data.
An analysis by subgroup has been conducted to investigate categories among individuals aged 50+ more vulnerable to COVID-19 pandemic distinguishing first and second year of the health crisis.
The remainder of this paper is organized as follows. Section 2 introduces data and propose a multidimensional well-being change index. Section 3 is devoted to empirical findings and discussion. Section 4 draws some conclusions.

2 Data and methods

The empirical analysis is based on data provided by the Survey of Health, Ageing and Retirement in Europe and Israel (SHARE), which is a longitudinal and interdisciplinary database gathering microlevel information on health, well-being, and socioeconomic characteristics for the population aged 50 or older in Italy. We focus on the waves 8 and 9.[1]
Health, employment, equivalent annual disposable income and the ability to make ends meet are used to construct the well-being change indices.[2] We also consider

[1] Wave 8 regular survey relates to pre-COVID period (from 10-2019 to 03-2020). Wave 8 corona survey collects data during the first year of COVID-19 (from 06-2020 to 08-2020), while wave 9 refers to the second year of COVID-19 (from 06-2021 to 08-2021).

[2] For the complete list of variables used in the analysis we refer to [8].

sociodemographic variables such as gender, age, and education level (ISCED classification) to investigate effects of COVID-19 pandemic by specific subgroups.
We compute three different measures to catch downward, upward and non-directional (overall) changes in the individual multidimensional well-being. Consider a population of individuals $i = 1, \ldots, n$ over periods of time t and $t-1$, and denote with x_t^{ik} and x_{t-1}^{ik} the value of the k-th well-being indicator at time t and $t-1$ respectively, with $k = 1, \ldots, K$. The individual downward well-being change index is defined as:

$$m_i = \frac{\sum_{k=1}^{K} \mathbb{1}(x_t^{ik} < x_{t-1}^{ik}) v_k}{\sum_{k=1}^{K} v_k}, \tag{1}$$

where v_k is the weight of each well-being indicator such that $\sum_{k=1}^{K} v_k = K$. In what follows, we assume equal weight of the well-being dimensions such that $v_k = 1/K$, for $k = 1, \ldots, K$.

The upward and the overall indices can be obtained from Equation (1) by replacing $<$ symbol with $>$ and $\neq$, respectively. The aggregate well-being change index, aimed to asses the intensity of the COVID-19 effects in a given subgroup or country, can be defined as the weighted mean of m_i:

$$M = \frac{\sum_{i=1}^{n} m_i w_i}{\sum_{i=1}^{n} w_i}, \tag{2}$$

where m_i is the individual well-being change index defined in Equation (1) and w_i represents the individual sample weight.

3 Empirical findings and discussion

Table 1 shows that in the first period of the pandemic the downward change of well-being is significantly higher than the upward change, meaning that for the elderly Italian population the net effect of the pandemic reveals to be a worsening of the well-being. In the second year of pandemic both downward and upward changes in well-being increases significantly (see Table 2).
Splitting the population by gender, we note that in the first year of pandemic males worsen their well-being more than females.
Tables 1 and 2 show also that well-being changes differ significantly by work-status, education and income class. In particular, multidimensional well-being of employed individuals is significantly more affected by the COVID-19 pandemic than the other work-status categories, with higher levels of downward changes. Moreover, having upper secondary education implies more pronounced negative changes in the well-being with respect to higher and lower educated individuals. Poorest income classes are less affected by downward changes than higher income classes both in the first and in the second year of pandemic.

	Overall			Downward			Upward		
	Index	95% CI		Index	95% CI		Index	95% CI	
Total	0.313	0.304	0.322	0.169	0.161	0.178	0.144	0.137	0.152
Gender									
Male	0.324	0.312	0.337	0.184	0.169	0.198	0.140	0.129	0.152
Female	0.304	0.292	0.315	0.156	0.147	0.166	0.147	0.138	0.156
Education									
Primary-lower secondary	0.311	0.302	0.319	0.156	0.146	0.165	0.155	0.147	0.163
Upper secondary	0.330	0.308	0.352	0.205	0.185	0.225	0.125	0.109	0.141
Tertiary	0.281	0.237	0.328	0.152	0.123	0.186	0.128	0.090	0.166
Work status									
Retired	0.299	0.291	0.307	0.138	0.131	0.146	0.160	0.152	0.169
Employed	0.345	0.327	0.362	0.228	0.208	0.248	0.116	0.100	0.133
Other	0.292	0.272	0.313	0.134	0.116	0.152	0.158	0.144	0.173
Income quantile									
First	0.292	0.266	0.319	0.097	0.076	0.117	0.195	0.175	0.216
Second	0.328	0.316	0.341	0.100	0.086	0.114	0.228	0.213	0.243
Thirth	0.306	0.287	0.325	0.155	0.139	0.170	0.151	0.133	0.170
Fourth	0.334	0.315	0.351	0.201	0.183	0.217	0.133	0.118	0.149
Fifth	0.305	0.285	0.326	0.242	0.222	0.262	0.063	0.050	0.076

Table 1: Well-being change indices, total and by subgroups (index and 95 % bootstrap confidence interval) in the first year of pandemic.

	Overall			Downward			Upward		
	Index	95% CI		Index	95% CI		Index	95% CI	
Total	0.381	0.370	0.391	0.190	0.180	0.200	0.190	0.182	0.199
Gender									
Male	0.380	0.363	0.395	0.197	0.179	0.214	0.183	0.170	0.195
Female	0.381	0.366	0.395	0.185	0.173	0.197	0.196	0.186	0.206
Education									
Primary-lower secondary	0.380	0.368	0.392	0.181	0.169	0.193	0.199	0.190	0.208
Upper secondary	0.393	0.372	0.413	0.219	0.200	0.237	0.174	0.157	0.191
Tertiary	0.350	0.302	0.398	0.175	0.143	0.209	0.175	0.140	0.209
Work status									
Retired	0.378	0.370	0.387	0.171	0.163	0.179	0.207	0.198	0.216
Employed	0.402	0.378	0.426	0.249	0.224	0.275	0.153	0.132	0.173
Other	0.359	0.338	0.381	0.154	0.139	0.169	0.206	0.191	0.221
Income quantile									
First	0.380	0.353	0.407	0.122	0.105	0.140	0.257	0.240	0.275
Second	0.406	0.392	0.420	0.135	0.122	0.148	0.271	0.257	0.285
Thirth	0.390	0.360	0.420	0.207	0.171	0.243	0.183	0.161	0.204
Fourth	0.389	0.370	0.407	0.213	0.199	0.228	0.175	0.157	0.193
Fifth	0.346	0.326	0.367	0.256	0.238	0.274	0.090	0.077	0.104

Table 2: Well-being change indices, total and by subgroups (index and 95 % bootstrap confidence interval) in the second year of pandemic.

Figure 1 shows the percentage of elderly individuals who suffer a worsening (left panel) and an improvement (right panel) in one or more dimensions of well-being. We note that during the second year of COVID-19 pandemic, the percentage of individuals who worsen in one dimension of their well-being is increased and, at the same time, also the percentage of individuals with an improvement in two or three dimensions slightly increases.

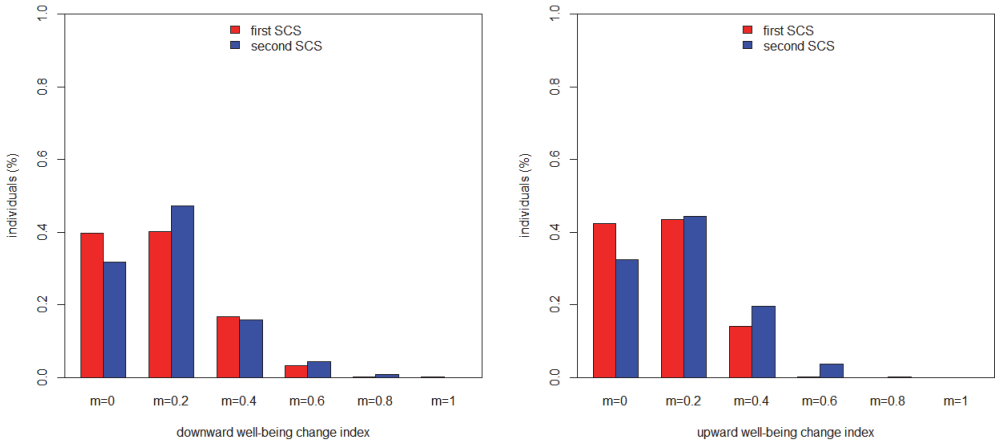


Fig. 1: Percentage of individuals who suffer worsening (left panel) or improvement (right panel) in multidimensional well-being for different index cutoff values: $m = 0$ refers to individuals who worsen in none of the well-being dimensions, $m = 0.2$ to individuals who worsen in one dimension, and so on. Cutoff $m = 1$ stands for a worsening in all the dimensions considered.

4 Conclusion

The paper contributes to the analysis of well-being of elderly Italians. Specifically, we compute a multidimensional index that captures changes in the level of individual well-being during first and second year of COVID-19 pandemic.
Findings suggest that employed and richer individuals suffer greater well-being losses, while gender is not decisive for discriminating against changes in individual well-being.
Since the local dimension plays a crucial role in well-being measurement ([11]), further researches will be aimed to include regional dimension.

References

1. Aaberge, R., and Brandolini, A. Multidimensional poverty and inequality. In A. B. Atkinson and F. Bourguignon (Eds.), Handbook of income distribution (pp. 141–216). Amsterdam: Elsevier (2015)
2. Atkinson, A. B., Bourguignon, F.: The comparison of multi-dimensioned distributions of economic status. The Review of Economic Studies **49 (2)**, 183–201 (1982)
3. Chakravarty, S. R. Analyzing multidimensional well-being: A quantitative approach. Hoboken: Wiley (2018)
4. Gigliarano, C., Mosler, K.: Constructing indices of multivariate polarization. The Journal of Economic Inequality **7 (4)**, 435–460 (2009)
5. Kolm, S. C.: Multidimensional egalitarianisms. The Quarterly Journal of Economics **91 (1)**, 1–13 (1977)
6. Maasoumi, E.: Multidimensioned approaches to welfare analysis. Springer, Dordrecht (1999)
7. Mueller, A. L., McNamara, M. S., Sinclair, D. A.:Why does COVID-19 disproportionately affect older people?. Aging (Albany NY) **12 (10)**, 9959–9981 (2020)
8. Polinesi, G., Ciommi, M., Gigliarano, C.: Elders and COVID-19: an analysis on multidimensional well-being changes in European countries. Mimeo (2022)
9. Sen, A. K. Equality of what? The Tanner Lecture on Human Values, I, 197–220 (1980)
10. Sen, A. K. Commodities and capabilities. Amsterdam: North-Holland (1985)
11. Stiglitz, J. E., Sen, A., Fitoussi, J.-P. . Report by the commission on the measurement of economic performance and social progress. Technical report, Institut National de la Statistique at des etudes economiques. (2010)

Exploring sustainable food purchasing behaviour using Italian scanner data

Un'esplorazione delle abitudini di consumo sostenibili tramite l'analisi dei dati scanner in Italia

Ilaria Benedetti[1], Alessandro Brunetti[2], Federico Crescenzi[3], Luigi Palumbo[4]

Abstract Due to the growing consumers' interest in "environmentally-friendly" products, in this paper we provide an exploration of the scanner data over the Italian provinces in term of share of organic products assortments and turnover share and how territorial distribution can reflect consumers' actual purchases (or observed behaviours) in assessing price premiums for organic products. To this aim, we estimated hedonic pricing models to evaluate the organic price premium for the 103 Italian provinces of selected food aggregates according to homogeneous market classification and by considering both branded and private label products.

Abstract *A seguito del crescente interesse dei consumatori per i prodotti biologici, in questo lavoro forniamo un'esplorazione dei dati scanner a livello provinciale italiano in termini di assortimenti di prodotti biologici e quota di fatturato e di come la distribuzione territoriale può riflettere gli acquisti effettivi dei consumatori (o i comportamenti osservati) nella valutazione dei premi di prezzo per i prodotti biologici. A questo scopo, abbiamo stimato modelli di prezzo edonici per valutare il premio di prezzo per il biologico per le 103 province italiane di aggregati alimentari selezionati secondo una classificazione di mercato omogenea e considerando sia i prodotti di marca che quelli a marchio privato.*

Key words: organic food consumption, scanner data, hedonic models.

[1] Ilaria Benedetti
University of Tuscia, email: i.benedetti@unitus.it
[2] Alessandro Brunetti
Istituto Nazionale di Statistica, email: albrunetti@istat.it
[3] Federico Crescenzi
University of Tuscia, email: federico.crescenzi@unitus.it
[4] Luigi Palumbo
University of Tuscia, email: luigi.palumbo@unitus.it

1 Introduction

Over the last decades, there has been growing consumer interest in "environmentally-friendly" products that are able to provide higher levels of personal and environmental well-being (Caron et al., 2018).

The increased demand of organic products, that are defined as fresh or processed food produced by organic farming methods without the use of synthetic chemicals, such as human-made pesticides and fertilizers, and without containing genetically modified organisms have been linked to consumer's environmental concern, farmers' welfare higher nutritional value and health, in recent studies (Laureti and Benedetti, 2018; Kushwah, 2019). Moreover, purchasing behaviour toward organic food products proved to be influenced by socio-economic and personal characteristics of the area in which individuals reside.

Food consumption is one of the most important areas which influence environmental sustainability since it is responsible for one third of a household total environmental impact (Vermier et al., 2020). More specifically organic food is often considered as a spearhead for transition towards more sustainable food production and consumption (Vittersø and Tangeland, 2015).

Therefore, in order to ensure long-term sustainability in the recovery of the global economy in the post COVID-19 pandemic era, societies of the future will be called upon to become increasingly inclusive and sustainable, including through green innovations.

Over the last three decades, there has been a steady increase in organic food and farming. In 2020, the number of organic producers in the EU increased by 1.6% compared to 2019.

In Italy the organic agricultural land is equal to 15.8% of the total cultivated land and organic farms in Italy represent 6.2% of total farms (Rapporto ISMEA, 2020) with high territorial heterogeneity. The highest number of organic operators is observed in the Southern regions (Sicily, Calabria and Apulia). The amount of Italian organic retail sale in 2019 was equal to 3,625 million of euros, with an equivalent spent per capita equal to 60 euros/month (compared to 84 euros/month at EU level[5]). The BioBank 2019 report[6] showed that in 2009, 45% of Italian consumers of organic products purchase their desired goods by visiting traditional shops, while the modern retail channel covers 19% of purchasing. However, between 2016 and 2018, organic sales in the modern retail channel, including supermarket and hypermarkets chains, (showed a double-digit increase, thus becoming the main sales channel for organic products. This acceleration was brought about by the massive offer of organic products in the modern trade outlets, in particular by the presence of private label products from modern distribution.

Although consumer interest in organic food has risen over time, resulting in a generally positive attitude toward these organic food products, the existing literature is stacked with studies relating factors affecting buying behaviour for organic food

5 https://www.organicseurope.bio/about-us/organic-in-europe/
6 https://www.biobank.it/?mh1=8&cs=8

items (Tandon et al. 2020). Yet, few research studies analyse the factors influencing how the markets respond to the organic food products growing presence .

The increasing availability of scanner data, providing highly detailed information on quantities and turnover for a huge number of products sold by chain stores on a weekly basis allows exploring sustainable food purchasing behaviour by also considering the territorial heterogeneity. The traditional data sources, such as the Household Budget survey, carried out by ISTAT for estimating Italian households expenditure, does not allow analysing sustainable consumption behaviour in terms of organic products purchasing behaviour.

Using 2018 Italian official scanner data we explore this new source of consumer data over the Italian territory in terms of assortments share and turnover share of organic products with the aim of analysing whether and to what extent territorial differences reflect consumers' actual purchases (or observed behaviors) in assessing price premiums for organic products.

In this paper, we estimated hedonic pricing models to evaluate the organic price premium for the 103 Italian provinces of selected food aggregates according to homogeneous market classification and by considering both branded and private label products.

2 Data and Methods

2.1 Italian Scanner data

In this paper, we use a scanner dataset obtained through an agreement between Istat and ASESD-Dagum Centre[7] in order to implement the tasks of the Makswell project[8] for the year 2018. ISTAT acquires data from hypermarkets and supermarkets located in all the Italian provinces and belonging to the most important 16 retail chains. ISTAT 2018 scanner data include data of turnover and quantities for each product code, using the Global Trade Item Number (GTIN[9]). The sample of outlets is representative of the entire universe of large-scale retail trade hypermarkets and supermarkets and includes approximately 1,800 hypermarkets (more than 500) and supermarkets (almost 1,300), concerning the grocery products sold in the most important retail chains (95% of modern retail chain distribution that covers 55.4% of total retail trade distribution for this category of products). More specifically, outlets have been stratified according to provinces (107), chains distribution (16) and outlet-types (hypermarket, supermarket) for a total of more than 800 strata. Probabilities of selection were assigned to each outlet based on the corresponding turnover value.

7 For further information about Dagum Center, please visit the web-site: http://www.centrodagum.it/
8 For further information about MAKSWELL project please visit web-site: https://www.makswell.eu/
9 GTIN (previously known as EAN) is the key code used to identify products and packages to sell them in the modern distribution.

For each GTIN, price are calculated by taking into account turnover and quantities observed in each province and type of outlet.

2.2 Method : the Dummy variable Hedonic Method

The hedonic analytical framework, developed by Lancaster (1966) and Rosen (1974), proved to be useful for studying the quality attributes of several food products. Heravi et al. (2003) provided the first application of scanner data in this framework. The Dummy variable Hedonic Method is akin to the Country Product Dummy method developed by Summers (1973), with the exception that the model incorporates a detailed set of variables on the quality characteristics. With the aim of exploring price premiums for organic products we applied hedonic price models to some basic headings included in the Italian scanner data.

For j=1,2…, m areas, i=1,2,…n items in a basic heading, p_{ij} represents the price of i-th item in j-th geographical area and ε_{ij} is the error term, the hedonic regression is given by:

$$lnp_{ij} = \beta_0 + \sum_{i=1}^{n} \beta_i Z_i + \sum_{j=1}^{m} \alpha_j D_j + \varepsilon_{ij}$$

Where D_j is a dummy variable equal to 1 for geographical area j and zero otherwise; Z_i is a vector of product characteristics (organic vs non-organic products , branded vs private label) βj are coefficients to be estimated, and ε is the error term. The antilogarithms of α_j are ordinary-least-squares (OLS) estimates of the area-specific price parities[10] with respect to the overall mean of the areas.

Following Roheim et al. (2011), the log-linear configuration is used in the estimation since it presents a twofold advantage with respect to other ones: it allows obtaining residuals that are approximately normally distributed and the interpretation of regression coefficients is more immediate: the dependent variable changes by $\left(e^{\beta} - 1\right) * 100$ for a one-unit increase in one of the regressors, holding all other variables fixed. Since for each item included in scanner data, quantity and turnover in each area are available in addition to the prices, following Diewert's (2005) proposal, we assume that a weighted least squares (WLS) regression is run with the expenditure shares in each geographical area serving as weight. In order to estimate price premium of organic products, we estimated coefficients on the explanatory dummy variable "organic vs traditional" then the explanatory variables were interacted with the area dummies.

[10] Price parities are spatial price index numbers. The concept price parity is used to measure the price level in one location compared to that in another location. More specifically, at the international level, purchasing power parities of currencies are defined as the number of currency units of a country that can purchase the same basket of goods and services that can be purchased with one unit of currency of a reference currency (World Bank, 2013).

3 Results and conclusions

In this section we report results for selected products aggregate: yoghurt, rice, eggs and flour. In preliminary data analysis, we observed that organic products generally have smaller packaging. For this reason, we did not consider the price per piece but the price per kg. Moreover, for each product aggregate we considered a homogeneous group of products according to the market classification[11]: skimmed and whole yoghurt, barn farms, free range and standard chicken eggs, durum and wheat flour.

Due to the limited available space, Figure 1 reports turnover share of organic products (a), quantity sold of organic products (b) number of organic products within the basic heading (c) for yoghurt product aggregate in the 103 Italian provinces.

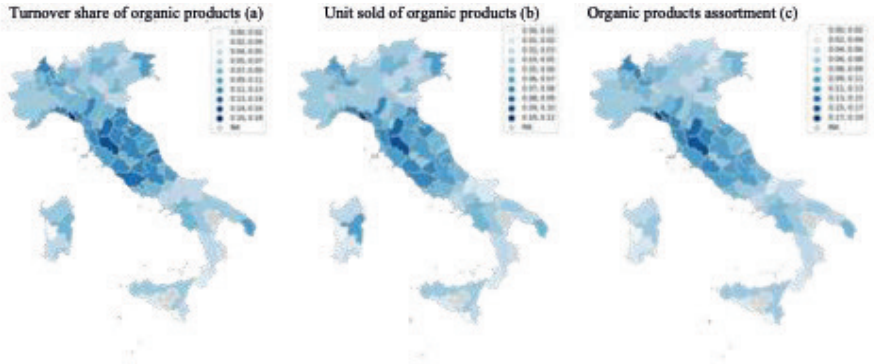


Figure 1: Organic products turnover share, unit sold and product assortment for yogurt product aggregate

From the hedonic regression models we can observe that the price premium for the selected organic products aggregate range from 141% (eggs) to 177% (flour). For yoghurt the price premium at national level is equal to 62.97% with statistical significant differences at provincial level. Indeed, as reported in Figure 2, the highest organic price premium is observed in Pordenone (150.58), Caserta (149.78) Gorizia (148.27), Perugia (148.27) and Reggio Emilia (148.03). While for rice, the organic price premium at national level is equal to 157% with significant differences at territorial level. These differences can be explained by considering living conditions at provincial level (disposable income, food expenditure and food expenditure share).

[11] Markets correspond to the lowest level of the classification of goods shared by industrial and distribution companies, which have been linked to the product aggregates of the ECOICOP classification.

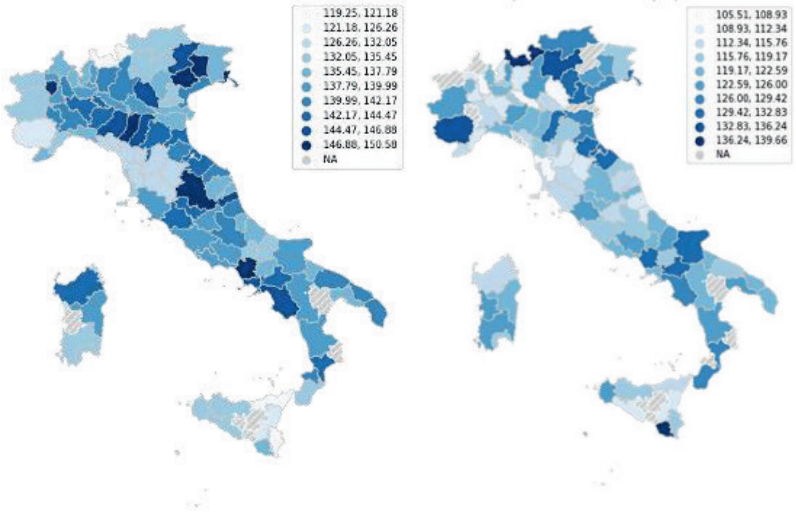


Figure 2: Organic products premium price at provincial level for yoghurt (left) and rice (right).

References

Caron, P., Ferrero y de Loma-Osorio, G., Nabarro, D., Hainzelin, E., Guillou, M., Andersen, I. and Verburg, G. (2018). Food systems for sustainable development: proposals for a profound four-part transformation. Agronomy for sustainable development, 38(4), 1-12.

Diewert, W. E. (2005). Weighted Country Product Dummy Variable Regressions and Index Number Formulae," Review of Income and Wealth, 51, 561–70.

Heravi, S., Heston, A., and Silver, M. (2003). Using scanner data to estimate country price parities: A hedonic regression approach. Review of Income and Wealth, 49(1), 1-21.

Kushwah, S., Dhir, A., Sagar, M., and Gupta, B. (2019). Determinants of organic food consumption. A systematic literature review on motives and barriers. Appetite, 143, 104402.

Lancaster, K. J. 1966. A new approach to consumer theory. J. Polit. Econ. 74:132–157

Laureti, T., and Benedetti, I. (2018). Exploring pro-environmental food purchasing behaviour: An empirical analysis of Italian consumers. Journal of Cleaner Production, 172, 3367-3378.

Tandon, A., Dhir, A., Kaur, P., Kushwah, S., & Salo, J.(2020). Why do people buy organic food? The moderating role of environmental concerns and trust. Journal of Retailing and Consumer Services, 57, 102247.

Summers, R. (1973). International Comparisons with Incomplete Data. The Review of Income and Wealth, March.

Vermeir, I., Weijters, B., De Houwer, J., Geuens, M., Slabbinck, H., Spruyt, A.and& Verbeke, W. (2020). Environmentally sustainable food consumption: A review and research agenda from a goal-directed perspective. Frontiers in Psychology, 11, 1603.

Vittersø, G., and Tangeland, T. (2015). The role of consumers in transitions towards sustainable food consumption. The case of organic food in Norway. Journal of Cleaner Production, 92, 91-99.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. J. Polit. Econ. 82:34–55.

World Bank. 2013. Measuring the Real Size of the World Economy: Th e Framework, Methodology, and Results of the Inter-national Comparison Program —ICP. Washington, DC: World Bank. DOl:10.1596/978-0-8213-9728-2).

The evaluation of heat vulnerability in Friuli Venezia Giulia

Vulnerabilità alle ondate di calore in Friuli Venezia Giulia

Laura Pagani, Maria Chiara Zanarotti and Anja Habus

Abstract Heat waves are leading cause of weather-related illness and death, in a context where their frequency, intensity and impact are expected to surge due to rising climate change, growing urbanisation and population ageing. This work develops a Heat Vulnerability Index by means of the composite indicator methodology with the aim to depict heat vulnerability in Friuli Venezia Giulia at the census tract level. The results show that heat vulnerability follows a spatial pattern with highest hazard in urban areas, lower risk in rural areas and lowest danger in mountainous areas. The performance interval approach is exploited to validate the Index.

Abstract *Le ondate di calore sono la principale causa di malattie e decessi legati ai cambiamenti climatici. La loro frequenza, intensità ed impatto sono destinati ad aumentare a causa del crescente surriscaldamento globale, urbanizzazione ed invecchiamento della popolazione. Questa analisi sviluppa un indice di vulnerabilità alle ondate di calore mediante la metodologia dell'indicatore composito per rappresentare la vulnerabilità, a livello di sezione censuaria, nella regione Friuli Venezia Giulia. I risultati mostrano che la vulnerabilità al calore è alta nelle aree urbane, media in quelle rurali e bassa nelle zone montane. L'approccio basato sull'intervallo di performance viene utilizzato per convalidare l'indice.*

Key words: Heat vulnerability index, heat waves, composite indicator

1 Introduction

With growing global warming, extreme climate events like heat waves (HWs) will increase in duration, frequency and intensity. Urban and metropolitan areas are particularly affected by HWs due to higher heat-absorbing capacity and reduced nighttime cooling capability of these environments with respect to the surrounding rural

Laura Pagani, Department of Economics and Statistics, University of Udine, Italy e-mail: laura.pagani@uniud.it

Maria Chiara Zanarotti, Department of Statistical Sciences, Catholic University of Milan, Italy e-mail: maria.zanarotti@unicatt.it

Anja Habus, e-mail: anja.habus@gmail.com

areas; this is the so-called Urban Heat Island (UHI) effect. HWs are becoming a significant public-health concern as the observed warming has raised heat-related morbidity and mortality of certain categories of individuals that suffer an excessive burden from heat-load. Specifically, the thermoregulatory system of elderly, children and ill individuals is impaired by high air temperatures and humidity levels. Nevertheless, targeted economic investments, mitigation and prevention policies can minimise health impacts of climate threats. Thus, there is a growing interest in understanding the determinants of heat vulnerability (HV) and identifying population and geographical areas more susceptible to adverse health impacts associated with HWs.

2 The Concept of Heat Vulnerability

HV is a latent and multifaceted concept, hence a readily available and comprehensive measure of this complex phenomenon does not exist. To carry out a meaningful measurement, a sound theoretical framework is designed and the determinants of HV are first identified. Secondly, a complete set of non-interchangeable indicators, i.e. the system of basic indicators (BIs), is collected for a comprehensive representation of the construct of interest.

2.1 Definition

The Intergovernmental Panel on Climate Change (IPCC) defines vulnerability as

> [...] the degree to which a system is susceptible to, or unable to cope with, adverse effects of climate change, including climate variability and extremes. Vulnerability is a function of the character, magnitude, and rate of climate change and variation to which a system is exposed, its sensitivity, and its adaptive capacity [2].

In addition, it outlines the three dimensions of vulnerability as follows: **Exposure** (E), the nature and degree to which a system is exposed to significant climatic variations; **Sensitivity** (S), the degree to which a system is affected, either adversely or beneficially, by climate-related stimuli; **Adaptive capacity** (AC), the ability of a system to adjust to climate change (including climate variability and extremes) to moderate potential damages, to take advantage of opportunities, or to cope with the consequences.

The IPCC framework is embraced to assess the subset of vulnerabilities associated with heat stress in the Friuli Venezia Giulia (FVG)[1] and guides the selection of the set of manifest variables that allow to quantify HV.

[1] FVG is Italy's north-easternmost region. Its landscape spans from the Carnic and Julian Alps to the Adriatic Sea, determining the subdivision into four main parts: mountainous-alpine terrain in the north, hilly to the south of the mountains and in the central part of the eastern border with Slovenia, plain from the centre to the coastal area in the south

2.2 The System of Basic Indicators

To reflect both the multifaceted nature of HV as well as the three dimensions identified by the IPCC, climate, environmental, health and socio-demographic data[2] are employed in the analysis. The statistical unit of reference is the census tract, that is the smallest entity of the municipality based on which the data collection of national census surveys is organised.

Table 1: The system of basic indicators

Class	BI	Description	Dimension	Range
Socio-demographic[a]	Under5	No. of individuals 5 years or less	S	$[0-15992]$
	Over65	No. of individuals aged 65 years or more	S	$[0-25942]$
	NoAC	No. of dwellings without air conditioning	E, AC	$[0-37100]$
Health[a]	Car	No. of individuals with cardiovascular diseases	S	$[0-51956]$
	Res	No. of individuals with respiratory diseases	S	$[0-6582]$
	End	No. of individuals with diabetes	S	$[0-5118]$
	Psy	No. of individuals with psychological disorders	S	$[0-11622]$
	MultiPat	No. of individuals with at least two of the above diseases	S	$[0-25921]$
Environmental	Imp	Degree of soil imperviousness	E, AC	$[1-4]$
	Lcpi	Largest patch of continuously built area	E, AC	$[0-1]$
	Lai	Leaf area index	E, AC	$[1-6]$
Climate	ThomMax	10-years average of daily Thom Index maxima	E	$[21-30]$
	ThomAvg	10-years average of daily Thom Index means	E	$[13-23]$

[a] Normalised by census tract's area (km^2)

As depicted in Table 1, sensitivity to heat is measured by data on population structure and healthiness condition. Climate BIs are indicators of exposure as the Thom Index represents the combined effect of both temperature and humidity on the warming and discomfort level perceived by the human body. Instead, availability of air conditioning and environmental variables do not find a clear-cut classification across literature: they are either proxies of exposure or adaptive capacity, depending on the subjective choice of the analyst. For instance, the number of dwellings without air conditioning can reflect indoor temperatures (E) or the willingness of individuals to seek relief from heat load (AC).

3 A Heat Vulnerability Index for FVG

To obtain a unidimensional measure of heat-risk the system of BIs is summarised into a single variable, called Heat Vulnerability Index (HVI), by means of the Composite Indicator (CI) methodology [3]. CIs construction is a complex task as it entails several steps and choices: BIs selection, functional form assessment, outliers and missing data treatment, normalisation, weights definition and aggregation.

[2] Climate data are provided by *ARPA FVG Osservatorio Meteorologico Regionale*; environmental, health and socio-demographic data are provided by *Regione Autonoma Friuli Venezia Giulia*

After testing several combinations of the aforementioned steps, health and socio-demographic data are log-transformed and BIs are normalised with the the min-max method. These are then aggregated in two stages based on a two-dimensional structure of HV (exposure and sensitivity). At a first stage, a cubic mean with equal weights is applied within HV dimensions. The adoption of a non-linear aggregation function allows the introduction of a partially compensatory approach. This choice is driven by the intention to preserve the impact of those features that make a census tract, and the related inhabitants, vulnerable to HWs either in terms of sensitivity or exposure. In other words, in case of imbalance between BIs, those with a high value do not compensate the ones with a low value. For instance, a census tract characterised by high density of elderly individuals is not fully compensated by a healthy population and thus is still highly sensitive to HWs. As the HVI is an indicator with negative polarity, meaning that an increase in the values of the index corresponds to a worsening of the phenomenon, an upward penalisation must be used. The cubic mean corresponds to a high upward penalisation. At a second stage a linear aggregation with equal weights is applied between HV dimensions. This means that a compensatory approach is instead chosen across dimensions and therefore low values of a dimension, e.g. sensitivity, linearly compensate high values of another dimension, e.g. exposure and vice versa.

4 Heat Vulnerability Patterns in FVG

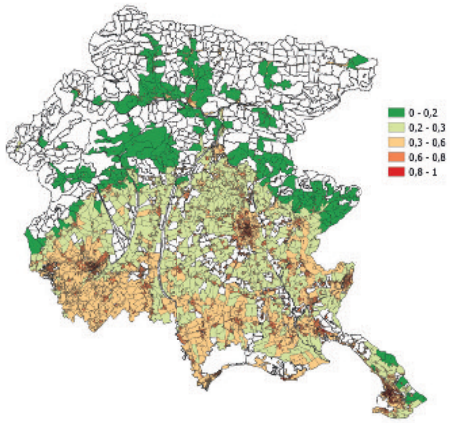


Fig. 1: Heat vulnerability map for FVG

Fig. 1 depicts HV patterns in FVG region (HVI ranges from 0 to 1, min. and max. vulnerability respectively)[3]. The map clearly shows that most urbanised and densely populated areas are the ones that record the highest values of HV: Trieste,

[3] The total number of census tracts is 6,835. Blank areas refer to census tracts having no inhabitants or that never recorded a discomfort from heat as per the Thom Index. The latter are located in the mountainous area, as in line with the orography of the region

Udine, Pordenone, Gorizia, Monfalcone and Sacile. Moreover, HVI increases in the surroundings of SR252, the regional road that connects Palmanova to Codroipo. Also some census tracts in the touristic and coastal cities of Lignano Sabbiadoro and Grado display high levels of HV, whereas other less urbanised census tracts in the same area do not record the same values, e.g. Lignano Riviera. This specific case highlights the impact of invasive tourism and high urbanisation. If furthermore suggests to policy maker to avoid the repetition of such policies in planning new touristic spots. The map points out that mountainous areas (Tolmezzo and Gemona) as well display high levels of HV due to elevated levels of urbanisation. As expected, census tracts located in the north are overall less vulnerable to HWs, whereas the HVI records intermediate levels in the central part of the region.

As previously mentioned, the CI methodology implies multiple decisions that might influence the resulting Index. To limit the impact of the aggregation methods, the performance interval (PI) approach proposed by Mazziotta and Pareto [1] is employed in this analysis. The researches suggest to compute, for each statistical unit, an interval of possible values rather than a single figure. The range of the PI depends on the level of compensability of BIs and their imbalances, and it generates a lower (LB) and upper (UB) bound for the HVI. Since the HVI has negative polarity, the LB corresponds to the hypothesis of full compensability (arithmetic mean of BIs), whereas the UB corresponds to the hypothesis of non-compensability (maximum across the BIs). The midpoint (MP) of the interval represents the case of a partially compensatory BIs.

Taking into account that the HVI is computed in two steps, the PI is adopted at the second level of aggregation, i.e. investigating different levels of compensability between the dimensions of exposure and sensitivity, whereas within the two HV dimensions the cubic mean is maintained. In addition, data are aggregated at a higher administrative level, i.e. municipality, to allow for a better interpretability of the results.

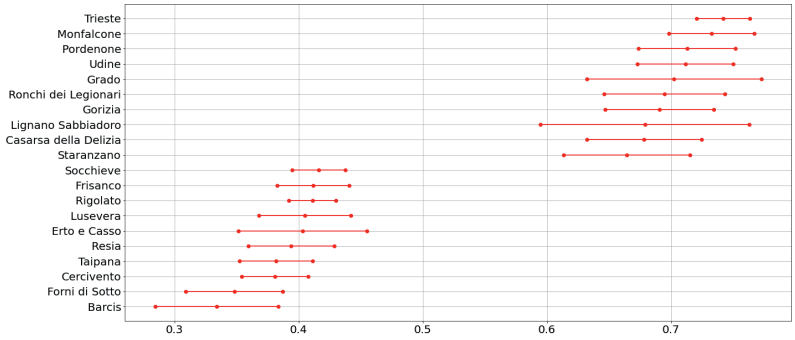


Fig. 2: Performance intervals - Top and bottom 10 municipalities

Fig. 2 displays the PIs of the top and the bottom 10 census tracts when ranked according to the MP[4]. It can be noticed how the ranking changes depending on the

[4] FVG counts 218 municipalities, determining the impossibility to effectively visualise all of them

chosen aggregation function. Grado is ranked 5^{th} based on a MP ranking but it is 1^{st} based on UB ranking. Beyond the impact of the aggregation function, Fig. 2 further confirms that the biggest cities in FVG display the highest values of vulnerability: Trieste, Monfalcone, Pordenone, Udine and Gorizia are all ranked in the first 10 positions, which confirms the validity of the built HVI and the observations resulting from Fig. 1. Conversely, municipalities with the lowest levels of vulnerability are all located in the Alps.

5 Conclusions

Measuring HV is a complex task due to its multidimensional and latent nature. However, the CI technique allows to summarise multifaceted phenomena into a single measure easily accessible to the general public and policy makers.

The resulting HVI for FVG shows that the census tracts recording the highest values are located in urbanised and highly populated areas, specifically in the cities of Trieste, Udine, Pordenone, Gorizia and Monfalcone. Few census tracts at risk are also in the tourist areas of Lignano Sabbiadoro and Grado. These outcomes are in line with expectations, as the synergies of climate and the UHI effect lead to higher daytime temperatures and reduced night-time cooling capacity of cities with respect to rural areas. This behaviour is driven by reduced availability of green areas and higher imperviousness levels of the urban landscape. Furthermore, Trieste, Udine and Pordenone, compared to the surrounding areas, record higher at risk population density, i.e. elderly, children and ill individuals. On the contrary, mountainous and rural areas record the lowest values of HV. These results are validated and confirmed through the PI approach: regardless the use of more or less compensatory techniques in aggregating the HVI, the set of risky areas previously detected remains the same.

Policy makers are suggested to redesign urban plans with the aim to change the way cities develop and grow. The current structure of urban areas should also be reshaped by introducing parks, tree-lined avenues or vegetated rooftops and incentives to steer economic development outside urban areas should be introduced in order to control the urbanisation process.

References

1. Mazziotta, M., Pareto, A.: Composite Indices Construction: The Performance Interval Approach. Soc. Indic. Res. (2020) doi: 10.1007/s11205-020-02336-5
2. McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S.: IPCC, 2001: Climate change 2001: impacts, adaptation and vulnerability, Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge (2001)
3. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffmann, A., Giovannini, E.: Handbook on constructing composite indicators: methodology and user guide. OECD publishing, Paris (2008)

Data Science for Functional and Complex Data

A parsimonious approach to representing functional data

Un approccio parsimonioso per rappresentare dati funzionali

Enea G. Bongiorno and Aldo Goia

Abstract The correction term appearing in a Small ball probability factorization for functional Hilbert data is considered and some properties are presented. Such term leads to a new local dimensionality reduction method that allows a parsimonious representation of data. For the sake of illustration, this approach is applied to the Tecator dataset.

Abstract *Vengono descritte alcune proprietà del fattore correttivo che appare in una fattorizzazione della probabilità di una piccola bolla per dati funzionali in spazi di Hilbert. Questo termine correttivo porta a definire un nuovo metodo di riduzione della dimensionalità locale che permette una rappresentazione parsimoniosa dei dati. A fini illustrativi, questo approccio è applicato al dataset Tecator.*

Key words: Small ball probability factorization, local dimension, correction term

1 Introduction

One of the main problem in the functional data analysis, that is the toolkit of statistical methodologies to treat sample of curves, surfaces or other objects taking values in infinite dimensional spaces (for a review, see e.g. the monographes [3], [4] or [5]), is the representation of the data in small dimension.

To achieve the goal, a typical approach is to use a truncated version of the Karhunen–Loève decomposition: given a separable Hilbert space $\mathscr{H}$ equipped with an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$ and a functional random element X

Enea G. Bongiorno
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: enea.bongiorno@uniupo.it

Aldo Goia
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: aldo.goia@uniupo.it

taking values in $\mathscr{H}$ with mean μ and covariance operator Σ, one can write

$$X \approx \mu + \sum_{j=1}^{d} \xi_j v_j \qquad \mathbb{E}[\xi_j \xi_k] = \lambda_j \delta_{jk} \tag{1}$$

where $d < +\infty$, $\xi_j = \langle X, v_j \rangle$ are the so-called Principal Components (PCs) of X, (λ_j, v_j) are the eigenelements of the covariance operator of X, and $\delta_{jk} = 1$ if $j = k$, and zero otherwise. The quality of the approximation provided by (1) is often measured by the so-called fraction of explained variance (FEV), that is

$$FEV(d) = \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j} 100\%$$

The proposed criterium is global: given a sample of functional data, a unique dimension d is selected for all the data. As a consequence, d could be too large for some of them and too small for other ones, thus producing inefficient or inadequate representations.

This paper aims to overcome this drawback, illustrating an approach to customize the choice of dimension for each element in a sample, through a local–based methodology. The latter exploits the properties of the correction term C_d appearing in the following Small-Ball Probability (SmBP) factorization (see [1] and [2]): given a positive integer d and a point $x \in \mathscr{H}$

$$\mathbb{P}(X \in B(x,h)) \sim f_d(x) V_d(h) C_d(x,h), \qquad h \to 0, \tag{2}$$

where $B(x,h)$ is the ball centred at x with radius h, $f_d(x)$ is the pdf of the first d PCs, V_d is the volume of the d–dimensional ball of radius h. In an intuitive way, for a fixed x, $C_d(x,h)$ provides a compensation for the use of the finite dimensional factorization $f_d V_d$. If that correction term is close to zero, it means that the selected dimension d is inadequate, because of the factorization $f_d V_d$ badly approximates the SmBP being x element of a space having dimension greater than d. On the other hands, if $C_d(x,h)$ reaches its maximum, d is a good choice to approximate x. These arguments allow to interpret $C_d(x,h)$ as a local measure of the quality of the representation of x as an element of a d-dimensional subspace of $\mathscr{H}$.

In this paper, this idea is described and applied. The outline is as follows: Section 2 illustrates the properties of the correction term that allow to interpret it as a quality index for a small-dimension representation of a functional data; in Section 3 a non-parametric estimate is introduced and an algorithm to select the dimensionality at x is described; finally, in Section 4 an application illustrates the advantages in using such an approach. More theoretical and computational details can be found in [1].

2 The correction factor in the SmBP factorization as quality index

This section collects some theoretical results that justify the use of the correction factor $C_d(x,h)$ as a measure of the quality in approximating x by means of a d-dimensional representation.

By definition, the correction term is:

$$C_d(x,h) = \mathbb{E}\left[\left(\left(1 - \frac{\|\Pi_d^{\perp}(X-x)\|^2}{h^2}\right)\mathbb{I}_{\left\{\|\Pi_d^{\perp}(X-x)\|^2 \leq h^2\right\}}\right)^{d/2} \middle| \Pi_d x\right] \tag{3}$$

where Π_d denotes the projector onto $\mathscr{H}_d = span\{v_1,\ldots,v_d\}$, and $\Pi_d^{\perp}$ is its orthogonal projector. Note that $C_d(x,h) \in (0,1]$.

It can be proven that, varying $x \in \mathscr{H}$, the term $C_d(x,h)$ reaches its maximum over $\mathscr{H}_d$, as stated below.

Proposition 1. *Fix $h > 0$ and a strictly positive integer d, and suppose that the assumptions that guarantee the existence of the factorization (2) hold. Assume that the r.v. $((1 - \|\Pi_d^{\perp}(X-x)\|^2/h^2)\mathbb{I}_{\left\{\|\Pi_d^{\perp}(X-x)\|^2 \leq h^2\right\}})^{d/2}$ is uncorrelated with $\{\Pi_d X = \Pi_d x\}$. Then, $C_d(x,h)$ admits a maximum $M_d(h)$ over $\mathscr{H}$ and it is achieved for any $x \in \mathscr{H}_d$.*

In other words, the maximum of $C_d(x,h)$ is reached for any x such that $\langle x, v_j\rangle = 0$ for any $j > d$. As a consequence, $C_d(x,h)$ helps to identify d to represent in small-dimension x: the closer $C_d(x,h)$ and $M_d(h)$ are, the more accurate the representation of x over the subspace $\mathscr{H}_d$ is, and adding further dimensions does not improve the quality of the representation.

Finally, the following characterization result can be stated:

Proposition 2. *Let X' be an independent copy of X, d a strictly positive integer and $h > 0$. Then the following statements are equivalent:*

i) $\mathbb{E}[C_d(X',h)] = 1$;
ii) $C_d(X',h) = 1$ *a.s.*;
iii) $\lambda_{d+1} = 0$;
iv) the process admits the d–dimensional representation $X = \sum_{j=1}^{d} \xi_j v_j$ *a.s.*.

3 A dimensionality selection algorithm

In order to make possible to use in practice the ideas described in the previous section, estimates of $C_d(x,h)$ and $M_d(h)$ must be defined. In this perspective, let $X_1,\ldots,X_n$ be a sample drawn from X. A Nadaraya–Watson type estimate of $C_d(x,h)$ is given by

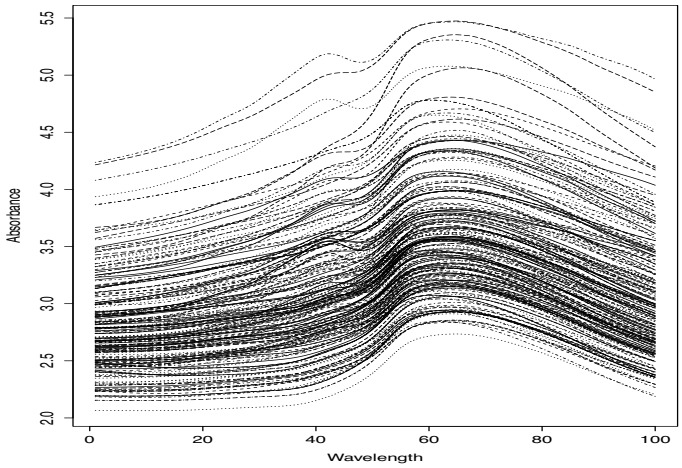


Fig. 1 The Tecator dataset.

$$\widehat{C}_{d,n}(x,h) = \sum_{i=1}^{n} \left(\left(1 - \frac{\|\widehat{\Pi}_d^{\perp}(X_i - x)\|^2}{h^2} \right) \mathbb{I}_{\left\{\|\widehat{\Pi}_d^{\perp}(X_i-x)\|^2 \le h^2\right\}} \right)^{d/2} \times \\ \times \frac{K(\|\widehat{\Pi}_d(X_i - x)\|/b)}{\sum_j K(\|\widehat{\Pi}_d(X_j - x)\|/b)}$$

where b is a bandwidth (in general depending on n), K a suitable kernel, $\widehat{\Pi}_d$ and $\widehat{\Pi}_d^{\perp}$ are the empirical estimates of the projectors Π_d and $\Pi_d^{\perp}$. For any d, an estimate of the upper bound $M_d(h)$ is provided by $\widehat{M}_{d,n}(h) = \max_i \widehat{C}_{d,n}(X_i,h)$.

At this stage, a procedure to select the local dimension can be defined. Given $\{\chi_j, j = 1,\dots,N\}$, possibly coincident with the sample, for each χ_j the local dimension $d_j^{\star}$, that should be used in (1), is selected as the smallest d for which $\widehat{C}_{n,d}(\chi_j,h)$ is close enough to $\widehat{M}_{d,n}(h)$. The proximity to this bound is quantified by considering if the relative measure $\widehat{C}_{n,d}(\chi_j,h)/\widehat{M}_{d,n}(h)$ is larger or smaller than a threshold $\alpha \in (0,1)$ suitably selected.

4 Application

To illustrate the performances of the local dimensionality selection algorithm described above, it is applied to the so–called Tecator dataset. It consists of near-infrared absorbance spectra of 215 chopped pieces of meat, discretized on 100 equally spaced wavelengths in the range $852-1050$ nm (these curves are visualized in the top panel of Figure 1).

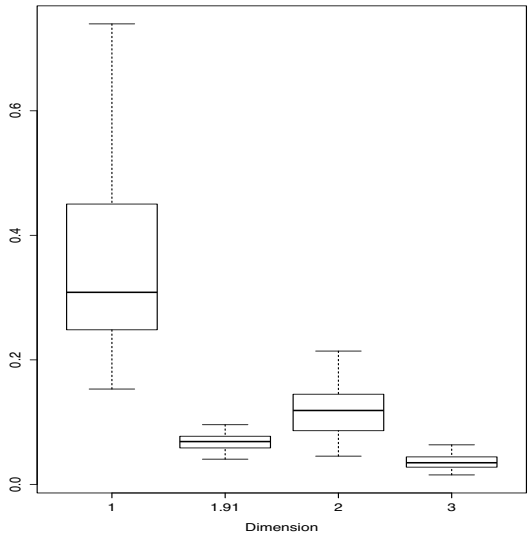


Fig. 2 Empirical distributions of the means of the ISEs, varying the dimension. The second (from the left) boxplot corresponds to the ISE computed when local dimensions are used.

The curves are rather smooth and a vertical shift appears: a good representation of them by using (1) can be obtained with the global dimension $d = 3$, that corresponds to a fraction of explained variance equal to 99.9%.

Clearly, that dimension could be too large for some of the curves in the dataset, and a parsimonious representations based on the algorithm, with a similar precision, can be adopted. In practice, the dataset is randomly split in two parts: the first one, containing 200 curves, is used to estimate the bounds $M_d(h)$ for $d = 1,\ldots,5$, whereas the remaining part $\left\{\chi_j, j = 1,\ldots,15\right\}$ is used to evaluate the local dimensions $d_j^\star$. The used kernel is the Epanechnikov one whereas the bandwidth is selected as the 10%–quantile of the distances between the curves in the training set projected by means of $\widehat{\Pi}_d$. The goodness of the approximation of χ_j by means of its $d_j^\star$-dimensional approximation $\chi_j^\star$ is measured by means of the Integrated Square Error (ISE), that is $\int_0^1 \left(\chi_j(t) - \chi_j^\star(t)\right)^2 dt$. The CV procedure is repeated 100 times: in each replication, the means of ISEs calculated by using both a global dimension d and the local one are computed, as well as the mean dimension $\overline{d}_m^\star$.

The choice of α is carried out by comparing the ISE behaviour varying α over a grid of possible values and the ISE obtained by using a global dimension. For this dataset, a good compromise is $\alpha = 0.87$ for which one gets $\overline{d}_m^\star = 1.91$ with a mean ISE equals to 0.068; if one compares this error with the one obtained when a global dimension is adopted, it is evident that the customization produces an efficient representation (see the distributions of the mean ISEs, multiplied by 100, in Figure 2).

References

1. Aubin, J.B., Bongiorno, E.G., Goia, A.: The correction term in a Small-Ball Probability factorization for random curves. Journal of Multivariate Analysis, In Press (2022)
2. Bongiorno, E.G., Goia, A.: Some insights about the Small Ball Probability factorization for Hilbert random elements. Statistica Sinica, **27**, 1949–1965 (2017)
3. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer Series in Statistics. Springer, New York (2006)
4. Horvath, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York (2012)
5. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd Edition. Springer Series in Statistics. Springer, New York (2005)

Mixed-effects high-dimensional multivariate regression via group-lasso regularization

Regressione multivariata con effetti misti per dati ad alta dimensionalità: un approccio con regolarizzazione di tipo group-lasso

Francesca Ieva, Andrea Cappozzo, and Giovanni Fiorito

Abstract Linear mixed modeling is a well-established technique widely employed when observations possess a grouping structure. Nonetheless, this standard methodology is no longer applicable when the learning framework encompasses a multivariate response and high-dimensional predictors. To overcome these issues, in the present paper a penalized estimation procedure for multivariate linear mixed-effects models (MLMM) is introduced. In details, we propose to regularize the likelihood via a group-lasso penalty, forcing only a subset of the estimated parameters to be preserved across all components of the multivariate response. The methodology is employed to develop novel surrogate biomarkers for cardiovascular risk factors, such as lipids and blood pressure, from whole-genome DNA methylation data in a multi-center study. The described methodology performs better than current state-of-art alternatives in predicting a multivariate continuous outcome.

Abstract *I modelli ad effetti misti sono ampiamente utilizzati nell'analisi di dati che possiedono una struttura a gruppi. Tuttavia, tale metodologia non è applicabile in contesti dove la variabile risposta è multidimensionale ed il numero di regressori elevato. Nel proporre una soluzione ai sopracitati problemi, nel presente lavoro viene introdotta una procedura di stima penalizzata per modelli ad effetti misti con risposta multivariata. In dettaglio, si propone di regolarizzare la verosimiglianza tramite una penalità di tipo group-lasso, forzando solo un sottoinsieme dei parametri stimati ad essere diverso da 0 per ogni componente della variabile risposta. La metodologia proposta viene poi utilizzata per creare nuovi surrogati per fattori di rischio cardiovascolare, come lipidi e pressione sanguigna, dai dati di metilazione del DNA dell'intero genoma in uno studio multicentrico. L'analisi così condotta dimostra risultati migliori rispetto alle attuali alternative nella previsione di un outcome continuo multivariato.*

Francesca Ieva, Andrea Cappozzo
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano, e-mail: francesca.ieva@polimi.it, andrea.cappozzo@polimi.it

Giovanni Fiorito
Department of Biomedical Sciences, Università di Sassari e-mail: gfiorito@uniss.it

Key words: Mixed-effects models, Multivariate regression, group-lasso penalty, penalized estimation

1 Introduction and motivation

Multivariate regression performs joint learning of a multidimensional response on a common set of predictors. When samples possess a hierarchical/temporal structure, data independence cannot be assumed a-priori and thus a Multivariate Mixed-Effects Model (MLMM) must be adopted [5]. An MLMM framework thus allows for the inclusion of a grouping structure within the model specification, a situation that often arises in multi-centric and/or longitudinal studies. With the advent of modern technologies, it is more and more common nowadays that in such studies a huge number of features is recorded, often greatly exceeding the available sample size. To this extent, regularization methods based on penalized estimation have been fruitfully adopted to overcome the resulting over-parameterization issue [7]. In particular, for univariate mixed-effects models, ℓ_1- penalization schemes have been devised to perform selection of fixed effects when dealing with high-dimensional data [4, 3]. By suitably leveraging the methodology proposed in [3], we extend it to the multivariate response framework including a group-lasso penalty in the model specification.

The remainder of the paper proceeds as follows: in Section 2 we introduce our new proposal and we discuss its main methodological aspects. Section 3 presents an application of our model in creating surrogate scores based on blood DNA methylation. Section 4 summarizes the novel contributions and highlights future research directions.

2 Group-lasso regularized mixed-effects multivariate regression

In an MLMM framework, the data-generating process for the n_j units in group j, with $\sum_{j=1}^{J} n_j = N$ and J the total number of groups, is assumed to be as follows:

$$\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{B} + \boldsymbol{Z}_j\boldsymbol{\Lambda}_j + \boldsymbol{E}_j, \tag{1}$$

where $\boldsymbol{Y}_j$, $\boldsymbol{X}_j$, $\boldsymbol{Z}_j$ respectively define the response, fixed and random effects design matrices. Further, $\boldsymbol{B}$ denotes the matrix of fixed coefficients, $\boldsymbol{\Lambda}_j$ the matrix of random effects in group j and $\boldsymbol{E}_j$ the group specific error term. The following distributions are assumed for the random quantities in (1):

$$\operatorname{vec}(\boldsymbol{\Lambda}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \quad \operatorname{vec}(\boldsymbol{E}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}), \quad j = 1, \ldots, J$$

with $\operatorname{vec}(\cdot)$ denoting the vec operator, $\boldsymbol{\Psi}$ is a positive semidefinite matrix incorporating variations and covariations between the responses and the random effects and

$\boldsymbol{\Sigma}$ is a covariance matrix capturing column-wise dependence in the multivariate error term $\boldsymbol{E}_j$. Thereupon, the distribution of the vectorized response can be written as follows:

$$\operatorname{vec}(\boldsymbol{Y}_j) \sim N\left((\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right), \quad j = 1, \ldots, J.$$

When dealing with high-dimensional data, the number of regressors (i.e., the rows of matrix $\boldsymbol{B}$) is generally much larger than the sample size N. Therefore, in order to still be able to make sensible inference on the parameters $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$, we propose to maximize the following penalized log-likelihood:

$$\begin{aligned}\ell_{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \log \phi\left(\operatorname{vec}(\boldsymbol{Y}_j), (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right) + \\ - \lambda \left[(1-\alpha)\sum_{c=1}^{r}\sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{l=2}^{p} ||\boldsymbol{b}_{l\cdot}||_2\right],\end{aligned} \tag{2}$$

where b_{lc} and $\boldsymbol{b}_{l\cdot}$ denote the element in position (l,c) and the l-th row of matrix $\boldsymbol{B}$, respectively. The penalty in (2) behaves like the lasso but on a whole group of coefficients. In details, for each covariate, the estimated parameters are either all zero or none are zero, and this behavior is preserved across all components of the response variable. This characteristic is particularly desirable when it comes to variable selection in multivariate regression, since features that are jointly related to the multidimensional response are automatically identified. The amount of shrinkage is determined by the penalty factor λ, whilst the mixing parameter α controls the weight associated to ridge and group-lasso regularizers. Maximization of (2) is performed via a tailored EM-type algorithm [1], in which standard fixed-effects routines are conveniently exploited within the M-step.

The devised framework is employed to build a multidimensional predictor of systolic and diastolic blood pressure, LDL and HDL cholesterol based on blood DNA methylation (DNAm): results are reported in the next section.

3 Application to DNAm biomarkers creation

DNAm biomarkers are obtained by regressing blood measured quantities (response variables) on methylation levels within CpG sites in the DNA sequence (dependent variables) [6]. The aim of this section is to build a multivariate DNAm biomarker for cardiovascular risk factors and comorbidities, considering Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), High Density Lipoprotein (HDL) and Low Density Lipoprotein (LDL) as responses, regressing them onto 13449 CpG sites (top 1% p-value based ranking) adjusting for sex and age. The employed dataset comes from the Italian component of the European Prospective Investigation into

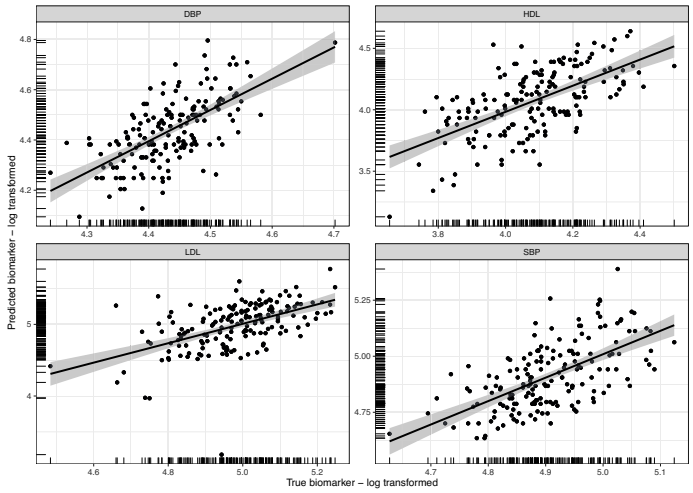


Fig. 1 Observed vs fitted scatterplots for the estimated biomarkers, namely log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and Systolic Blood Pressure (SBP), EPIC Italy test set. Linearly smoothed conditional means and associated standard deviations are superimposed in each facet.

Cancer and Nutrition (EPIC) study [2], comprised of $J = 4$ geographical sub-cohorts identified by the centre of recruitment. We employ $N_{tr} = 401$ training samples to fit the model in (2) including a random intercept component, validating its performance on $N_{te} = 173$ test units. The root mean squared error (RMSE), computed on the test set for the four-dimensional response, is reported in Table 1. Together with our proposal (denoted as MLMM Group-lasso in the table), results for two competing methods are reported, namely fixed-effect group lasso and univariate elastic-net [8]. For each method, the penalty factor λ was tuned via 10-fold CV on the training set, while the the mixing parameter α was kept fixed and equal to 0.5.

As it clearly stands out from Table 1, our proposal achieves better predictive performances for all components in the response variable with respect to the competing models. The reason behind this result is two-fold. On one hand MLMM Group-lasso performs better than its fixed-effects counterpart as the heterogeneity induced by the centre of recruitment is properly taken into account by means of a random intercept. On the other hand, solving the four regression problems jointly and imposing a group structure on the coefficients leads to better prediction performance than fitting four univariate models separately as done for the elastic-net procedure. The good predictive performance of the proposed model is highlighted in Figure 1, where we report for each biomarker the observed vs fitted scatterplots. All components of the response exhibits positive linear correlation between measured and predicted values in the test set, with Pearson's correlation coefficients always higher than 0.5.

The employment of the MLMM Group-lasso not only produces moderate improvements in terms of prediction accuracy, but it is also supported by biological

Table 1 Root Mean Squared Error (RMSE) for different penalized regression models, EPIC Italy test set. Bold numbers indicate lowest RMSE for each component of the four dimensional response.

Model	DBP	HDL	LDL	SBP
MLMM Group-lasso	**0.102**	**0.2139**	**0.278**	**0.1172**
Group-lasso	0.112	0.2238	0.286	0.1229
Univariate elastic-net	0.1064	0.2292	0.2884	0.1271

reasons. In fact, the pleiotropic effect suggests that multiple correlated phenotypes will likely affect the same set of CpG sites, motivating the adoption of a group-lasso penalty. Furthermore, DNAm biomarkers creation stands on the rationale that the resulting surrogate should be study-invariant: by incorporating a random intercept in the model specification the center effect can still be captured, while maintaining generalizability of the method to external cohorts.

4 Conclusion

The present work has introduced a novel penalized mixed-effects multivariate regression framework, able to model a multidimensional response with high-dimensional covariates and grouped data structure. By means of a group-lasso regularizer, we have achieved excellent predictive accuracy when creating a DNAm surrogate of cardiovascular risk factors, outperforming state-of-the-art alternatives. Such surrogates possess some advantages over their blood-measured counterparts, as they can directly take into account genetic susceptibility and subject specific response to risk factors.

In the devised framework we have implicitly assumed low-dimensionality in the response variable. A direction for future research may concern the inclusion of custom penalties to cope with situations in which both the response and the design matrix are high-dimensional. Feasible solutions are currently being investigated and they will be the object of future work.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. R. Stat. Soc. Ser. B **39**(1), 1–22 (1977) doi:10.1111/j.2517-6161.1977.tb01600.x
2. Riboli, E., Hunt, K., Slimani, N., Ferrari, P., Norat, T., Fahey, M., Charrondière, U., Hémon, B., Casagrande, C., Vignat, J., Overvad, K., Tjønneland, A., Clavel-Chapelon, F., Thiébaut, A., Wahrendorf, J., Boeing, H., Trichopoulos, D., Trichopoulou, A., Vineis, P., Palli, D., Bueno-de Mesquita, H., Peeters, P., Lund, E., Engeset, D., González, C., Barricarte, A., Berglund, G., Hallmans, G., Day, N., Key, T., Kaaks, R., Saracci, R.: European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. **5**(6b), 1113–1124 (2002) doi:10.1079/PHN2002394

3. Rohart, F., San Cristobal, M., Laurent, B.: Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. Comput. Stat. Data Anal. **80**, 209–222 (2014) doi:10.1016/j.csda.2014.06.022
4. Schelldorfer, J., Bühlmann, P., De Geer, S.V.: Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. Scand. J. Stat. **38**(2), 197–214 (2011) doi:10.1111/j.1467-9469.2011.00740.x
5. Shah, A., Laird, N., Schoenfeld, D.: A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. J. Am. Stat. Assoc. **92**(438), 775–779 (1997) doi:10.1080/01621459.1997.10474030
6. Singal, R., Ginder, G.D.: DNA Methylation. Blood **93**(12), 4059–4070 (1999) doi:10.1182/blood.V93.12.4059
7. Vinga, S.: Structured sparsity regularization for analyzing high-dimensional omics data. Brief. Bioinform. **22**(1), 77–87 (2021) doi:10.1093/bib/bbaa122
8. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. **67**(5), 768–768 (2005) doi:10.1111/j.1467-9868.2005.00527.x

The integration of immigrants in Italy: a multidimensional perspective

Albanian, Romanian and Italian women's fertility intentions: a comparative analysis among migrants, stayers and natives

Le intenzioni di fecondità delle donne albanesi, rumene e italiane: un'analisi comparativa tra migranti, non-migranti e nativi

Thaís García-Pereiro and Anna Paterno

Abstract This paper analyses fertility intentions of Albanian and Romanian women in Italy following an origin-destination perspective. Short-term intentions to have a child of Albanian and Romanian women living in Italy are compared not only to those of Italian women (host country) but also to those of stayers in sending countries (Albania, Romania). To account for differences and similarities among groups, the final dataset mixes microdata coming from several sources (FFS and SCIF for Italy, DHS for Albania, Eurobarometer for Romania). Ordinal regression models on fertility intentions of natives, migrants and stayers are aimed at testing theoretical models of adaptation and socialization, while controlling for independent variables that literature has proven important determinants of the intention to have one or another child.

Abstract *Questo articolo analizza le intenzioni di fecondità delle donne albanesi e romene immigrate in Italia adottando una prospettiva che considera sia l'origine sia la destinazione. Si comparano le intenzioni di avere un figlio a breve termine delle donne albanesi e rumene che vivono in Italia non solo con quelle delle donne italiane (paese ospitante) ma anche con quelle delle donne rimaste nei paesi di origine (Albania, Romania). Per tenere conto delle differenze e delle analogie tra i gruppi, il dataset finale unisce microdati provenienti da diverse fonti (FFS e SCIF per l'Italia,*

[1] Thaís García-Pereiro, Università degli Studi di Bari; email: t.garcia.pereiro@uniba.it

Anna Paterno, Università degli Studi di Bari; email: anna.paterno@uniba.it

The authors acknowledge the financial support provided by the MiuR-PRIN Grant N. 2017W5B55Y (GDR project "The Great Demographic Recession", PI: Daniele Vignoli).

DHS per l'Albania, Eurobarometro per la Romania). I modelli di regressione ordinale sulle intenzioni di fecondità di native, migranti e non-migranti sono volti a testare la validità delle teorie note come "adaptation" e "socialization", controllando al contempo variabili indipendenti che la letteratura ha dimostrato importanti determinanti delle intenzioni di avere uno o un altro figlio.

Key words: fertility intentions, natives, migrants, stayers, Italy, data linkage.

Introduction

The literature on migrant fertility in Europe has significantly grown in recent years (De Valk and Liefbroer, 2007; Mussino and Strozza, 2012; Kleinepier et al., 2015; Robards and Berrington, 2016; Impicciatore et al., 2020; Tønnessen and Mussino, 2020; Lindström et al., 2020). However, as stated by Puur et al. (2018), studies on migrant fertility have neither sufficiently stressed differences between migrants' and natives' reproductive decision-making processes nor considered the link between fertility intentions and the migratory background. There are some empirical analyses about migrants' fertility intentions (Carlsson, 2018; Mussino et al., 2021; Alderotti and Trappolini, 2021) but, to the best of our knowledge, only the study by Puur et al. (2018) intentionally developed an origin-destination perspective comparing Russian women living in Estonia, Estonians in the host country and Russians in the country of origin, applying the one-origin and one-destination approach.

This paper is aimed at contributing to the existing literature by filling the need of studies on fertility intentions using a comparative origin-destination perspective while disentangling and quantifying the influence of adaptation and socialization theories on fertility intentions of natives, Albanian and Romanian migrants (long and short term), and stayers (non-migrants).

Theoretical background and hypotheses in brief

Our aim is to verify if the adaptation and socialization theoretical approaches fit in order to describe differences on fertility intentions between Italian women, migrants (Albanian and Romanian women) in Italy and non-migrants (stayers) in their countries of origin (Albania and Romania). Therefore, we formulate the following research hypotheses:

RH1 Adaptation. According to the adaptation perspective, migrants' reproductive choices in the host countries tend to become similar to those of natives over time (Kulu 2005, Gabrielli et al. 2007). If this hypothesis prevails, long-term migrants' fertility intentions will be like those of comparable Italians.

RH2 Socialization. For the socialization perspective, the social environment experienced during childhood strongly impacts future reproductive decisions. Thus,

socialization norms and values experienced in the country of origin will tend to prevail (Andersson, 2004; Kulu and Milewski, 2007). If this hypothesis prevails, fertility intentions of both short-term and long-term migrants will resemble those of comparable stayers and differ from those of Italians.

We are fully aware of the importance of other theoretical approaches (e.g. the selection and disruption theories) in explaining reproductive behaviours following migration (Kulu, 2005; Milewski, 2010; Mussino and Strozza, 2012; Wolf and Mulder, 2019). We were not able to test for selection given data availability, while, regarding disruption, we are planning to search for differences considering recently arrived migrants in a successive phase of our research. We also consider that any combination of our research hypotheses can be verified, given that adaptation and socialization might not be mutually exclusive theories.

Data and methods

Data were drawn from several sources.

1. Natives (n = 4,276): Families, Social Subjects and Life Cycle (FSS) survey, conducted in 2016 by the Italian Institute of Statistics (ISTAT),
2. Migrants (Albanian and Romanian women living in Italy) (n = 2,476): Social Condition and Integration of Foreign Citizens (SCIF) survey again carried out by ISTAT between 2011-2012.
3. Albanian stayers (n = 7,852): last available Demographic and Health Survey (DHS) conducted in 2017/2018 by the Albanian Institute of Statistics (INSTAT).
4. Romanian stayers (n = 210): Eurobarometer 75.4 survey conducted in 2011.

Only comparable information was harmonized in one unique dataset. The final sample was restricted to 14,814 women aged 18-44 due to data limitations and models included harmonized variables.

Our dependent variable is respondents' intentions to have a child within the next three years coded in: intended to have a(nother) child, undecided, and not intended to have a(nother) child. Migrants were further distinguished according to the length of their stay in Italy in two categories (recent migrants: ten years or less since migration; and long-term migrants: more than 10 years). Our main independent variable is a combination of the migratory background and the length of stay of migrant women, coded 1 for Italian, 2 for Albanian who are long term migrants, 3 for Albanian who are short term migrants,4 for Romanian who are long term migrants, 5 for Romanian who are short term migrants, and 6 and 7 respectively for Albanian and Romanian stayers.

We estimated ordinal regression models considering migratory background as the main independent variable and other independent variables already identified by the literature as important determinants of fertility intentions (Carlsson, 2018; Puur et al., 2018; Mussino et al., 2021): age groups (18-24, 25-29, 30-34, 35-39, 40-44), parity

(childless, 1, 2 and 3 or more), partnership status (never married, married, not married), educational attainment (primary or less, secondary[1], tertiary) and labour market status (employed, unemployed, inactive). For a more accurate interpretation of our results, we also computed adjusted predictions for prototypical cases (by migratory background, at mean values). Descriptive statistics are shown in Table 1.

Table 1: *Descriptive statistics of dependent and independent variables included in empirical analyses.*

Independent variables		Fertility intentions		
		Intended	Undecided	Not intended
Migratory background	Italian	8.67 (7.3)	46.26 (47.05)	28.76 (45.42)
	Albanian LongTmig	0.81 (9.12)	3.06 (40.43)	2.46 (50.46)
	Albanian ShortTmig	2.29 (21.52)	4.46 (49.11)	1.72 (29.37)
	Romanian LongTmig	0.94 (12.11)	2.74 (41.18)	2.00 (46.71)
	Romanian ShortTmig	6.60 (16.75)	16.26 (48.33)	7.57 (39.43)
	Albanian stayer	80.12 (37.89)	26.35 (18.20)	55.26 (71.90)
	Romanian stayer	0.57 (10.00)	0.87 (29.36)	2.24 (45.58)
Age groups	18-24	44.20 (48.51)	23.22 (29.86)	10.84 (21.64)
	25-29	27.07 (39.46)	21.75 (37.14)	8.83 (23.40)
	30-34	17.40 (25.06)	21.75 (36.70)	14.60 (38.25)
	35-39	8.32 (11.80)	18.33 (30.44)	22.39 (57.75)
	40-45	3.02 (3.04)	14.95 (17.62)	43.34 (79.34)
Parity	Childless	59.33 (39.15)	53.21 (41.12)	16.44 (19.73)
	1	29.49 (36.55)	23.06 (33.48)	13.30 (29.97)
	2	9.80 (8.86)	19.11 (20.23)	43.13 (70.90)
	3+	1.37 (2.45)	4.62 (9.64)	27.13 (87.91)
Partnership status	Never married	43.13 (31.95)	51.12 (44.44)	17.56 (23.61)
	Married	51.95 (21.98)	42.01 (20.86)	74.45 (57.16)
	Not married	4.91 (17.89)	6.87 (29.35)	7.99 (52.77)
Labour market status	Employed	37.71 (21.27)	49.18 (32.49)	45.08 (46.24)
	Unemployed and other situation	62.29 (28.10)	50.82 (26.85)	54.92 (45.05)
Educational attainment	Primary or less	25.32 (25.14)	12.35 (14.36)	33.50 (60.50)
	Secondary	40.24 (19.34)	61.99 (34.90)	52.34 (45.75)
	Tertiary	34.45 (38.17)	25.66 (33.30)	14.16 (28.53)
Source	FSS2016_IT	8.67 (7.53)	46.26 (47.05)	28.76 (45.42)
	SCIF2011/12_IT	10.64 (15.95)	26.51 (46.57)	13.74 (37.48)
	DHS2017_AL	80.12 (37.89)	26.35 (14.60)	55.26 (47.52)
	EB2011_RO	0.57 (10.00)	0.87 (18.10)	2.24 (71.90)
N	*14,814*	*3,713*	*4,349*	*6,752*

Notes: column percentages, (row percentages).

Preliminary results

In the first part of this section, we focus our attention on the association between our main independent variable of interest, the migratory background, and fertility intentions[2]. Table 2 reports average marginal effects (AMEs) of native, migrant (short

[1] DHS data did not allow to further distinguish among levels of secondary education.

[2] According to our results, control variables positively affecting the intention to have another child are having a partner and holding a higher educational level, while this intention is negatively affected by increasing age, parity and being unemployed. Results are not shown here but available upon request.

and long term) and stayer women on the probability of declaring a specific category of their short-term fertility intentions.

In general, results show that migrant women are more likely to intend to have a(nother) child than Italian women. Considering the greatest differences between migrants and Italian women, Albanian short-term migrants were almost 7% more likely to want to have a child (first included) than Italians. The probability of wanting a child was always higher than Italians, but considerably lower than for Albanian short-term migrants, for Albanian long-term migrants, Romanian short-term migrants and Romanian long-term migrants, in this order. Albanian non-migrants were around 18% more prone to intend to have a child in the next three years than Italian women, while Romanian stayers were 4% less likely than native women.

Table 2: *Average marginal effects coming from ordinal regression models on fertility intentions by migratory background.*

Migratory background	Fertility intentions	AME	sig.
Ref. Italian			
Albanian LongTmig	Not intended	-.108	***
	Undecided	.073	***
	Intended	.035	***
Albanian ShortTmig	Not intended	-.189	***
	Undecided	.120	***
	Intended	.069	***
Romanian LongTmig	Not intended	-.076	**
	Undecided	.052	**
	Intended	.023	**
Romanian ShortTmig	Not intended	-.092	***
	Undecided	.063	***
	Intended	.029	***
Albanian stayer	Not intended	-.351	***
	Undecided	.176	***
	Intended	.176	***
Romanian stayer	Not intended	.191	***
	Undecided	-.148	***
	Intended	-.043	***

To directly compare all migratory categories among them, we computed adjusted predictions for prototypical cases, holding control variables at their mean values as presented in Figure 1. As aimed at comparing seven different groups, we plotted prediction for one outcome only: being intended to have a (or another) child.

Focusing on Albanian women, the intentions to have a child differ between long-term and short-term migrants, with a probability around 11% and 15%, respectively. For Albanian stayers the likelihood of wanting another child is much higher than for migrants, being approximately 24%. All these figures are higher than those observed among Italian women and Romanian migrants.

Turning the attention to this last group, differences between short and long-term migrants become less evident (8% for the first and 9% for the second). However, the probability of being intended to have a child is significantly higher than the one observed among stayers (3-4%) but slightly higher than for Italian women (6%).

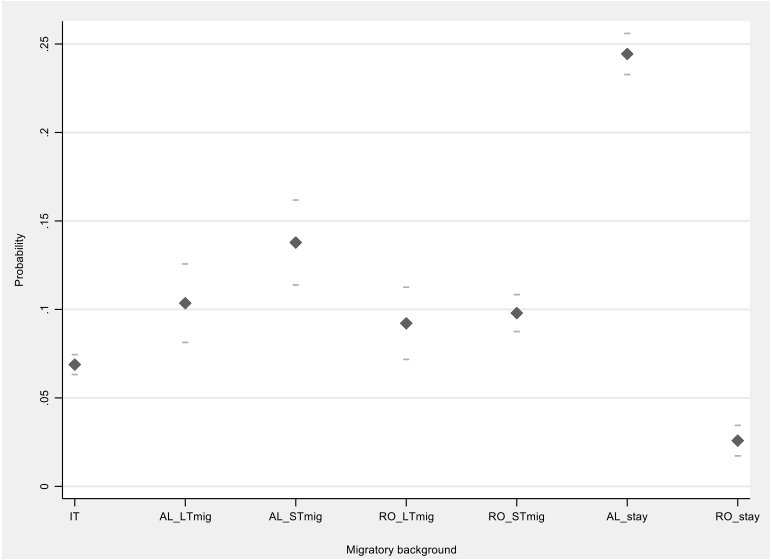


Figure 1: *Adjusted predictions for being intended to have a (another) child by migratory background (95% confidence intervals) while holding control variables at their mean values.*

Notes: own elaboration merged data.

Brief discussion of findings

As previous research considering fertility behaviors of migrant women in Italy (Mussino and Strozza 2012, Impicciatore et al. 2020, Mussino et al. 2021, Alderotti and Trappolini 2021), our results show that fertility intentions differ between natives and Albanian and Romanian migrants and stayers, but also among Albanian migrants, when considering time passed since migration.

The likelihood of wanting a (another child) of Romanian migrant women, independently of the duration of the stay, is similar to the likelihood of Italian women, but much higher than the one observed for stayers. These results are guiding us to find support for our first research hypothesis, *Adaptation,* that seems to fit well in the case of Romanian migrant women.

Mixed results are found for migrants coming from Albania. On one hand, Albanian migrant women are more likely to want to have a(nother) child than Italian (and Romanian) women, while Albanian stayers have the highest probability of wanting another child. This is supporting our second research hypothesis, *Socialization*, which seem to resemble -more accurately- fertility intentions of Albanian women when compared to stayers and natives. On the other, short-term migrants are more prone to want a child than long-term migrants, whose fertility intentions tend to resemble those of Italian women. This finding, instead, is in line with our first research hypothesis, *Adaptation*, given that fertility intentions of Albanian women tend to converge to those of Italian women as their length of stay in Italy increases.

These results might be reflecting current fertility behaviours of migrant populations, especially when we look at the values of group-specific Total Fertility Rates (TFR), but not only. In the case of Romanian women living in Italy, their TFR is higher than the one of women living in Romania, and this also holds for their fertility intentions. When analyzing migrants from this country, we must also take into account that their fertility intentions and outcomes might also be related to the elevated share of mixed unions, in particular, of Romanian women married to Italian men. Regarding Albanian women, their TFR tend to be much higher than the one registered among Romanian women, holding a level much more similar to the one of Italian women but lower than non-migrants period fertility (Mussino and Strozza, 2012).

References

1. Alderotti, G. and Trappolini, E.: Health status and fertility intentions among migrants. International migration, 00, 1--14 (2021)
2. Andersson, G.: Childbearing after migration: fertility patterns of foreign-born women in Sweden. International Migration Review, 38(1), 364--392 (2004)
3. Carlsson, E.: Fertility Intentions across Immigrant Generations in Sweden. Do Patterns of Adaptation Differ by Gender and Origin?. Comparative Population Studies, 43 (2018)
4. De Valk, H. A., and Liefbroer, A. C.: Timing preferences for women's family-life transitions: Intergenerational transmission among migrants and Dutch. Journal of Marriage and Family, 69(1), 190--206 (2007)
5. Gabrielli, G., Paterno, A., and White, M.: The impact of origin region and internal migration on Italian fertility. Demographic Research, 17, 705-740 (2007).
6. Impicciatore, R., Gabrielli, G., and Paterno, A.: Migrants' fertility in Italy: A comparison between origin and destination. European Journal of Population, 1--27 (2020).
7. Kleinepier, T., de Valk, H. A., and van Gaalen, R.: Life paths of migrants: A sequence analysis of Polish migrants' family life trajectories. European Journal of Population, 31(2), 155--179 (2015)
8. Kulu, H.: Migration and fertility: Competing hypotheses re-examined. European Journal of Population, 21(1), 51--87(2005)
9. Kulu, H., and Milewski, N.: Family change and migration in the life course: An introduction. Demographic research, 17, 567--590 (2007)
10. Milewski, N. (2010). Immigrant fertility in West Germany: Is there a socialization effect in transitions to second and third births?. European Journal of Population/Revue européenne de Démographie, 26(3), 297-323.
11. Mussino, E., and Strozza, S.: The fertility of immigrants after arrival: The Italian case. Demographic Research, 26, 99--130 (2012)
12. Mussino, E., Gabrielli, G., Ortensi, L. E., and Strozza, S.: Fertility Intentions Within a 3-Year Time Frame: a Comparison Between Migrant and Native Italian Women. Journal of International Migration and Integration, 1--28 (2021)
13. Puur, A., Vseviov, H., and Abuladze, L.: Fertility intentions and views on gender roles: Russian women in Estonia from an origin-destination perspective. Comparative Population Studies, 43 (2018)
14. Robards, J., and Berrington, A.: The fertility of recent migrants to England and Wales. Demographic Research, 34, 1037--1052 (2016)
15. Tønnessen, M., and Mussino, E.: Fertility patterns of migrants from low-fertility countries in Norway. Demographic Research, 42, 859--874 (2020)
16. Wolf, K., & Mulder, C. H. (2019). Comparing the fertility of Ghanaian migrants in Europe with nonmigrants in Ghana. Population, Space and Place, 25(2), e2171.

Does self-employment in the origin-country affect self-employment after migration? Evidence from Italy and Spain

Quale influenza del lavoro indipendente nel paese di origine sul lavoro indipendente dopo la migrazione? Il caso dell'Italia e della Spagna

Floriane Bolazzi and Ivana Fellini

Abstract According to the *home-country self-employment (HCSE) hypothesis*, immigrants' propensity for self-employment would depend on the diffusion of self-employment in home country. However, previous evidence is contrasting and the studies using information on the individual experience of self-employment, rather than the self-employment rate in the origin country, are very few. Using information on the last occupation before migration for a sample of immigrants in Italy and Spain, the article shows that pre-migration self-employment experience is not associated with higher chances of employability in the destination country, but it is associated with a higher probability of being self-employed after migration. In contrast with the HCSE hypothesis, we find no relationship with the self-employment rate in home country.

Abstract *Secondo l'ipotesi sull'influenza del lavoro indipendente del paese di origine, gli immigrati provenienti da paesi in cui il lavoro indipendente è molto diffuso sarebbero più probabilmente lavoratori indipendenti nel paese di destinazione. Tuttavia, le evidenze sono contrastanti e gli studi che considerano l'esperienza individuale di lavoro indipendente nel paese di origine, anziché il tasso di occupazione indipendente, sono pochi. Sfruttando l'informazione sull'ultima occupazione svolta nel paese di origine presso un campione di immigrati in Spagna e in Italia, l'analisi mostra che l'aver avuto un'esperienza di lavoro indipendente prima della migrazione non è associato a una maggiore probabilità di trovare un lavoro nel paese di destinazione ma è associato a una maggiore probabilità di avere un lavoro indipendente. Contrariamente a quanto previsto dall'ipotesi sull'influenza della diffusione del lavoro indipendente nel paese di origine, non sembra esserci relazione con il tasso di lavoro indipendente nel paese di origine.*

Key words: self-employment, migration, human capital, Italy, Spain

[1] Floriane Bolazzi, Università Milano-Bicocca; email: floriane.bolazzi@unimib.it
Ivana Fellini, Università Milano-Bicocca; email: ivana.fellini@unimib.it

1 Introduction

According to the *home-country self-employment hypothesis* (HCSE hypothesis, Yuengert, (1995); Fairlie and Meyer, (1996)), immigrants from countries where self-employment is widespread are more likely to become self-employed in destination countries. This hypothesis relies on the assumption that the entrepreneurial human capital immigrants have acquired in the origin country is a specific form of human capital making it more likely to enter self-employment after migration (Borjas (1986); Kloosterman and Rath, (2001)). The studies exploring this relationship, however, are few and evidence is controversial (Fairlie and Lofstrom, (2015)) partly due to the lack of individual data on work before migration. Most of the previous studies used the self-employment rate in the country of origin to approximate the individual experience of self-employment (Yuengert, (1995); Fairlie and Meyer, (1996); van Tubergen, (2005); Hammarstedt and Shukur, (2009)) while only a minority has used direct information on pre-migration experience of self-employment (Akee et al., (2013); Garcia-Diez and Perez-Villadoniga, (2013); Tibajev, (2019)).

The contribution of the paper is twofold. First, it extends the analysis to the influence of pre-migration self-employment on the employability of immigrants in the destination country. No previous study, to the best of our knowledge, has explored this relation, while a rich literature on entrepreneurship assumes stronger initiative and capacity of the self-employed and the entrepreneurs (Parker, (2004)). Exploring whether self-employment experiences in the origin country have positive effects on the employment chances of immigrants moving to a new country is a novelty. Second, the paper explores the relationship between self-employment in the origin and in the destination countries combining the use of individual and aggregate information on self-employment as very few studies do.

1.1 Research questions

In this paper, we aim to assess whether the experience of self-employment in the origin country affects immigrants' employment and self-employment chances in Italy and Spain, two Southern European receiving countries with similar structural and institutional characteristics. Indeed, both Italy and Spain traditionally record very high rates of self-employment among natives and are characterised by a very fragmented productive asset, a dualistic labour market where informal and irregular labour play an important role. In both countries, immigrants as well record high rates of self-employment, although lower than those of natives, contrary to what happens in older European immigration countries where self-employment is often a last resort option for immigrants and has scarcer appeal for natives (OECD, (2010, 2017)). Italy and Spain also share a similar migration history (Baldwin-Edwards, (2012); King et al., (2000)): they have become immigration countries only in recent decades, and both the characteristics of immigrants and the model of their economic

incorporation in the labour market has several common features (Fullin and Reyneri, (2011)).

More in detail, we want to test whether:

(i) the experience of individual self-employment in the origin country is associated with higher chances of being employed rather than unemployed after migration, under the hypothesis that the specific human capital provided by the experience of pre-migration self-employment (resourcefulness, knowledge of the labour market, work experience) positively affects immigrants' labour market incorporation in the receiving country, with differences due to different cultural contexts of origin.

(ii) the experience of individual self-employment before migration is associated with a higher probability of being self-employed rather than employee after migration, under the hypothesis that the specific human capital and the work experience provided by pre-migration self-employment positively affect the chances of entering self-employment also in the receiving country, with differences due to different cultural contexts of origin.

iii) the aggregate diffusion of self-employment in the origin country is positively associated with the probability of being self-employed after migration, as predicted by the HCSE hypothesis.

2 Data and methods

The analytical sample consists of 7,545 immigrants in Italy and 8,805 in Spain for whom information has been collected, respectively, by ISTAT (the Italian National Institute of Statistics) in 2011-2012 and by INE (the Spanish National Institute of Statistics) in 2007-2008. The two surveys collected retrospective information on labour market status, employment and occupation in the country of origin; the job at the time of the survey; the first job after arrival if the job at the time of the survey is different from the first one.

Immigrants are defined as foreign-born individuals who did not hold host-country citizenship at birth. We consider individuals aged between 15 and 55 at arrival and between 18 and 60 at the time of the interview who have had at least one work experience in their country of birth or in their country of origin if they moved from a country different from that of birth[2].

[2] We exclude transit countries where immigrants stayed for less than 3 months in Italy and less than 6 months in Spain. However, the large majority of immigrants moved directly from their country of birth without transiting to tiers countries (94.2% of the sample in Italy and 86.4% in Spain).

They are grouped in the following areas of origin: Western Countries of the European Union (EU15) and other highly developed countries (HD), Latin America, East Europe, Asia, Middle East and North Africa (MENA), and other African and Andean countries (Bolivia, Colombia, Ecuador and Peru). Although rarely used, the distinction between Latin America and Andean countries is relevant as Latin Americans present characteristics similar to immigrants from developed countries while Andeans do not (Reher and Requena, (2009); Fellini and Guetto, (2019)).

The analysis is two-fold: in the first step, we explore the influence of the individual experience of self-employment before migration on both the probability to be employed and to be self-employed in the receiving country. In the second step, we explore the relationship between the self-employment rate in the country of origin and the individual likelihood of being self-employed after migration.

In the first step, we estimate three logit models with different dependent variables and common independent and control variables. The *dependent variables* are:

- being employed rather than unemployed[3] at the time of the survey (model 1),
- being self-employed rather than employee at the time of the survey (model 2);
- being self-employed rather than employee in the first job after arrival (model 3)

The *independent variable* is the last occupation before migration: self-employed vs employee. The definition of self-employed in the country of origin (versus employee) has been harmonized for Italy and Spain considering all the situations different from wage employment. In the Spanish survey, this includes independent workers, own-account workers without employees, entrepreneurs with employees, members of a cooperative, family worker, or being in other professional situations. In the Italian survey, this includes entrepreneurs, independent professionals, own-account worker (with and without employees), family worker, member of cooperative, collaborators. The definition of self-employment in the receiving country has been instead restricted in order to exclude likely forms of bogus self-employment (i.e. collaborators in Italy).

In all models, the independent variable is interacted with immigrants' area of origin assuming that the effect of self-employment in origin could partly depends on the area of origin. All models are controlled for several individual characteristics. The *control variables* are: sex; age at migration (15-29, 30-44, 45 and more); education in the home country (low (primary), medium (lower and upper secondary) and high (tertiary)); the years spent in the receiving country since migration (up to 5 years; 6 to 10 years; 11 years or more); being married at the time of arrival (dummy); presence of children at the time of arrival (dummy); having found a job in the destination country before migration (dummy); degree of knowledge of the language of the destination country before migration (none, medium, good)[4]; reasons for migrating (four separate dummies[5]: economic, religious or political family, and other reasons); skill level of the last job before migration (ISCO 1-3; ISCO 4-8; ISCO 5-9); perceived working status at migration (employed; unemployed; inactive). In model 2, we also control for the status of the present job,

[3] Inactives are excluded.

[4] Only included in Italy because in Spain it highly correlates with the area of origin.

[5] Respondents could choose more than one option.

whether it coincides with the first job taken after arrival or a subsequent job. Indeed, more than half of the sample have changed jobs, at least once, before the survey (50.4% in Italy and 65.1% in Spain), that is the job at the time of the survey is not the first after arrival.

In the second step, we estimate a logit model (model 4) where the *dependent variable* is the probability of being self-employed at the time of the survey in the destination country and the *independent variable* is the self-employment rate in the birth country. The self-employment rate is computed as the average value of self-employment rate in the birth country over the period 1991-2011. The model controls, in a first specification (4a), for the home-country income level, based on the World Bank Classification of comparable measures of GDP per capita (2019). Due to the overlap of informal employment with many forms of self-employment in developing countries, in a second specification (4b), the model controls for the contribution of informal activities to GDP, computed as mean value, from 1991 to 2007.

3 Pre-migration self-employment and post-migration employment chances

In order to assess immigrants' employability in the destination country, we can consider their employment chances before migration and at the time of the survey. The information on the last job immigrants had in the origin country can be combined with the information on their labour market status at the time of migration (employed, unemployed or inactive), to assess their employment rates in the two moments and by work experience in the origin country (Table 1). The pre-migration employment rate is then the share of employed at migration on the total sample of immigrants who have had at least a work experience in the origin country and the employment rate at present is the share of employed at interview on the same sample. Overall, in Italy, the employment rate of immigrants increases after migration (77.1% at present vs 61.2% at migration), while in Spain it remains more or less the same (75.8% at present vs 76.9% at migration) (Table 1). In Italy, however, those immigrants whose last work experience before migration has been in self-employment record higher employment rates, both before migration and at present, compared to those who were employees before migration. The employment chances at present, however, have increased relatively more for those who were employees than for those who were self-employed (16 percentage points vs 3 percentage points). In Spain as well, those immigrants whose last work experience before migration has been in self-employment record higher employment rates, both at migration and at present, compared to employees before migration. The decrease in the employment chances after migration is similar in the two groups.

Table 1: Employment rate (%)* pre-migration and at present* by work experience pre-migration

	Italy		Spain	
	Pre-migration	**At present**	**Pre-migration**	**At present**
Employee	60.2	76.6	75.8	74.9
Self-employed	66.9	79.9	80.7	78.7
Total	61.2	77.1	76.9	75.8

* the share of employed at migration and at present on the total sample of immigrants who have had at least a work experience in the origin country.
Notes: weights applied

When controlling for the immigrants' heterogeneity (model 1), having been self-employed rather than employee in the last job before migration is not significantly associated with a higher probability to be employed at present (Table 2). Both in Italy and Spain, the difference in the probability of being employed rather than unemployed between immigrants who were self-employed and immigrants who were employees before migration not only is very small (2.6 percentage points) but also statistically not significant. With the only exception of immigrants from MENA countries, who have a significantly higher probability of being employed when they have experienced self-employment before migration, self-employment in origin seems not to affect immigrants' employment opportunities in the two receiving Southern European countries.

Table 2: Average Marginal Effects of having been self-employed before migration on the probability of being employed at present (overall and by area of origin)

	Italy		Spain	
	dy/dx	*Std. Err.*	*dy/dx*	*Std. Err*
Dep. Variable: employed at present				
All immigrants	0.026	0.014	0.026*	0.010
EU15+HD	0.003	0.030	0.048*	0.018
Latin	0.010	0.028	0.015	0.020
East-Europe	0.019	0.021	0.039	0.026
Asia	0.018	0.031	-0.006	0.038
MENA	0.092*	0.036	0.095*	0.035
Other Africa	0.009	0.044	0.047	0.063
Andeans	0.004	0.038	-0.010	0.018

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: based on logit estimates (95% Std. Err.) controlled for sex, age, marital status at arrival, having children at arrival, years since migration, language proficiency at arrival, reason for migration, job found before migration, skill level of occupation.

4 Pre-and post-migration self-employment experience

Overall, in Italy, 15.1% of immigrants who have had a work experience in the origin country were self-employed in the last job before migration while in Spain the figure is 21.6% (Table 3). In both countries, the self-employment rate after migration decreases to 11.5% for Italy and 13.0% for Spain. In both Italy and Spain, the share of self-employed at present is much larger for those who have been self-employed in their last job before migration. In Italy, the difference with respect to those who have been employees before migration is of 8 percentage points (18.3% vs 10.2%). In Spain, it is even larger with 12 percentage points of difference (22.4% vs 10.2%).

Table 3: Self-employment rate (%) at present* by work experience pre-migration

	Italy	**Spain**
Employee	10.2	10.2
Self-employed	18.3	22.4
Total	11.5	13.0

* the self-employment rate at present is the share of those who are self-employed on those who are employed at present.
Notes: weights applied

Table 4: Self-employment rates (%) before migration and at present* by area of origin

	Italy		**Spain**	
	Pre-migration	**At present**	**Pre-migration**	**At present**
EU15+HD	12.4	26.0	16.6	30.7
Latin	17.8	13.2	19.8	15.4
East-Europe	9.7	8.9	13.6	7.7
Asia	19.4	18.1	23.3	18.2
MENA	21.9	15.1	29.4	10.9
Other Africa	34.0	8.2	44.3	10.1
Andeans	26.7	5.4	25.7	8.4
Total	15.1	11.5	21.6	13.0

The pre-migration self-employed rate is the share of those who have had an experience of self-employment before migration on those who have had a work experience; the self-employment rate at present is the share of those who are self-employed on those who are employed at present.
Notes: weights applied

If we consider the share of those who had an experience of self-employment before migration and the share of those who are self-employed at present, not only there is a difference across immigrant groups, but also in the variation of the pre-and post-migration self-employment rates. In both countries, the rate of self-employment increases only for immigrants from EU15 & HD, while for

all the other groups it decreases (Table 4). In Italy, the decrease for Eastern Europeans and Asians is very small, it is large for Latins and immigrants from the MENA countries, and it is even larger for Andeans and immigrants from other African countries. In Spain, the decrease is larger for all, compared to Italy. The drop for Latins, East Europeans and Asians is relatively lower than for immigrants from other African countries, MENA countries and Andeans.

When controlling for the immigrants' heterogeneity (model 2), in both countries, the individual experience of self-employment before migration is associated with a significantly higher probability of being self-employed at present (Table 5). In Italy, the probability is 7.1 percentage points higher than for those who were employees before migration. In terms of predicted probabilities, immigrants who were self-employed before migration have 17.5% chance of being self-employed at present against 10.4% chance of those who were employees before migration. In Spain, the difference in the probability is 14.0 percentage points: immigrants who were self-employed before migration have 24.3% chance of being self-employed at present against 10.3% chance for those who were employees.

Table 5: Average marginal effects of pre-migration self-employment on the probability of being self-employed at present (all immigrants and by area of origin)

	Italy		**Spain**	
	dy/dx	*Std. Err.*	*dy/dx*	*std. Err*
Dep. Variable: self-employed at present				
All immigrants	0,071***	0,019	0,140***	0.016
EU15+HD	0.350*	0.129	0.313***	0.057
Latin	0.196	0.116	0.126***	0.036
East-Europe	0.044	0.030	0.203***	0.044
Asia	0.049	0.044	0.182	0.105
MENA	0.047	0.037	0.052	0.036
Other Africa	0.033	0.033	0.064	0.066
Andeans	0.108	0.079	0.068**	0.023

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: based on logit estimates (95% Std. Err.) controlled for sex, age, marital status at arrival, having children at arrival, years since migration, language proficiency at arrival, reason for migration, job found before migration, skill level of occupation, perceived working status at migration, status of present job (first, subsequent).

The association between pre-migration self-employment and the probability of being self-employed at present is positive for all areas of origin in both countries, but the statistical significance is beyond the 10% level for most areas in Italy. There is a clear and significant distinct pattern for EU15 and other HD immigrants whose probability of being self-employed at present is largely affected by pre-migration self-employment with, both in Italy and Spain, more than 30 percentage points of difference with respect to employees before migration. In Spain, the relationship is significant also for Latin, East-Europeans and Andeans. In Italy instead, the average marginal effects for other immigrants groups are not statistically significant. Among immigrants from other areas than EU15 and other HD countries, there seems to be

minor differences across area of origin. However, estimates do not allow assessing whether differences by areas of origin are actually relevant.

Table 6 presents a similar analysis using the first job after arrival, restricting previous analysis to the probability of being self-employed in the first job only[6]. The estimated coefficient of self-employment in origin for all immigrants is very close in magnitude than the coefficient in Table 5. Compared to model 2, the coefficient is smaller for EU15 and HD in Italy but almost the same in Spain, it is statistically significant for MENA, in both countries, and for Andeans in Italy. The major difference concerns immigrants from MENA countries suggesting that previous experience in self-employment of immigrants from MENA countries most likely increases their probability of being self-employed in the first job after arrival in Italy and Spain, but less certainly in subsequent jobs.

Table 6: Average marginal effects of pre-migration self-employment on the probability of being self-employed in the first job after arrival (all immigrants and by area of origin)

	Italy		**Spain**	
	dy/dx	*std err*	*dy/dx*	*std err*
Dep. Variable: self-employed in the first job				
All immigrants	0.068***	0.014	0.148***	0.015
EU15+HD	0.225**	0.110	0.386***	0.055
Latin	0.128	0.078	0.124***	0.029
East-Europe	0.031	0.019	0.184***	0.039
Asia	0.077	0.037	0.143	0.097
MENA	0.090***	0.033	0.088***	0.032
Other Africa	0.046	0.029	0.056	0.063
Andeans	0.131**	0.065	0.065***	0.019

* p < 0.05, ** p < 0.01, *** p < 0.001

Notes: based on logit estimates (95% conf. interval) controlled for sex, age, marital status at arrival, having children at arrival, years since migration, language proficiency at arrival, reason for migration, job found before migration, skill level of occupation, perceived working status at migration,

5 Self-employment diffusion in the origin-country and post-migration self-employment chances

While we have evidence of the association between the pre-and post-migration individual experience of self-employment, in both Italy and Spain, there is no evidence of an association between the self-employment rate in the country of origin[7] and the probability to be self-employed in the destination country. The

[6] Fist jobs correspond either to the same job at the time of the survey for those who still performed the first job they found after arrival or a different job for those who changed job at least once since arrival.

[7] The self-employment rate is based on the country of birth rather than the country of origin before entering the receiving country: 5.84% of the sample in Italy and 13.63% in Spain transited trough tier countries.

estimates of the association, controlling either for the income level of the origin country or for the size of the informal sector, is negligible and statistically not significant (Table 7). Coming from countries where self-employment is widespread seems not to increase the probability of being self-employed in the country of destination. Therefore, the self-employment rate is not a good proxy of the entrepreneurial human capital acquired through individual experiences of self-employment, contrary to what the HCSE hypothesis assumes[8].

The probability of being self-employed in the receiving country, however, varies significantly with the income level of the origin country. Lower levels of national income are associated with a lower probability of immigrants to become self-employed post-migration. Similarly, the size of the informal sector is negatively associated with the probability of being self-employed after migration.

These results support previous evidence showing that the probabilities of being self-employed increase after migration for immigrants from "Western countries", with a similar level of economic development to Spain and Italy, but not for immigrants from developing countries, although the latter have on average a higher self-employment rate in origin than the former (Table 3). Moreover, in Table 5, we observed that the influence of self-employment in origin on the probability of being self-employed in destination is higher for immigrants from EU15 and other HD countries.

Table 7: Probability of being self-employed at present (log-odds and std. err. for independent variable and country level controls)

	ITALY		**SPAIN**	
	[4a]	[4b]	[4a]	[4b]
Origin-country self-employment rate	0.00249	0.00222	-0.00431	0.00078
	(-0.00339)	(-0.004)	(-0.00408)	(-0.00415)
Country income level				
(ref. High income)				
Low income	-1.287		-2.613**	
	(-0.761)		(-0.795)	
Lower middle income	-0.651*		-0.830***	
	(-0.254)		(-0.2)	
Upper middle income	-0.354		-0.964***	
	(-0.206)		(-0.154)	
Informal sector (%GDP)		0.0453***		-0.0223**
		(-0.00837)		(-0.00689)
Observations	5689	5615	6226	6213

Log-odds. Standard errors in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
Notes: Logistic regressions controlled for sex, age, marital status at arrival, having children at arrival, years since migration, language proficiency at arrival, reason for migration, job found before migration, skill level of occupation, perceived working status at migration.

[8] The self-employment rate in the country of origin significantly correlates with the probability that immigrants of the sample have been self-employed before migration, thus partially excluding biases due to migration selection.

6 Concluding remarks

Compared to previous studies exploring the relationship between pre-and post-migration self-employment, this paper adds at least two novelties. First, not only it looks at the influence that pre-migration self-employment could have on self-employment after migration but also on the overall employment chances. Second, we consider two similar countries and the results are close, thus suggesting some soundness of the estimated relations.

Three main evidences have emerged from our analysis.

First, the analysis shows that the probability of being employed is not significantly associated with the previous experience of self-employment in the origin country. Therefore, the evidence does not support the hypothesis that the (entrepreneurial) human capital acquired in the origin country positively affects the employment chances in the destination countries Immigrants from MENA countries are the only exception in both Italy and Spain, suggesting potential self-selection among this group of immigrants (Tibajev, (2019)).

Second, we found that, overall, self-employment in origin is significantly associated with higher chances of being self-employed after migration. The self-employment specific human capital seems to positively affect the chances of becoming self-employed in the destination country. However, the relationship is strong and significant for "Western immigrants" while for the other immigrant group the magnitude is positive but lower and statistically uncertain. For "non-Western" immigrants, differences by areas of origin seem to not affect significantly the relationship between pre-and post-migration self-employment.

The level of human capital transferability can play a role in the difference between "Western" and "non-Western" immigrants. Coming from a country with a similar level of economic development and comparable labour market features, economic culture and social norms, could facilitate the transferability of the human capital consolidated in origin. Moreover, in developing countries, the nature of most of the self-employment jobs, performed as a survivalist strategy (e.g. street-vendors, porters, cleaners etc.), may not provide a real entrepreneurial human capital and may not favour its transmission. The language proximity could also be an advantage in terms of transferability as the Latin Americans case suggests.

Finally, no relationship between the probability of being self-employed in the receiving country and the self-employment rate in the origin country has been found, contrary to the expectations of the home-country self-employment hypothesis. If consider this evidence together with the previous one, we can suggest that while the diffusion of the self-employment rate does not influence self-employment chances after migration, the individual experience does. The evidence has at least two possible implications. First, using the aggregate self-employment

rate in the origin country to approximate the individual experience of self-employment can be erroneous, possibly due to immigrants' selection. Second, the accumulation of the self-employment specific human capital might be more relevant than the exposure to a context in which the culture of self-employment is widespread.

References

1. Akee, R. K., Jaeger, D. A., & Tatsiramos, K. (2013). The persistence of self-employment across borders: New evidence on legal immigrants to the United States. Economics Bulletin, 33(1), 126-137.
2. Baldwin-Edwards, M. (2012). The Southern European model of immigration. European Immigrations, 149.
3. Borjas, G. J. (1986). The self-employment experience of immigrants. The Journal of Human Resources, 21(4), 485-506
4. Fairlie, R. W., & Meyer, B. D. (1996). Ethnic and racial self-employment differences and possible explanations. Journal of human resources, 757-793.
5. Fairlie, R. W., & Lofstrom, M. (2015). Immigration and entrepreneurship. In Handbook of the economics of international migration (Vol. 1, pp. 877-911). North-Holland.
6. Fairlie, R. W., & Woodruff, C. (2007). Mexican entrepreneurship: A comparison of self-employment in Mexico and the United States. In Mexican immigration to the United States (pp. 123-158). University of Chicago Press.
7. Fairlie, R., & Woodruff, C. M. (2010). Mexican-American entrepreneurship. The BE Journal of Economic Analysis & Policy, 10(1).
8. Fullin, G., & Reyneri, E. (2011). Low unemployment and bad jobs for new immigrants in Italy. International Migration, 49(1), 118-147.
9. Garcia-Diez M.M., Perez-Villadoniga M.J. (2013). A New Insight into the Home-Country Self-Employment Hypothesis: The Case of Spain. Economic Discussion Papers. Universidad de Oviedo. Available online at: http://economia.uniovi.es/investigacion/papers
10. Hammarstedt, M., & Shukur, G. (2009). Testing the home-country self-employment hypothesis on immigrants in Sweden. Applied Economics Letters, 16(7), 745-748.
11. Kloosterman, R. C. (2010). Matching opportunities with resources: A framework for analysing (migrant) entrepreneurship from a mixed embeddedness perspective. Entrepreneurship and Regional Development, 22(1), 25-45.
12. Kloosterman, R., & Rath, J. (2001). Immigrant entrepreneurs in advanced economies: mixed embeddedness further explored. Journal of ethnic and migration studies, 27(2), 189-201.
13. King, R. (2000). Southern Europe in the changing global map of migration. In Eldorado or Fortress? Migration in Southern Europe (pp. 3-26). Palgrave Macmillan, London.
14. Light, I. (1984). Immigrant and ethnic enterprise in North America. Ethnic and racial studies, 7(2), 195-216.
15. Lofstrom, M., & Wang, C. (2009). Mexican-American self-employment: a dynamic analysis of business ownership. In Ethnicity and Labor Market Outcomes. Emerald Group Publishing Limited.
16. OECD (2010) Open for Business: Migrant Entrepreneurship in OECD Countries. OECD Publishing.
17. OECD (2017) Employment—Self-employment rate—OECD Data. Available at: https://data.oecd.org/emp/self-employment-rate.html〉
18. Parker, S. C. (2004). The economics of self-employment and entrepreneurship. Cambridge university press.
19. Sowell, T. (1996). Migrations and cultures: A worldview (No. 04; JV6217, S6.).
20. Temkin, B. (2009). Informal Self-Employment in Developing Countries: Entrepreneurship or Survivalist Strategy? Some Implications for Public Policy. Analyses of Social Issues and Public Policy, 9(1), 135-156.
21. Tibajev, A. (2019). Linking self-employment before and after migration: Migrant selection and human capital. Sociological Science, 6, 609-634.

22. Van Tubergen, F. (2005). Self-employment of immigrants: A cross-national study of 17 western societies. Social forces, 84(2), 709-732.
23. Yuengert, A. M. (1995). Testing hypotheses of immigrant self-employment. Journal of human resources, 194-204.

The impact of integration on immigrants' health behaviours in Italy

L'impatto dell'integrazione sui comportamenti correlati alla salute tra gli immigrati in Italia

Giovanni Minchio, Raffaella Rusciani, Teresa Spadea

Abstract The adoption of unhealthy behaviours is influenced by a plethora of determinants, particularly among immigrants. Portraying a behavioural profile is necessary for preventing the deterioration of immigrants' health capital. Recent studies offer a 'micro-level' approach for measuring integration at an individual level through survey data. We used data from a national survey involving more than 15000 first generation immigrants and analysed six individual integration indicators capturing different aspects of social life. We assessed the impact of these integration indicators on obesity, daily alcohol consumption and smoking, accounting also for sociodemographic and migration characteristics, through multivariable Poisson models. Results vary depending on the studied outcome and indicator. In general, lower levels of integration increase the risk of obesity but protect women from smoking and drinking alcohol.

Abstract *L'adozione di comportamenti dannosi per la salute è influenzata da una moltitudine di fattori, in particolare tra gli immigrati. Descrivere il profilo comportamentale è necessario per prevenire il deterioramento della salute degli immigrati. Studi recenti offrono un approccio 'micro' per misurare l'integrazione a livello individuale su informazioni rilevate da questionari. Abbiamo usato i dati di un'indagine nazionale su più di 15000 immigrati di prima generazione e analizzato sei indicatori individuali di integrazione in diversi campi della vita sociale. Abbiamo stimato gli effetti di questi indicatori di integrazione su obesità, consumo giornaliero di alcol e fumo, attraverso modelli multivariati di Poisson che tenessero conto anche di caratteristiche sociodemografiche e del percorso migratorio. I risultati variano in funzione dell'esito e dell'indicatore considerati. In generale, livelli più bassi di integrazione aumentano il rischio di obesità ma proteggono le donne dal fumo e dal consumo di alcolici.*

Key words: Immigrants; integration; health-related behaviours; Italy; socioeconomic factors; migration pathway

[1] Giovanni Minchio, Department of Sociology and Social Research, University of Trento, Via Giuseppe Verdi, 26, 38122 Trento, Italy; email: giovanni.minchio@unitn.it

Raffaella Rusciani, Epidemiology Unit, ASL TO3 Piedmont Region, Via Sabaudia 164, 10095 Grugliasco (TO), Italy; email: raffaella.rusciani@epi.piemonte.it

Teresa Spadea, Epidemiology Unit, ASL TO3 Piedmont Region, Via Sabaudia 164, 10095 Grugliasco (TO), Italy; email: teresa.spadea@epi.piemonte.it

1 Introduction

In the light of the literature suggesting that adopting healthy lifestyles may lower the overall mortality risk by 66% with respects to engaging harmful behaviours (Loef 2012), the definition of the behavioural risk profile of a population appears essential to prioritize actions for prevention and health care services organization. Specifically among migrants, which are characterized by good health conditions at the moment of migration (the so-called 'healthy migrant effect') (Razum 2000), the promotion of healthy behaviours would be necessary to interrupt the deterioration of their health capital and to prevent future poor health outcomes (Ikram 2015).

The mechanisms of risky behaviours adoption among immigrants are very complex, operating in different phases of their life course, going from reflecting past exposures to the acculturation process and the present social disadvantage in working and living conditions, which they share with the lowest socioeconomic classes of the host population (Spallek 2011, Acevedo-Garcia 2012). Such mechanisms are mainly influenced by characteristics of the migration pathway, along with cultural and socioeconomic factors and the level of inclusion and integration in the new country (McKay 2003, Spadea 2018).

The past three decades experienced the growth of attention on understanding the integration process of immigrants in European countries, leading to the definition of the Migrant Integration Policy Index (MIPEX), a multi-dimensional score of migrants' opportunities for full participation in various areas of social life (e.g. labour market access or political participation), regularly updated for 56 countries worldwide (available at https://www.mipex.eu/). Several authors combined the MIPEX 'macro level' approach with individual survey data, investigating the association between immigrants' health, individual socioeconomic position and contextual integration by comparing health outcomes among European countries grouped by different integration levels (Malmusi 2014, Giannoni 2016). However, such a methodology cannot take into account the level of participation in social life of each subject, which may depend not only on national policies, but also on specific individual life experiences. Therefore, recent studies have also offered a 'micro-level' approach for measuring integration at an individual level through survey data, by aggregating scores on integration-related variables collected for each interviewee. This approach provides more detailed and flexible indicators, which, similarly to MIPEX, cover various areas of social life (Blangiardo 2013, 2018).

In 2011-2012, the National Institute of Statistic (Istat) conducted the first national survey on 'Social conditions and integration of foreign citizen in Italy' (https://www.istat.it/en/archivio/191097), investigating all aspects of individuals' migration history. The survey provided a unique opportunity for analysing the impact of integration on health behaviours, jointly accounting for sociodemographic factors and characteristics of the migration pathway. Using these data, in the first phase of our research, we found that low levels of individual integration, as measured by the sense of loneliness in Italy and language difficulties with the doctor, had an additional negative impact on immigrants' health behaviour, even after adjusting for sociodemographic characteristics and migration pathway (Minchio, 2022). The effect of the two integration indicators, however, was not consistent between genders and

for all behaviours, leaving space to more in-depth analyses. Therefore, thanks to the detail in its integration-related questions, we followed up on the previous study, with the aim of exploring the impact of integration along its main dimensions: cultural, political, social and economic. In particular, we aimed at investigating the additional impact on unhealthy lifestyles of each dimension of integration, assessing also the possible interaction with the socioeconomic position of immigrants.

2 Data and methods

We used data from the Istat multipurpose survey 'Social conditions and integration of foreign citizen in Italy', conducted in 2011-2012 on a representative sample of about 12000 households with at least one foreigner resident in Italy. We selected all the individuals born abroad and with foreign citizenship at birth (first generation of immigrants); health behaviours were enquired only among people aged 15 years or more and we further decided to set an upper age limit to 64 years, both because the proportion of older foreigners was very small (about 3%) and to reduce the possible impact of the salmon bias, suggesting a selective return to countries of origin of older and sicker migrants (Razum 2006). The final study population consisted of 6947 men and 8783 women.

The health-related behaviours investigated in the survey were obesity, smoking and daily alcohol consumption. Obesity is derived from the body mass index (BMI), including subjects with BMI >= 30, and the reference category is normal weight (20=<BMI<25). The smoking habit variable compares current smokers only with people who have never smoked: we excluded former smokers because no information on smoking duration or quitting date is available, so this category may be too heterogeneous and comparisons may be unreliable. Daily alcohol consumption, finally, groups subjects who declare that they drink at least one glass of beer, wine or spirits every day, and is treated as a dichotomous (yes/no) variable.

Integration was measured by six indicators: the four composite indexes of cultural, political, social and economic integration developed by Blangiardo and Mirabelli (2018), and the two simple indicators used in the previous analysis: 'sense of solitude in Italy' and 'language difficulties with the doctor', obtained by combining difficulties in explaining symptoms with those in understanding therapeutic prescriptions. Shortly, the cultural index represents the level of acculturation in Italy, e.g. language knowledge, reading or watching media in Italian, or eating Italian food; the political index reflects attention to the Italian political life as opposed to political events in the country of origin; the social index reflects the ease of access to social and health services, and participation in recreational, political party or volunteering activities; finally, the economic index considers various characteristics of the employment status, such as the occupational condition, the type of contract or any discrimination suffered at work. The specific components of the integration indexes are available upon request. The composite indexes vary between -1 and 1 and are constructed on the basis of the observed frequencies for each composing variable. In the multivariable analysis, we transformed the cultural, political and social integration indexes in

categorical variables based on their quintiles ('low' for any observation belonging to the first quintile, 'medium' for those in the second, third and fourth quintiles and 'high' for those in the fifth quintile). The economic integration index, on the other hand, was classified in a dichotomous variable (below/above the median), due to its skewed distribution with frequencies concentrated in only a few values on either side of the median, which made the classification into quintiles unfeasible.

Along with integration, we analysed two other classes of possible determinants of health-related lifestyles and two possible confounders. The first class included the socio-demographic characteristics, represented by family composition, marital status, occupational condition and educational level. The second one grouped indicators of the migration pathway, and specifically the area of birth, the length of residence in years, the reason for migration, and the type of transportation used. The classification used for each variable can be seen in Table 1. Finally, as possible confounders, we included the age at the interview in 10-year age classes and four macro-regions of residence to account for possible geographical variations.

2.1 Statistical methods

To explore in detail the independent impact of the integration indexes on health behaviours and their association with the other determinants, we used a three-step procedure: for each outcome and separately by gender, we started from the adjusted model including the socio-demographic determinants, the migration characteristics and the confounders (six models M0). We then added the six integration indicators one at a time (thirty-six models M1: three outcomes by six indicators by gender), and finally tested their interaction with the educational level, which represents the socioeconomic position (SEP) before arriving in Italy and was not considered in any of the composite integration indicators.

We applied robust multivariable Poisson models estimating prevalence ratios (PR), which has been proved to be the best choice for analysing data from a cross-sectional study (Barros 2003). The multivariable Poisson models were performed using generalised linear models with a log link function (Espelt 2016) and robust standard errors computed using the heteroscedasticity-consistent 'HC1' estimator in the R 'sandwich' package (Zeileis 2006). We tested the significance of the interaction terms of the integration indicators with education using the Wald tests performed in the R 'aod' package (Lesnoff 2012). For the models showing a significant Wald test of the interaction term, we further stratified the corresponding M1 model by educational level, to assess how the effect was modified. All the analyses were weighted by Istat normalized survey weights. To overcome problems of collinearity, since the occupational condition is included in the economic integration index, we excluded occupation from all the models that studied the effects of economic integration.

Main results and data cleaning have been carried out using R version 4.1.2 (R core team 2021).

3 Results

The distribution of the study population according to sociodemographic and migration characteristics is reported in table 1. This population consists predominantly of people arriving from Eastern Europe and the new EU countries (56%), particularly women, followed by immigrants from Africa (20%), Asia (13%), and South America (7%). There is also a small proportion of immigrants from developed countries (4%), which were not considered in the following multivariable analyses. The participants have a medium-high educational level, with less than 15% with only primary education; women are more educated (45% with higher qualifications vs. 32% among men), but less occupied (55% vs. 78%). The great majority has lived in Italy for longer than 5 years (83%). Women arrived later and at an older age, mostly for family reasons, while 65% of men have moved to look for a job. A forced migration and an uncomfortable trip are infrequent in this resident population (3% and 7%, respectively).

Table 2 reports the prevalence of unhealthy behaviours by gender and level of integration according to the six indicators. Obesity has a prevalence of 7.5% in both genders, while the other behaviours are much more common in men than in women (33.6% vs. 15.8% for smoking and 28.0% vs. 9.4% for daily alcohol use). Looking at the prevalence by level of integration, we observe a first striking difference by gender: the prevalence of obesity is in fact generally higher among the most integrated men, with variations ranging from 5.0% to 8.4%, but it is higher among less integrated women, with larger oscillations (from 4.1% among culturally integrated women to 9.8% among those with a high sense of loneliness). On the other side, smoking and drinking are more frequent among women with the highest levels of integration, consistently across the various indicators, with the only exception of the sense of solitude that has a reversed trend of smoking (although differences between feelings of loneliness are slight and not significant). Among men, daily alcohol consumption is more common among the most integrated, while smoking does not show a clear and consistent pattern for the different integration indicators.

The multivariable analysis built on the variables in table 1 (models M0) confirmed previous results for the sociodemographic and migration characteristics (Minchio (2022), detailed results are available upon request). Compared to couples with children, all the other family typologies are generally at higher risk of unhealthy behaviours, as well as not married people compared to married living with spouse, with the only exception of not married men, who appear protected from obesity. Being outside the labour market is a risk factor for obesity but is protective for smoking and drinking, while a low educational level is significantly associated only with obesity among women. Non-European citizens always show healthier behaviours, except for African women and South American men and women, who are more likely to be obese than Europeans. The risk of unhealthy behaviours increases with time from arrival, particularly among women, and for people declaring a forced migration or an uncomfortable trip; on the other hand, women arriving for family reasons resulted protected from alcohol use and smoking.

Table 1: Socio-demographic characteristics and migration pathway among resident immigrants aged 15-64, by gender. Italy, 2011-2012

		Men (n=6947)			Women (n=8783)		
		N	%*	95% CI	N	%*	95% CI
A. Socio-demographic characteristics							
Family composition	Couples with children	3583	51.7	50.0-53.4	4139	50.0	48.5-51.4
	Couples without children	871	9.3	8.4-10.2	1282	13.4	12.5-14.4
	Lone parents	862	12.5	11.5-13.6	1523	16.3	15.4-17.5
	Single people	1181	21.0	19.4-22.6	1630	18.4	17.3-19.5
	Other	450	5.5	4.9-6.3	209	1.8	1.5-2.2
Marital status	Married living with spouse	3431	46.7	45.2-48.5	4255	50.7	49.2-52.1
	Married living at distance	577	8.8	7.9-9.9	460	5.6	5.0-6.3
	Not married	2939	44.3	42.6-46.0	4068	43.7	42.3-45.2
Educational level	High school or more	1995	32.4	30.8-34.0	3735	45.2	43.8-46.7
	Middle school	3755	52.7	51.1-54.4	3958	43.7	42.3-45.1
	Up to primary school	1197	14.9	13.7-16.1	1090	11.1	10.3-12.0
Occupational condition	Occupied	5433	78.0	76.5-79.3	4879	55.2	53.8-56.7
	Job seekers	670	10.4	9.4-11.5	793	9.7	8.8-10.6
	Inactive	844	11.6	10.6-12.7	3111	35.1	33.7-36.5
B. Migration pathway							
Area of birth	New EU countries	1745	24.2	22.8-25.7	2947	30.0	28.9-31.5
	Eastern Europe and Balkans	1738	22.2	20.9-23.5	2403	26.1	24.9-27.3
	North Africa and Middle East	1338	18.5	17.2-20.0	1012	11.3	10.4-12.3
	Sub-Saharan Africa	500	7.7	6.8-8.6	372	5.1	4.4-5.7
	Asia	1100	17.2	15.9-18.5	948	12.4	11.4-13.5
	South America	327	6.5	5.6-7.4	719	9.7	8.8-10.6
	Developed countries	199	3.8	3.2-4.7	382	5.3	4.6-6.0
Length of residence in Italy (years)	0-4	1090	15.0	13.8-16.3	1631	17.8	16.7-18.9
	5-10	2177	32.4	30.8-34.0	3413	39.0	37.6-40.4
	11-14	1989	30.1	28.6-31.7	2420	28.8	27.5-30.1
	15+	1691	22.5	21.2-23.9	1319	14.5	13.5-15.5
Reason for migration	Work	4660	65.1	63.4-66.6	4179	45.2	43.8-46.6
	Family	1662	25.5	24.1-27.0	3981	47.5	46.1-48.9
	Forced	353	4.6	4.0-5.4	184	1.9	1.6-2.3
	Other	270	4.8	4.0-5.6	439	5.4	4.8-6.1
Type of transportation	Comfortable	2781	46.7	45.0-48.4	3801	49.7	48.2-51.1
	Average	3308	44.5	42.8-46.1	4668	48.1	46.6-49.5
	Uncomfortable	776	8.8	8.0-9.7	279	2.3	1.9-2.8

* percentages do not correspond to the Ns because they are weighted by a weight that accounts for the sampling design

Table 2: Prevalence of unhealthy behaviours and 95% confidence intervals by gender and level of integration according to the different indicators. Italy, 2011-2012

		Obesity				Daily alcohol consumption				Smoking habit			
		Men (n=483)		Women (n=607)		Men (n=1899)		Women (n=759)		Men (n=2390)		Women (n=1553)	
		%*	95% CI	%*	95% CI	%*	95% CI	%*	95% CI	%*	95% CI	%*	95% CI
		7.5	6.7-8.4	7.5	6.8-8.4	28.0	26.5-29.5	9.4	8.6-10.3	33.6	32.0-35.2	15.8	14.8-16.9
Cultural Integration	High	8.2	6.3-10.1	4.1	2.9-5.3	33.0	29.6-36.4	14.3	12.2-16.4	36.1	32.6-39.6	20.1	17.6-22.6
	Medium	7.7	6.6-8.9	9.1	8.0-10.3	27.4	25.5-29.4	9.1	8.0-10.2	33.9	31.8-35.9	16.7	15.3-18
	Low	5.7	3.7-7.6	6.8	5.2-8.4	23.3	19.9-26.6	4.7	3.2-6.1	29.2	25.5-32.8	8.0	6.2-9.8
Political Integration	High	8.2	6.2-10.3	7.8	6.1-9.4	32.2	28.6-35.7	12.9	10.9-14.9	35.2	31.6-38.7	21.0	18.5-23.6
	Medium	7.5	6.4-8.6	7.5	6.4-8.5	27.5	25.6-29.5	8.5	7.5-9.6	32.6	30.6-34.7	14.3	13.0-15.6
	Low	6.6	4.8-8.5	7.5	5.9-9.1	24.3	20.9-27.8	8.3	6.3-10.2	34.9	31.0-38.7	14.3	12.2-16.5
Social Integration	High	7.3	5.4-9.1	6.3	4.7-8.0	26.2	22.8-29.5	12.0	9.9-14.1	31.3	27.8-34.9	16.0	13.7-18.3
	Medium	8.2	7.1-9.4	7.9	6.9-8.9	28.8	26.8-30.7	9.0	7.9-10.0	34.4	32.4-36.4	16.3	14.9-17.6
	Low	5.0	3.4-6.7	7.7	5.9-9.6	26.9	23.3-30.4	8.1	6.3-10.0	32.9	29.2-36.5	14.0	11.9-16.1
Economic Integration	High	7.8	6.6-8.9	7.1	5.8-8.5	30.6	28.5-32.8	12.2	10.6-13.9	33.9	31.7-36.1	16.9	15.1-18.7
	Low	7.1	5.9-8.4	7.8	6.8-8.7	24.9	22.8-27.1	8.0	7.0-8.9	33.2	30.8-35.5	15.3	14.0-16.5
Sense of solitude in Italy	Not Any	8.4	7.2-9.5	7.3	6.4-8.3	27.9	26.0-29.9	9.6	8.4-10.7	34.2	32.1-36.3	15.4	14.0-16.8
	Low	5.7	4.2-7.2	6.6	5.2-8.1	29.7	26.6-32.8	9.6	8.0-11.2	30.2	27.2-33.2	15.7	13.9-17.5
	High	6.9	4.7-9.2	9.8	7.5-12.2	25.2	21.3-29.1	8.6	6.5-10.8	36.6	32.1-41.0	17.5	14.7-20.3
Language difficulties with the doctor	Not Any	8.4	7.2-9.5	6.7	5.7-7.6	29.0	27.0-31.0	11.0	9.8-12.1	35.2	33.1-37.3	18.8	17.4-20.2
	Some	6.8	5.0-8.6	9.4	7.6-11.2	25.5	22.7-28.4	8.5	7.0-10.1	31.8	28.7-34.9	13.1	11.2-14.9
	Many	6.3	3.9-8.6	8.7	6.4-11	22.7	18.4-27.0	3.9	2.1-5.8	27.5	23.0-32.0	5.1	3.7-6.6

* percentages do not correspond to the Ns because they are weighted by a sampling design weight

3.1 The impact of integration

The distributions of the integration indexes vary by gender. On average, cultural and social integration are greater among women, while political and economic integration among men, with statistically significant differences. The density distributions, however, are quite similar, except for the economic integration index, which has a high concentration around the highest values, particularly among men. Almost 60% of the interviewees report that they do not feel alone or have had any language difficulties with doctors, with a slight male predominance.

The estimates of the integration indicators from models M1 are displayed in table 3. In general, lower levels of integration are associated with a higher risk of obesity in both genders, while less integrated women seem to be protected from daily alcohol consumption. None of the integration indexes have hardly any effect on men's smoking habit. On the other side, the association with integration of drinking among men and smoking among women depends on the specific dimension considered.

After multivariable adjustment, a medium level of cultural integration increases the risk of being obese by 34% in men, compared to those in the highest level (5th quintile), while women in the medium and low level are 89% and 46% more likely to be obese, respectively. On the contrary, the same women are 29% and 35% less likely to drink alcohol daily. Low levels of political integration protect women from daily alcohol and from smoking, with risk reductions above 20%, and lower the risk of drinking alcohol by 14% among men. Low social integration is a risk factor for obesity (although significantly only among men in the medium level, PR=1.33) and for alcohol use among men (PR =1.20 in the low level). Conversely, less integrated

women are protected from alcohol consumption. Economic integration has significant effects only among women, increasing by 40% the risk of being obese and by 18% the risk of smoking in those with a level of integration below the median. The sense of solitude is the only indicator showing an effect on men's smoking, with a protection of 13% among those with some sense of loneliness. On the other hand, men who feel lonely are 16% more likely to drink alcohol and women 26% more likely to smoke. Like the economic indicator, facing language difficulties with the doctor has an effect only among women, increasing the risk of obesity (by 31% and 42% in the two levels) while protecting them from smoking (with 17% and 34% risk reductions).

When we added the education-by-integration interaction terms into the M1 models, only one term was significant in the Wald test, i.e. cultural integration for smoking among women. We therefore analysed the cultural indicator separately in the three educational levels and compared the results (detailed results are available upon request). This comparison revealed that the medium and low levels of integration, which were not associated to smoking overall (table 3, PR=1.00 and PR=0.84, respectively), were significantly protective only among less educated women (PR=0.44 and PR=0.28, respectively).

Table 3. Impact of the integration indicators on unhealthy behaviours by gender. Prevalence ratios (PR) and 90% confidence intervals. Italy 2011-2012

		Obesity				**Daily alcohol consumption**				**Smoking habit**			
		Men		**Women**		**Men**		**Women**		**Men**		**Women**	
		PR*	**90% CI**	**PR***	**90% CI**	**PR***	**90% CI**	**PRR***	**90% CI**	**PR***	**90% CI**	**PRR***	**90% CI**
Cultural Integration	High	1		1		1		1		1		1	
	Medium	**1.34**	**1.06-1.68**	**1.89**	**1.45-2.45**	0.95	0.86-1.06	**0.71**	**0.59-0.85**	0.99	0.90-1.10	1.00	0.87-1.14
	Low	1.24	0.81-1.50	**1.46**	**1.04-2.05**	1.09	0.92-1.28	**0.65**	**0.48-0.90**	0.95	0.83-1.10	0.84	0.67-1.06
Political Integration	High	1		1		1		1		1		1	
	Medium	1.11	0.87-1.41	1.00	0.82-1.23	0.94	0.84-1.05	**0.73**	**0.61-0.87**	0.95	0.86-1.04	**0.82**	**0.72-0.94**
	Low	1.11	0.81-1.50	0.95	0.75-1.21	**0.86**	**0.73-1.00**	**0.77**	**0.59-0.99**	1.03	0.91-1.17	0.88	0.75-1.03
Social Integration	High	1		1		1		1		1		1	
	Medium	**1.33**	**1.04-1.69**	1.22	0.97-1.54	1.10	0.98-1.24	**0.75**	**0.63-0.90**	1.08	0.96-1.20	1.10	0.95-1.27
	Low	1.06	0.75-1.49	1.19	0.90-1.59	**1.20**	**1.03-1.40**	0.82	0.63-1.07	1.08	0.94-1.24	1.14	0.95-1.36
Economic Integration **	High	1		1		1		1		1		1	
	Low	1.12	0.92-1.35	**1.40**	**1.17-1.69**	0.95	0.87-1.04	1.01	0.84-1.21	1.05	0.97-1.14	**1.18**	**1.05-1.34**
Sense of solitude in Italy	Not any	1		1		1		1		1		1	
	Low	0.86	0.67-1.10	0.86	0.70-1.05	**1.16**	**1.05-1.28**	0.98	0.77-1.26	**0.87**	**0.79-0.96**	1.02	0.90-1.16
	High	1.07	0.80-1.44	1.19	0.96-1.48	1.10	0.97-1.25	0.98	0.81-1.18	1.09	0.97-1.22	**1.26**	**1.08-1.47**
Language difficulties with the doctor	Not any	1		1		1		1		1		1	
	Some	1.04	0.81-1.34	**1.31**	**1.08-1.60**	1.09	0.97-1.21	0.71	0.47-1.07	0.98	0.89-1.08	**0.83**	**0.72-0.96**
	Many	1.13	0.80-1.61	**1.42**	**1.09-1.87**	1.14	0.97-1.35	0.88	0.72-1.08	0.91	0.78-1.06	**0.66**	**0.50-0.86**

* estimates from robust Poisson models mutually adjusted for all the variables in table 1 plus age, residence macro-region and each integration indicator separately

** not adjusted for occupational condition

4 Discussion

Unhealthy behaviours are more common among men than women, apart from obesity that is equally distributed between genders. The impact of integration on heath behaviours varies widely both by gender and depending on which outcome and integration indicator are analysed, portraying integration as a heterogeneous phenomenon.

Lower levels of integration increase the probability of being obese among both women and men, and consistently for most indicators. These results conflict with the evidence on the positive, or non-significant, association between acculturation and obesity found in other Western countries (Alidu 2018, Dijkshoorn 2008). A possible explanation for this inconsistency is the peculiarity of Italy of being a Mediterranean country characterized by a high prevalence of a healthy diet (Denoth 2016), while the majority of immigrants arrive from areas with a high prevalence of obesity (Marques 2018). Therefore, higher levels of integration and acculturation allow immigrants to benefit from a faster knowledge and transition to the healthier Mediterranean diet, more widespread and affordable in Italy than in their countries, thus reducing obesity (Denoth 2016).

Among women, lower levels of cultural, political and social integration protect from daily alcohol consumption and smoking. Hence, it appears that poorly integrated women, who are also often outside the labour market and less in contact with the host population, are less vulnerable to acquiring drinking and smoking habits, generally more frequent in the host population than in the countries of origin (Hosper 2007, Lara 2005, Acevedo-Garcia 2012). This was particularly true for less educated women, as reflected in the stratified analysis: the observed interaction, in fact, results in an amplification of the protective effect among less educated and less culturally integrated women.

The protective effect, however, is counterbalanced by the excess risk of smoking observed among women who have low levels of economic integration and those who feel lonely in Italy. Similarly, some sense of solitude and low social integration increase the risk of daily alcohol consumption among men. Such findings could be linked to the effects of marginalization on men's alcohol abuse (Rehm 2015) and women's smoking habit (Lehmiller 2012). More generally, immigrants less integrated in the host country, who tend to form close social networks within their communities of origin, reveal poorer health than those who have a stronger mix with natives (Rostila 2010).

We also observed important differences by gender, with men's behaviours less affected by integration indicators. Indeed, men – who arrived earlier and at a younger age for work reasons – are more likely than women to have completed their acculturation process, and particularly their political and economic indexes display distributions more concentrated around higher values, therefore less capable to detect a significant impact on behaviours. On the other hand, women, although less represented in the labour market, have higher levels of cultural and social integration, probably due to their parenting role, which provides them with alternative social networks.

4.1 Conclusions

To our knowledge this is the first study on the impact of acculturation on health-related behaviours, conducted at the national level on a representative sample of the Italian population. However, it finds its main limitation in the sample composition: since the survey included only residents, characterized by a medium-long stay in Italy, respondents embody the more integrated part of the foreign population. This also contributes to explain the limited impact of some integration indicators.

Nonetheless, a few conclusions are worthy noticing.

First, the important differences observed between men and women in the acculturation and integration processes warn about the need to keep a gender approach in monitoring and analysing these phenomena.

Secondly, in some cases greater integration with the host population leads to the adoption of unhealthy behaviours. This happens because immigrants very often tend to mix, and therefore to share lifestyles, with the lower socio-economic segments of the population (Malmusi 2010), where unhealthy behaviours are more prevalent (Mackenbach 2008). This should not lead to slowing down the integration process but rather to monitor its potential negative effects, targeting preventive interventions not only at immigrants but also at the most disadvantaged groups of the native population.

Finally, particular attention should be paid to the areas of more marginalized individuals, characterized by a higher prevalence of smoking among women and alcohol consumption among men.

Acknowledgements

The analyses here presented are part of the project "Immigration, integration, settlement. Italian-Style", financed by the Italian Ministry for Education, University and Research under the Grant: PRIN 2017N9LCSC_003. GM holds a research grant within this project. The sponsor had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. All the authors could access and verify the data, approved the manuscript and are responsible for the views expressed in it.

References

1. Alidu, L., Grunfeld, E.A.: A systematic review of acculturation, obesity and health behaviours among migrants to high-income countries. Psychol Health. 33, 724-745 (2018) doi: https://doi.org/10.1080/08870446.2017.1398327
2. Acevedo-Garcia, D., Sanchez-Vaznaugh, E.V., Viruell-Fuentes, E.A., Almeida, J.: Integrating social epidemiology into immigrant health research: A cross-national framework. Soc Sci Med. 75, 2060-2068 (2012) doi: https://doi.org/10.1016/j.socscimed.2012.04.040.
3. Barros, A.J.D., Hirakata, V.N.: Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 3, 1-13 (2003) doi: https://doi.org/10.1186/1471-2288-3-21.
4. Blangiardo, G.C.: Per misurare l'integrazione. [For measuring integration] Libertà Civili. 2, 24-39 (2013)
5. Blangiardo, G.C., Mirabelli, S.M.: Misurare l'integrazione. [Measuring Integration]. In: Vita e Percorsi di Integrazione degli Immigrati in Italia [Immigrants' life and integration paths in Italy]. Pp. 361-381. ISTAT (2018). Available at: https://www.istat.it/it/files//2019/05/Vita-e-percorsi.pdf
6. Denoth, F., Scalese, M., Siciliano, V. et al.: Clustering eating habits: frequent consumption of different dietary patterns among the Italian general population in the association with obesity, physical activity, sociocultural characteristics and psychological factors. Eat Weight Disord. 21, 257–268 (2016). doi: https://doi.org/10.1007/s40519-015-0225-9
7. Dijkshoorn, H., Nierkens, V., Nicolaou, M.: Risk groups for overweight and obesity among Turkish and Moroccan migrants in The Netherlands. Public Health. 122, 625-630 (2008) doi: https://doi.org/10.1016/j.puhe.2007.08.016.
8. Espelt, A., Marí-Dell'Olmo, M., Penelo, E., Bosque-Prous, M.: Applied Prevalence Ratio estimation with different Regression models: An example from a cross-national study on substance use research. Adicciones. 29, 105-112 (2016) doi: https://doi.org/10.20882/adicciones.823
9. Ikram, U.Z., Malmusi, D., Juel, K., Rey, G., Kunst, A.E.: Association between integration policies and immigrants' mortality: an explorative study across three European countries. PLoS One. 10, e0129916 (2015) doi: https://doi.org/10.1371/journal.pone.0129916.
10. Giannoni, M., Franzini L, Masiero G.: Migrant integration policies and health inequalities in Europe. BMC Public Health.16-463 (2016) doi: https://doi.org/10.1186/s12889-016-3095-9.
11. Lara, M., Gamboa, C., Kahramanian, MI., Morales LS., Hayes-Bautista DE: Acculturation and Latino health in the United States: a review of the literature and its sociopolitical context. Annu Rev Public Health. 26, 367-397 (2005) doi: https://doi.org/10.1146/annurev.publhealth.26.021304.144615
12. Lehmiller, J.J.: Perceived marginalization and its association with physical and psychological health. J Soc Pers Relat. 29, 451-469 (2012) doi: https://doi.org/10.1177/0265407511431187
13. Lesnoff, M., Lancelot R.: aod: Analysis of Overdispersed Data. R package version 1.3.1. (2012) Available at: http://cran.r-project.org/package=aod
14. Loef, M., Walach, H.: The combined effects of healthy lifestyle behaviors on all cause mortality: a systematic review and meta-analysis. Prev med. 55, 163-170 (2012) doi: https://doi.org/10.1016/j.ypmed.2012.06.017.

15. Mackenbach, J.P., Stirbu, I., Roskam, A.J., et al.: European Union Working Group on Socioeconomic Inequalities in Health. Socioeconomic inequalities in health in 22 European countries. N Engl J Med. 358, 2468-2481 (2008) doi: https://doi.org/10.1056/NEJMsa0707519.
16. Malmusi, D.: Immigrants' health and health inequality by type of integration policies in European countries. Eur J Public Health. 25, 293-299 (2015) doi: https://doi.org/10.1093/eurpub/cku156
17. Malmusi, D., Borrell, C., Benach, J. Migration-related health inequalities: showing the complex interactions between gender, social class and place of origin. Soc Sci Med. 71, 1610–1619 (2010) doi: https://doi.org/10.1016/j.socscimed.2010.07.043
18. Marques, A., Peralta, M., Naia, A., Loureiro, N., Gaspar de Matos, M.: Prevalence of adult overweight and obesity in 20 European countries. Eur J Public Health. 28, 295–300 (2014) doi: https://doi.org/10.1093/eurpub/ckx143
19. McKay, L., MacIntyre, S., Ellaway, A.: Migration and health: a review of the international literature. Glasgow: MRC Soc Public Health Sci Unit (2003) Available at: https://www.researchgate.net/profile/Anne-Ellaway/publication/320865881_Migration_and_Health_A_review_of_the_international_literature/links/59ff75570f7e9b9968c69ae5/Migration-and-Health-A-review-of-the-international-literature.pdf
20. Minchio, G., Rusciani, R., Costa, G., Sciortino, G., Spadea, T.: Health behaviours and their determinants among immigrants residing in Italy. Preprint. medRxiv. (2022) doi: https://doi.org/10.1101/2022.03.14.22272345
21. R Core Team.: R: A language and environment for statistical computing. R Found Stat Comput. (2021) Available at: https://www.R-project.org/
22. Razum, O.: Commentary: Of salmon and time travellers - musing on the mystery of migrant mortality. Int J Epidemiol. 35, 919-21 (2006) doi: https://doi.org/10.1093/ije/dyl143.
23. Razum, O., Zeeb, H, Rohrmann, S.: The 'healthy migrant effect' - not merely a fallacy of innacurate denominator figures. Int J Epidemiol. 29, 191-192 (2000) doi: https://doi.org/10.1093/ije/29.1.191.
24. Rehm, J., Allamani, A., Della Vedova, R. et al.: General Practitioners Recognizing Alcohol Dependence: A Large Cross-Sectional Study in 6 European Countries. Ann Fam Med. 3, 28-32 (2015) doi: https://doi.org/10.1370/afm.1742
25. Rostila, M.: Birds of a feather flock together--and fall ill? Migrant homophily and health in Sweden. Sociol Health Illn. 32, 382-399 (2010) doi: https://doi.org/10.1111/j.1467-9566.2009.01196.x
26. Spadea, T., Rusciani, R., Mondo, L., Costa, G.: Health-Related Lifestyles Among Migrants in Europe. In: Rosano, A. (ed), Access to Primary Care and Preventative Health Services of Migrants, pp. 57-64. SpringerBriefs in Public Health. Springer, Cham (2018)
27. Spallek, J., Zeeb, H., Razum, O.: What do we have to know from migrants' past exposures to understand their health status? A life course approach. Emerg Themes Epidemiol. 8, 6 (2011) doi: https://doi.org/10.1186/1742-7622-8-6.
28. Zeileis, A.: Object-Oriented Computation of Sandwich Estimators. J Stat Softw. 16, 1-16. (2006) doi: https://doi.org/10.18637/jss.v016.i09

Migration, gender, and the distribution of paid and unpaid labour. Preliminary perspectives on foreign couples in Italy

Migrazione, genere e divisione del lavoro retribuito e non retribuito. Prospettive preliminari sulle coppie di stranieri in Italia

Rocco Molinari, Agnese Vitali, Ester Gallo

Abstract This paper looks at the intra-couple distribution of paid and unpaid labour, taking as a case in point foreign heterosexual couples in Italy. First, it provides a descriptive outlook of whether and how partners in foreign couples engage in paid and unpaid labour, distinguishing by their macro area of origin. Second, the paper focuses on two aspects of the division of unpaid labour among partners: keeping the house in order and childcare. For each of these two dimensions, we run logistic regression models aimed at identifying the correlates of a gender-traditional division of unpaid work. Results show that, despite differences by area of origin and religion, the intra-couple division of paid labour and the migration pathway followed by couples prove to be associated with gender imbalance in domestic work and childcare tasks.

Abstract *Questo studio esplora la divisione del lavoro retribuito e non retribuito nelle coppie straniere eterosessuali in Italia. In primo luogo, lo studio fornisce un quadro descrittivo della partecipazione dei partners di coppie straniere al lavoro retribuito e non retribuito. In secondo luogo, lo studio analizza due aspetti della divisione del lavoro non retribuito: tenere la casa in ordine e occuparsi della cura dei figli. Per ciascuna di queste dimensioni sono sviluppati dei modelli di regressione logistica che identificano i fattori correlati a una divisione dei ruoli di genere tradizionale. I risultati mostrano che, nonostante le differenze per area di provenienza e appartenenza religiosa, la divisione del lavoro retribuito e il percorso migratorio della coppia sono aspetti associati allo squilibrio di genere nel lavoro domestico e nelle attività di cura dei figli.*

Key words: Domestic work, Childcare, Female breadwinner, Immigrants

1 Introduction

Empirical studies on the division of paid and unpaid work among migrant couples are scarce, and knowledge is especially scarce for Italy, partly due to its relatively recent history as a migrant-receiving country, partly due to scarcity of representative data, which allow studying migrant couples. Italy, traditionally a migrant-sending country, has become, in the past decades, a country of destination for migrants coming from a multitude of origins, from Eastern Europe to North Africa, Asia and Latin America. Despite the increase in the number of foreign residents - which account today for the 8.7% of the country population - we know little about the characteristics of migrant families, in particular in terms of their gender-role division in paid and unpaid work. It should be note that, in the European comparison, Italy scores poorly in terms of gender equality and female labour force participation, while various countries of origin of migrants in Italy score considerably higher. For instance, according to the Global Gender Gap Report 2021 (World Economic Forum, 2021), Moldova, Albania and Philippines – i.e. some of the major countries of origin of migrants in Italy – rank 28th, 25th and 17th, respectively, in terms of gender equality, while Italy is only nr. 63rd. It is therefore particularly interesting to study the gender dynamics of foreign couples in Italy.

2 Background

International literature has recently paid attention to the gendered division of labour or gender-egalitarian attitudes among migrant families (Pessin and Arpino, 2018; Blau et al. 2020). This literature sheds light on the role of culture for the allocation of unpaid labour among migrant families. By comparing foreigners coming from gender-conservative countries of origin to natives in comparatively more gender-egalitarian settings, this work found that migrant couples tend to behave according to the predominant gender-role culture typical of their country of origin (Frank and Hou 2015; Carriero 2021).

However, the distribution of tasks within migrant couples might be the result of different intertwining factors, above and beyond the role of the prevailing culture in the country of origin – a characteristic which, in this study, is measured by the macro-area of origin. In particular, research on native couples found that a more gender-egalitarian distribution of domestic work and childcare is expected among younger and higher-educated couples (Altintas and Sullivan 2016), with no children or a small family size, and when women are employed and in particular in dual-earning couples. Differently, in female-breadwinning couples where the woman is employed and the man is not, due to gender display, women tend to actually do more than their non-employed partner, especially in gender-conservative countries such as Italy (Aassve et al. 2014). In this respect, female-breadwinning couples tend to be less gender egalitarian if compared to dual earner ones. We expect that such

associations hold also for foreign couples. Moreover, specifically to foreign couples, we expect to find a more gender-egalitarian distribution of domestic work and childcare when the woman migrated before her husband (Torosyan, Gerber and Goñalons-Pons 2015; Gallo and Scrinzi 2016); if the couple resides in a more gender-egalitarian setting (Hondagneu-Sotelo 1992), i.e., in the case of Italy, the North of the country; and among non-Muslim couples.

3 Study Objectives

The present paper draws from the above discussion and aims at contributing to it by pursuing two main objectives: (1) to describe the division of paid and unpaid labour among immigrant couples in Italy, disentangling eventual differences on the basis of the area of origin; (2) to study the correlates of the intra-couple distribution of unpaid work among foreign couples in Italy.

4 Data and Methods

To test our hypotheses, we use data provided by the Social Condition and Integration of Foreign Citizens (SCIF) survey, a representative household survey of the foreign population in Italy collected in 2011-12 by the Italian National Institute of Statistics (ISTAT).[1] The survey is based on a random sample of households that include at least one foreign national. Within each household, all members have been separately interviewed.

The SCIF survey collects, among other things, socio-demographic and economic characteristics of household members. Furthermore, the SCIF survey has a module on gender roles and the intra-couple distribution of unpaid work. Each female respondent being part of a couple was asked to specify how housework and childcare activities are distributed among partners (mostly her, mostly her partner, or both partners equally). Unpaid activities include several items, e.g. cleaning, cooking, ordering, shopping, childcare.

We rely on a SCIF sub-sample of married and unmarried first-generation immigrant couples living in the same household whose members are aged 18-65 (N=3,372).[2] We exclude mixed couples, consisting of one Italian and one immigrant partner or two immigrant partners of different nationalities, due to the small sample size of these two groups. We also exclude immigrant couples from EU15 and other highly developed non-EU countries, which generally follow different integration trajectories, for the same reason.

1 https://www.istat.it/en/archivio/191097.

2 Sample size refers to couples.

As dependent variables we consider the intra-couple distribution of two unpaid activities: *keeping the house in order*, as an indicator of unpaid domestic work, and, limited to the sample of couples with children, *childcare*. Originally the survey included, for each of these tasks, 5 possible answers to the question "who spend more time on this unpaid activity?": exclusively her; mostly her; exclusively him; mostly him; equally. We reaggregated these answers and obtained two dichotomous variables that equal 1 when the woman mostly or exclusively contributes within the couple, and 0 otherwise (either if members equally contribute or the man mostly/exclusively contributes).

As independent variables, we consider: the *intra-couple distribution of paid work* (male breadwinner, i.e. only the male partner is employed; dual earner, i.e. both partners are employed; female breadwinner, i.e. only the woman employed; no earner, i.e. none of the partners employed); the couple's macro-area of *origin* (Eastern-Europe; Latin America; Asia; Middle East and North Africa; Other Africa); *years since migration* (0-5; 6-10; 11-15; 16 or more); the *intra-couple migratory history* observing dates of access in Italy (man accessed first; woman accessed first; accessed the same year); an indicator of *religion affiliation* as a dummy variable that equals 1 for couples with at least one Muslim partner; a measure of *intra-couple education* (both man and women lower secondary education or less; both man and woman upper secondary or tertiary education; only man upper secondary or tertiary education; only woman upper secondary or tertiary education); the *woman's age* (18-29; 30-39; 40-49; 50-65); presence of *children* and the couple's macro-*area of residence* in Italy (North-west; North-east; Centre; South and Islands).

We develop two logistic regression models on our dichotomous dependent variables.

5 Results

Table 1 shows strong differences in the intra-couple distribution of paid work by macro-area of origin. Immigrant couples from Latin America show the highest proportion of dual earning and female breadwinning, confirming results by Bueno and Vidal-Coso (2019) for Spain, whereas couples from the Middle East and North Africa the lowest proportion. Amongst immigrant couples from the MENA and (to a lesser extent) from other African countries, male breadwinning is largely the prevailing arrangement. Furthermore, foreign couples from Eastern Europe and Asia have a slightly higher proportion of dual earning than male breadwinning. Finally, Asian foreign couples, along with Latin American ones, show a lower percentage of no earning.

Observing the intra-couple distribution of unpaid work (Table 2), we notice that gender imbalance (women do more than men) among foreign couples is stronger in domestic activities than in childcare. In both cases, couples with men contributing more than women to housework and childcare are uncommon. Furthermore, large differences by origin emerge. The largest gender imbalance in housework is observed among foreign couples from MENA (in 87% of couples the woman is

solely or mostly in charge of domestic work) followed by those from Eastern Europe (80%). By contrast, immigrant couples from Latin America show the lowest percentage (about 60%). Similarly, the division of childcare activities differs between couples from the MENA and Latin America, the latter showing relatively lower levels of gender imbalance.

To study the association between explanatory variables and gender imbalance in the distribution of domestic work and childcare, we present exponentiated coefficient estimates (odds ratios) of the two logistic regression models separately (Table 3). Considering domestic work, we notice that differences among immigrant groups observed in descriptive analyses are partly maintained in the multivariate framework: once controlling for the other variables, all the considered immigrant groups (including immigrants from the MENA) show a statistically significant lower relative risk of a positive gender imbalance compared to immigrant couples from Eastern Europe, with Latin Americans showing the sharpest reduction. This means that some differences amongst groups observed at the descriptive level remain unexplained in the model.

Results identify some migratory background characteristics as important correlates of the distribution of unpaid work. On the one hand, the intra-couple migratory pathway matters: if the woman migrated to Italy before the man, the couple shows a more egalitarian distribution of domestic work among partners, compared to couples for which men migrated before women. On the other hand, religion affiliation is also important: for Muslim couples the likelihood that women are responsible for most of unpaid labour is considerably higher compared to non-Muslim couples.

Furthermore, the intra-couple division of paid work is strongly associated with the intra-couple division of unpaid work: the woman's labour market participation reducing gender imbalance in domestic activities. Either dual earning and (particularly) female-breadwinning couples show a statistically significant lower risk of gender imbalance compared to male-breadwinning couples. Hence, after controls, for our sample of foreign couples, we do not find evidence of any gender display among female-breadwinner couples as predicted by the literature on native population.

Finally, we notice that other socio-demographic characteristics of foreign couples are associated with the intra-couple distribution of domestic work. Couples in which both members are upper-secondary or tertiary educated show a lower gender imbalance compared to couples in which both the man and the woman are lower educated. Moreover, for childless couples we observe a lower relative risk, i.e. a more egalitarian distribution of housework.

Results differ when we move to explain the correlates of the intra-couple division of childcare activities. First, none of the migratory background characteristics considered is associated with the intra-couple division of childcare. Nonetheless, similarly to the case of unpaid domestic work, the intra-couple distribution of paid work does matter: couples with an employed woman show a relatively reduced gender gap in childcaring. Again, female breadwinning, just like dual earning, are associated with a more egalitarian distribution of childcare. Finally,

foreign couples in which only women are highly educated and in which women are relatively older show a less unequal distribution of childcare activities.

Table 1: Intra-couple distribution of paid work by macro-area of origin

	Dual earner	*Male breadwinner*	*Female breadwinner*	*No earner*	*tot*	*N*
Latin	59.6	26.2	11.4	2.8	100	141
Eastern-Europe	45.7	41.7	6.9	5.7	100	1,880
Asia	47.5	43.5	5.2	3.8	100	501
MENA	14.4	71.5	4.2	9.9	100	666
Other Africa	34.4	52.4	7.4	5.8	100	189

Source: *SCIF* 2011-12.

Table 2: Intra-couple distribution of unpaid work by area of origin: domestic work and childcare activities

	Domestic work				*Child-care*			
	Mostly her	*Equal*	*Mostly him*	*TOT*	*Mostly her*	*Equal*	*Mostly him*	*TOT*
Latin	62.1	36.4	1.4	100	39.8	58.3	1.9	100
Eastern-Europe	81.0	17.3	1.7	100	46.4	50.7	2.9	100
Asia	73.8	23.2	3.0	100	44.5	52.3	3.3	100
MENA	87.7	11.0	1.4	100	57.3	39.5	3.2	100
Other Africa	75.7	22.8	1.6	100	45.1	51.2	3.7	100

Source: *SCIF* 2011-12.

Table 3: Logistic regression model estimates on the intra-couple distribution of (1) domestic work and (2) child-care activities (=1 if woman contributes more)

	(1) domestic work		*(2) child-care*	
	Odds Ratios	*Std. Err.*	*Odds Ratios*	*Std. Err.*
Origin				
Eastern-Europe	ref.		ref.	
Latin	0.56**	0.11	0.89	0.19
Asia	0.58***	0.08	0.95	0.12
MENA	0.67*	0.13	1.09	0.15
Other Africa	0.57**	0.11	0.83	0.15
Years since migration				
0-5 years	ref.		ref.	
6-10 years	1.24	0.20	0.79	0.13
11-15 years	1.18	0.20	0.74	0.13
16 of more	1.28	0.24	0.95	0.17
Intra-couple migratory history				

man accessed first	ref.		ref.	
woman accessed first	0.61***	0.08	0.98	0.13
same year	0.97	0.12	0.82	0.09
Religion				
non-Muslim	ref.		ref.	
Muslim	1.79***	0.29	1.08	0.13
Intra-couple paid work distr.				
male breadwinner	ref.		ref.	
dual earner	0.46***	0.05	0.54***	0.05
female breadwinner	0.17***	0.03	0.31***	0.06
no earner	0.53**	0.10	0.79	0.13
Intra-couple education				
M&W lower secondary or less	ref.		ref.	
M&W upper secondary or tertiary	0.66***	0.08	0.82	0.08
M upper secondary or tertiary	0.86	0.14	1.05	0.14
W upper secondary or tertiary	1.00	0.16	0.77*	0.1
Woman's age				
18-29	ref.		ref.	
30-39	0.82	0.11	0.85	0.09
40-49	0.93	0.14	0.63***	0.08
50-65	0.91	0.17	0.38***	0.08
Children				
no	ref.			
yes	1.51***	0.17		
Area of residence				
North-West	ref.		ref.	
North-East	1.02	0.14	0.62***	0.08
Centre	1.42*	0.21	1.19	0.16
South and Islands	1.55***	0.20	0.79*	0.09
Observations	3,372		2,622	
Pseudo R-squared	0.105		0.054	

** p<0.05, ** p<0.01, *** p<0.001*

Source: *SCIF* 2011-12.

6 Conclusions

The analysis developed in this paper aimed at mapping how paid and unpaid labour are distributed among foreign couples in Italy. In doing so it provides a comparison with previous studies mainly based on North America or Western Europe. These studies mostly stressed how migrant couples tend to behave according to the predominant gender-role culture typical of their country of origin and highlighted patterns of adaptation of immigrant couples to the receiving contexts.

We partly confirmed that relevant differences between origin groups in the intra-couple distribution of unpaid work persist among foreign couples in Italy, even accounting for other socio-demographic factors, at least in relation to domestic unpaid activities. Our data also allowed to specify more deeply the role of cultural background: Muslim foreign couples in Italy are substantively less gender

egalitarian in the division of unpaid domestic work than non-Muslim ones. Furthermore, when religion affiliation is taken into account, differences between origin groups decrease substantively, especially considering the case of foreign couples from the MENA, that in the descriptive analysis showed the highest gender gaps.

However, we also revealed how gender unbalances in the distribution of unpaid domestic work and childcare are associated with other factors, partly connected with contextual changes triggered by migration. First, dual-earner foreign couples in Italy and, to a larger extent, female-breadwinning ones are substantively more egalitarian either in the division of housework and childcare. Second, foreign couples in which the woman migrated before the husband/partner show a more gender-egalitarian distribution of domestic work.

Therefore, our findings suggest that the predominant gender-role culture might be transformed though migration. Women who migrate first, who are highly educated, and who actively participate in the receiving labour markets, might act as pioneer subjects within the families and communities and this might lead, in destination countries, to changes in the distribution of paid/unpaid labour within the couple.

References

1. Altintas, E., Sullivan, O.: Fifty years of change updated: Cross-national gender convergence in housework. Demogr. Res., 35, 455–470 (2016)
2. Aassve, A., Fuochi, G., Mencarini, L.: Desperate housework: Relative resources, time availability, economic dependency, and gender ideology across Europe. J. Fam. Issues, 35(8), 1000–1022 (2014)
3. Blau, F.D., Kahn, L.M., Comey, M. et al.: Culture and gender allocation of tasks: Source country characteristics and the division of non-market work among US immigrants. Rev. Econ. Househ., 18, 907–958 (2020)
4. Bueno, X., Vidal-Coso, E.: Vulnerability of Latin American migrant families headed by women in Spain during the Great Recession: A couple-level analysis. J. Fam. Issues, 40(1), 111–138 (2019)
5. Carriero, R.: The role of culture in the gendered division of domestic labor: Evidence from migrant populations in Europe. Acta Sociol., 64(1), 24–47 (2021)
6. Frank, K., Hou, F.: Source-country gender roles and the division of labor within immigrant families. J. Marriage Fam., 77(2), 557–574 (2015)
7. Gallo, E., Scrinzi, F.: Migration, masculinities and reproductive labour: Men of the home. London (2016)
8. Hondagneu-Sotelo, P.: Overcoming patriarchal constraints: The reconstruction of gender relations among Mexican immigrant women and men. Gend. Soc., 6(3), 393–415 (1992)
9. Pessin, L., Arpino, B.: Navigating between two cultures: Immigrants' gender attitudes toward working women. Demogr. Res., 38, 967–1016 (2018)
10. Torosyan, K., Gerber, T.P., Goñalons-Pons, P.: Migration, household tasks, and gender: Evidence from the Republic of Georgia. Int. Mig. Rev., 50(2), 445–474 (2016)
11. World Economic Forum: Global gender gap report, March (2021)

Sampling techniques for big data analysis

Non-probability samples and big data: how to use them?

Campioni non probabilistici e big data: come usarli?

Pier Luigi Conti

Abstract As a consequence of the data deluge phenomenon, non-probability samples have recently received increasing attention. Drawbacks related to the naive use of non-probability samples are illustrated, and proposed remedies are reviewed and discussed.

Abstract *In conseguenza del fenomeno del "data deluge", l'uso di campioni non probabilistici ha di recente ricevuto un'elevata e crescente attenzione. In questo lavoro si illustrano inconvenienti e limitazioni legati ad un uso "ingenuo" di dati derivanti da campioni non probabilistici, e si discutono i possibili rimedi.*

Key words: non-probability samples, weighting, imputation, sampling design, ignorability.

1 Introduction

In the last twenty years, there has been a huge increase of available data, coming from Official Statistics as well as from different private sources. It is the well known phenomenon of *data deluge*, whose effect has been the creation of a (more or less) steady flow of data from private archives to public databases. Sources include not only data in researchers' private archives, but also data generated from online transactions, emails, videos, audios, images, click streams, logs, search queries, health records, social networking interactions, sensors and mobile phones, etc.. In other terms, we are in the big data era.

Using the above data is of primary interest, mainly because their are frequently freely available. However, their effective use is Statistical Inference poses relevant challenges. Probability sampling is actually the basic ingredient of the correct way

Pier Luigi Conti
Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy, e-mail: pierluigi.conti@uniroma1.it

of acquiring data from populations, and it may be considered as the touchstone for data collection processes, at least since [1]. The theory of sampling from finite population offers an excellent background for probability sampling, including stratification, clusters of units, unequal inclusion probabilities, balancing equations, etc.; cfr. [2] for an excellent account on methods of sampling. On the other hand, the use of non-probability samples has frequently produced dramatic failures. A well known case is that of USA Presidential Elections in 1936, where the candidates were Alfred Landon (Republican Party) and Franklin D. Roosevelt (Democratic Party). To predict the winner, the *Literary Digest* (a respected magazine with a long history of accurate predictions of winners of presidential elections) implemented one of the largest and most expensive polls ever conducted. Based on telephone directories in USA, lists of magazine subscribers, and other sources, a mailing list of about 10 million names was created. Every name on that list was mailed a mock ballot and asked to return the marked ballot to the magazine. About 2.4 million people mailed back their ballots. The *Literary Digest* prediction was that Landon would get 57% of the vote, and Roosevelt 43%. On the other hand, on the basis of a (probability) sample of 50000 people, Dr. George Gallup predicted that Roosevelt would win reelection hands down. The actual results of the election were 62% for Roosevelt and 38% for Landon. The estimation error in the *Literary Digest* poll was 19%, probably the largest ever in a major public opinion poll. Practically all of the estimation error was the result of *sample bias*.

There were two main causes of the *Literary Digest* failure: *selection bias* and *nonresponse bias*. The main lesson learned from this failure is that bad sampling methods cannot be cured by increasing the size of the sample. More recent failures of polls are the 2006 Italian Parliamentary Election, the 2015 Israeli Knesset Election, as well as many others; cfr. [3] and references therein.

2 Basic aspects and notation for probability and non-probability samples

Let $\mathscr{U}_N$ be a finite population of size N. The *character of interest* is denoted by $\mathscr{Y}$, and its value for unit i by y_i; furthermore, let $\mathbf{y}_N = (y_1, \ldots, y_N)$.

A sample $\mathbf{s}$ is a subset of $\mathscr{U}_N$. Denote by D_i the *sample membership indicator* of unit i, which is going to be 1 (0) whenever $i \in \mathbf{s}$ ($i \notin \mathbf{s}$). With no loss of generality, D_i may be seen as a Bernoulli random variable (r.v.); clearly, $\mathbf{s} = \{i \in \mathscr{U}_N : D_i = 1\}$. Denote further by $\mathbf{D}_N$ the N-dimensional r.v. of components $(D_1, \ldots, D_N)$. A (unordered, without replacement) sampling design P is the probability distribution of the random vector $\mathbf{D}_N$: $p(\mathbf{s}) = P(D_i = 1 \,\forall\, i \in \mathbf{s},\ D_i = 0 \,\forall\, i \notin \mathbf{s})$. In probability sampling, the joint distribution of $\mathbf{D}_N$ is *known*; it is essentially introduced by the Statistician as a (randomization) device to draw a sample from the population.

The expectations $\pi_i = E_P[D_i]$ and $\pi_{ij} = E_P[D_i D_j]$ are the first and second order inclusion probabilities, respectively. The suffix P denotes the sampling design used to select the sample $\mathbf{s}$. The sample size is $n_s = D_1 + \cdots + D_N$. In probability

sampling, first order inclusion probabilities are frequently taken proportional to an auxiliary variable X: $\pi_i \propto x_i$, where x_i is the value of X for unit i ($i = 1, \ldots, N$). The rationale of this choice is simple: if the values of the variable of interest are positively correlated with (or, even better, approximately proportional to) the values of the auxiliary variable, then the Horvitz-Thompson estimator of the population mean will be highly efficient. More generally, in probability sampling the sampling design is constructed on the basis of *design variables*, namely statistical variates known in advance for all population units. Examples of design variables are strata / clusters indicators, variables used to construct inclusion probabilities, variables used in balancing equations, etc.. In the sequel, we will denote by $X_1^d, \ldots, X_k^d$ the design variables, and by x_{il}^d the value of variable X_l^d for unit i. Furthermore, let $\mathbf{X}_N^d$ be the $N \times k$ matrix of x_{il}^d values. The probability of drawing a sample $\mathbf{s}$, given $\mathbf{X}_N^d$ and $\mathbf{y}_N$, is denoted by $p(\mathbf{s}|\mathbf{X}_N^d, \mathbf{y}_N)$. The sampling design is *non-informative* if $p(\mathbf{s}|\mathbf{X}_N^d, \mathbf{y}_N)$ does not depend on $\mathbf{y}_N$: $p(\mathbf{s}|\mathbf{X}_N^d, \mathbf{y}_N) = p(\mathbf{s}|\mathbf{X}_N^d)$. Specifying $p(\mathbf{s}|\mathbf{X}_N^d)$ is equivalent to specify the probability distribution of $\mathbf{D}_N$.

Next, let us denote by $\mathbf{y}_\mathbf{s}$ the y-values corresponding to sampled units. In symbols: $\mathbf{y}_\mathbf{s} = \{y_i;\ i \in \mathbf{s}\}$. *Sample data* are composed by the pair $(\mathbf{s}, \mathbf{y}_\mathbf{s})$, namely by the observed y-values together with the corresponding (sampled) units. The knowledge of $(\mathbf{s}, \mathbf{y}_\mathbf{s})$ is equivalent to the knowledge of $\{(i, y_i);\ i \in \mathbf{s}\}$.

The y-values in the finite population parameter $\mathbf{y}_N$ may be thought as generated by a superpopulation model, namely by a N-dimensional r.v. $\mathbf{Y}_N$ possibly depending on unknown parameters $\boldsymbol{\theta}$. In the sequel, we will denote by $f(\mathbf{y}_N|\mathbf{X}_N^d; \boldsymbol{\theta})$ the density of $\mathbf{Y}_N$ given the design variables $\mathbf{X}_N^d$. Furthermore, $\mathbf{Y}_\mathbf{s}$ will denote the set of r.v.s Y_i with $i \in \mathbf{s}$.

Sometimes, together with the values of the character of interest, the values of p covariates playing the role of *auxiliary variables* $X_1, \ldots, X_p$ are observed. From now on, we will denote by x_{il} the value of variable X_l for unit i, and by $\mathbf{X}_N$ be the $N \times p$ matrix of x_{il} entries. Furthermore, $\mathbf{X}_\mathbf{s}$ denotes the sub-matrix of $\mathbf{X}_N$ composed by rows corresponding to sampled units. Note that some (or even all) of the auxiliary variables may be included among the design variables.

In general, two different inferential goals may be considered.

1. *Descriptive inference*, referring to statistical inference on finite population parameters, *i.e.* of functions of $\mathbf{y}_N$ such as the population mean $\bar{y}_N = \sum_{i=1}^N y_i/N$.
2. *Analytic inference*, referring to statistical inference on superpopulation parameters.

In non-probability samples the probability of drawing sample $\mathbf{s}$ is typically unknown. Generally speaking, the probability of selecting a sample $\mathbf{s}$ might depend on $\mathbf{y}_N$ (character of interest), on covariates $X_1, \ldots, X_p$, and on unknown parameters. There are various types of non-probability samples. A broad classification is made in [4].

a. Convenience sampling (non-probability sampling based on recruiting participants, for instance volunteer sampling, river sampling, etc.).
b. Sample matching (sample units are selected in order to match some important population characteristics, for instance quota sampling).

c. Network sampling, where sampled units are asked to contact other population units connected with them.

A major drawback of non-probability samples is that there are no real design variables. One may at most hope to identify a set of covariates that affect the selection process, and that play the role of "pseudo-design" variables. In the sequel, they will be denoted by $X_1, \ldots, X_p$, with the notation introduced above. In non-probability sampling (cfr. [5]) the probability of selecting a sample $\mathbf{s}$ could depend on $\mathbf{X}_N$, $\mathbf{y}_N$, as well as on unknown parameters $\boldsymbol{\psi}$. In symbols, such a probability will be denoted by $p(\mathbf{s}|\mathbf{X}_N, \mathbf{y}_N; \boldsymbol{\psi})$. In addition, let $f(\mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta})$ be be density function of $\mathbf{y}_N$ given $\mathbf{X}_N$, generally depending on unknown parameters $\boldsymbol{\theta}$. The joint density of $\mathbf{s}$ and $\mathbf{y}_N$ is then equal to:

$$f(\mathbf{s}, \mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta})p(\mathbf{s}|\mathbf{X}_N, \mathbf{y}_N; \boldsymbol{\psi}). \quad (1)$$

Eqn. (1) makes it clear the difference between probability and non-probability samples. In case of a (non-informative) probability sampling design, the term $p(\mathbf{s}|\mathbf{X}_N, \mathbf{y}_N; \boldsymbol{\psi})$ reduces to $p(\mathbf{s}|\mathbf{X}_N^d)$, which is *known*, and (1) becomes

$$f(\mathbf{s}, \mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_N|\mathbf{X}_N^d; \boldsymbol{\theta})p(\mathbf{s}|\mathbf{X}_N^d). \quad (2)$$

As a consequence, if $\mathbf{y}_\mathbf{s}$ denotes the set of y-values for sample units, if $\bar{\mathbf{s}}$ are the non-sampled population units, and if $\mathbf{y}_{\bar{\mathbf{s}}}$ are the corresponding y-values, under (2), two main conclusion holds for probability samples.

(*i*) Descriptive inference may be based on the distribution $p(\mathbf{s}|\mathbf{X}_N^d)$, which is known.
(*ii*) Analytic inference may be based on the conditional distribution $f(y_\mathbf{s}|\mathbf{X}_N^d; \boldsymbol{\theta})$, which is obtained as marginal of $f(y_N|\mathbf{X}_N^d; \boldsymbol{\theta})$ w.r.t. $\mathbf{y}_{\bar{\mathbf{s}}}$. In more detail, under the above conditions the sampling design is *ignorable*, and $f(y_\mathbf{s}|\mathbf{X}_N^d; \boldsymbol{\theta})$ is equivalent to $f(y_\mathbf{s}|\mathbf{s}, \mathbf{X}_N^d; \boldsymbol{\theta})$; cfr. [6] for limitations of the concept of ignorability of sampling design.

A comparison of (1) and (2) makes it clear that neither (*i*) nor (*ii*) hold in case of non-probability samples. In other terms, we cannot safely make neither analytic inference on superpopulation parameters by ignoring the (non-probabilistic) sampling mechanism, nor descriptive inference without accounting for the (unknown) $p(\mathbf{s}|\mathbf{X}_N, \mathbf{y}_N; \boldsymbol{\psi})$.

A careful analysis of the effect of non-probability sampling on the estimation of the population mean $\bar{y}_N$ is in [7]. Consider the sample mean $\bar{y}_\mathbf{s} = \sum_{i=1}^N y_i D_i / \sum_{i=1}^N D_i$, and let $f_N = \sum_{i=1}^N D_i / N$ be the sampling fraction. By elementary algebra, it is easy to that the estimation error is equal to

$$\bar{y}_\mathbf{s} - \bar{y}_N = \sqrt{f_N^{-1} - 1}\sqrt{S_{\bar{y}N}^2}\,Corr_N(y, D)$$

where

$$S^2_{yN} = \frac{1}{N-1} = \sum_{i=1}^{N}(y_i - \bar{y}_N)^2 \ \text{(finite population variance of } y_i\text{s)}$$

$$S^2_{DN} = \frac{1}{N-1} = \sum_{i=1}^{N}(D_i - f_N)^2 \ \text{(finite population variance of } D_i\text{s)}$$

$$Corr_N(y, D) = \frac{1}{N-1}\sum_{i=1}^{N} y_i(D_i - f_N) \Big/ (S_{yN}S_{DN}) \ \text{(finite population correlation of } y_i\text{s and } D_i\text{s)}.$$

Consider next simple random sampling without replacement (SRS, for short). As well known, the variance of the sample mean is

$$V_{SRS}(\bar{y}_{\mathbf{s}}) = \frac{1-f_N}{f_N} S^2_{yN}/N$$

and hence, if $MSE_{NP}(\bar{y}_{\mathbf{s}}) = E_{NP}[(\bar{y}_{\mathbf{s}} - \bar{y}_N)^2]$ denotes the Mean Squared Error w.r.t. the non-probabilistic design actually used, we get

$$\frac{MSE_{NP}(\bar{y}_{\mathbf{s}})}{V_{SRS}(\bar{y}_{\mathbf{s}})} = NE_{NP}[Corr_N(y, D)^2] = N D_I. \tag{3}$$

As remarked in [7], the term $D_I = E_{NP}[Corr_N(y, D)^2]$ is essentially related to the data quality, or better to the *quality of the data collection process*. The behaviour of D_I is radically different in probabilistic and non-probabilistic sampling. In the first case, $D_I = O(N^{-1})$ (for instance, under SRS $D_I = 1/(N-1)$), so that we get control on the behaviour of the mean squared error of $\bar{y}_{\mathbf{s}}$. Under non-probabilistic sampling, D_I is generally $O(1)$, and hence $\frac{MSE_{NP}(\bar{y}_{\mathbf{s}})}{V_{SRS}(\bar{y}_{\mathbf{s}})}$ increases as the population size does, unless $f_N \to 1$ as $N \to \infty$. Unfortunately, even in the case of VERY BIG data, this does not reasonably occur. As a result, independently of the sampling fraction, non-probabilistic sampling does generally produce a mean squared error much higher than SRS. A sampling fraction 5% obtained by SRS generally gives an estimation error better than a sampling fraction 60% obtained by non-probabilistic sampling. Eqn. (3) also stresses that the population size N plays the role of a "magnifying lens" that amplifies defectives of non-probability sampling.

Even worse, the problem is not in the sample mean. For instance, as a remedy one may think to use for sample units weights w_i, and to replace the sample mean by the Hájek estimator

$$T_H = \sum_{i=1}^{N} w_i D_i y_i \Big/ \sum_{i=1}^{N} w_i D_i.$$

Unless very special conditions (hardly ever met in practice) occur, considerations similar to the above ones hold.

These drastic conclusions are mitigated by the presence of non-responses. Some authors (cfr., *e.g.*, [8]) remark that non-responses make probability samples similar, in many respects, to non-probability samples. To tell the truth, there are impor-

tant differences. First of all, in probability samples extra-sample information, such as the values of the design variables, is available for both respondents and non-respondents. In the second place, in high quality sample surveys, paradata such as call records documenting on day, time, outcome of each contact attempt, mode of contact attempts, etc., are available for both respondents and non-respondents; cfr. [9], [10]. Paradata, that are frequently highly correlated with response rates, can be used to adjust for non-responses. For instance, in [11], call-back modelling incorporating information on the Level Of Effort (LOE) is used to produce weight adjustments that can compensate for the non-response bias in the respondent-only variables. The adopted model is Missing At Random (MAR), but this assumption can be tested.

In the subsequent sections, different approaches to make non-probability sample data "usable" are shortly reviewed.

3 Superpopulation approach

Another approach proposed in the literature is based on a superpopulation regression model. For the sake of simplicity, suppose the goal is to estimate the population total $Y = \sum_{i=1}^{N} Y_i$. With the notation already introduced, suppose further that for the N-dimensional random vector $\mathbf{Y}_N$, a regression model

$$\mathbf{Y}_N = \mathbf{X}_N \boldsymbol{\beta} + \boldsymbol{\varepsilon}_N \tag{4}$$

holds, where $\mathbf{X}_N$ is a $N \times m$ matrix containing the values x_{ij} of m regressors $X_1, \ldots, X_m$ for all population units, $\boldsymbol{\varepsilon}_N$ is the vector of errors, and $\boldsymbol{\beta}$ is the m-dimensional vector of regression coefficient. In [3] a predictive approach to the estimation of the population total Y is considered. Denote, as usual, by $\mathbf{y}_\mathbf{s}$ the vector of sample units, and by $\mathbf{X}_\mathbf{s}$ the sub-matrix of $\mathbf{X}_N$ composed by the rows corresponding to sample units. In practice, the vector $\boldsymbol{\beta}$ is estimated through its OLS estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_\mathbf{s}^T \mathbf{X}_\mathbf{s})^{-1} \mathbf{X}_\mathbf{s}^T \mathbf{y}_\mathbf{s}$, and then Y is predicted through

$$\widehat{Y} = \sum_{i \in \mathbf{s}} y_i + \sum_{i \in \bar{s}} \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \tag{5}$$

$\bar{\mathbf{s}}$ being the set of population units that are non in the sample $\mathbf{s}$. In the above mentioned paper, the problem of estimating the (model-based) variance of (5) is also studied.

A nice feature of the oulined approach is its simplicity. Its main weakpoint is that it is essentially based on the assumption of *ignorability* of the sampling design (cfr. [6]). In the present case, ignorability basically means that the conditional distribution of $\mathbf{Y}_\mathbf{s}$ given $\mathbf{X}_\mathbf{s}$ and $\mathbf{s}$ coincides with the conditional distribution of $\mathbf{Y}_\mathbf{s}$ given $\mathbf{X}_\mathbf{s}$. Intuitively speaking, the selection process of units from the population does not play any role.

If the ignorability assumption does not hold, then (5) can be viewed as a predictor of the (expectation of the) total Y *conditionally on* $\mathbf{s}$. However, it is doubtful whether this kind of inference is of real interest.

Criticisms to the assumption of ignorability are made in [6], where it is also shown that the estimator $\widehat{\boldsymbol{\beta}}$, under a non-ignorable sampling design, could be biased and inconsistent, as well. In case of non-probability samples, the assumption of ignorability cannot be tested, so that one has entirely to rely on a very strong, non-testable assumption with the *caveat* that its violation could produce, as an effect, an estimator (5) severely biased and inconsistent.

4 Weighting non-probability sample data through modelization of the selection process

Weighting data techniques are used as a remedy for bias due to the uncontrolled nature of the sample unit selection process. This approach is used in several papers, under various assumptions and developments: cfr. [12], [13], [8], [3] and references therein. The common feature of the above papers is the use of propensity score (cfr. [14]) to construct weights to account for bias in the non-probability sample.

The method requires availability of a reference sample $\mathbf{s}^*$ collected through a non-informative sampling design (or a census), with common variables $X_1, \ldots, X_p$ explaining the selection of units of both the non-probability sample $\mathbf{s}$ and the probability sample $\mathbf{s}^*$. Let $\mathbf{x}_i$ be vector of X variables for unit i, assume that units are independently selected in both $\mathbf{s}$ and $\mathbf{s}^*$, and denote by D_i, D_i^* the corresponding sample membership indicators of unit i in $\mathbf{s}$ and $\mathbf{s}^*$, respectively. In all the above papers it is tacitely admitted that design variables are available for all units in $\mathbf{s}^*$, as well as for all units in $\mathbf{s}$, and that both probability and non-probability samples are non-informative, conditionally on $\mathbf{x}_i$, $i \in \mathbf{s} \cup \mathbf{s}^*$.

The idea is to consider the non-probability sample $\mathbf{s}$ as the set of treated subjects in an observational study, and the reference sample $\mathbf{s}^*$ as the set of untreated subjects. The propensity score is then defined as

$$p(\mathbf{x}_i) = P(D_i^* = 1 | \mathbf{x}_i;\ i \in \mathbf{s} \cup \mathbf{s}^*). \tag{6}$$

In [13] a logistic model is used for the propensity score (6). The estimated propensity score $\widehat{p}(\mathbf{x})$ is then used to partition $\mathbf{s}$, $\mathbf{s}^*$ into C classes $\mathbf{s}_c$, $\mathbf{s}_c^*$; $c = 1, \ldots, C$, based on increasing vales of $\widehat{p}(\mathbf{x})$ ($C = 5$ classes are used in the above paper). Next, for each class c an adjustment factor

$$f_c = \frac{\sum_{i \in \mathbf{s}_c^*} w_i^* \big/ \sum_{i \in \mathbf{s}^*} w_i^*}{\sum_{i \in \mathbf{s}_c} w_i \big/ \sum_{i \in \mathbf{s}} w_i},\ \ c = 1, \ldots, C$$

where w_i, w_i^* are appropriate weights (for units in the non-probability and probability sample, respectively). Finally, for each unit in $\mathbf{s}$ an adjusted weight

$$w_i^{adj} = f_c w_i;\ i \in \mathbf{s}_c,\ c = 1, \ldots, C. \tag{7}$$

are defined, and used as weights of units in the non-probability sample.

Variation on similar ideas are in [3], [8]. Again, units in the non-probability sample are considered as "treated", whilst units in the reference probability sample are considered as "untreated". Let T_i be equal to 1 for treated units, and $T_i = 0$ for untreated units. Assuming that sampling fractions in $\mathbf{s}$, $\mathbf{s}^*$ are small (in case of big data, this could be questionable), it can be easily shown that

$$P(D_i = 1|\mathbf{x}_i) = c\,P(D_i^* = 1|\mathbf{x}_i)\,\frac{P(T_i = 1|\mathbf{x}_i)}{1 - P(T_i = 1|\mathbf{x}_i)} \tag{8}$$

where c is a proportionality constant. Now, the term $P(D_i^* = 1|\mathbf{x}_i)$ is known, or estimable *via* a regression of inclusion probabilities w.r.t. $\mathbf{x}_i$s in the probability sample $\mathbf{s}^*$. The term $P(T_i = 1|\mathbf{x}_i)$ is essentially a propensity score, that can be estimated *via* logistic regression, or by other methods from $\mathbf{s} \cup \mathbf{s}^*$. This would lead to adjustment weights equal to

$$w_i^{adj} = \frac{1}{\widehat{P}(D_i^* = 1|\mathbf{x}_i)}\,\frac{1 - \widehat{P}(T_i = 1|\mathbf{x}_i)}{\widehat{P}(T_i = 1|\mathbf{x}_i)}. \tag{9}$$

The weakpoint of methods based on propensity score as those outlined here is that they rely on the *strong ignorability condition* by Rosenbaum and Rubin [14]:

1. T_i and Y_i are independent conditionally on $\mathbf{x}_i$, for all $i \in \mathbf{s} \cup \mathbf{s}^*$;
2. $0 < P(T_i = 1|\mathbf{x}_i) < 1$ for each $\mathbf{x}_i$.

In practice, the above two assumption imply that (i) we may ignore sample membership for inference, and (ii) the population correlation among D_is and y_is is negligible (its expectation in $O(N^{-1/2})$, in Meng's [7] language).

Strong ignorability hardly ever holds for all study variables of interest,. Furthermore, as shown in a simulation study in [12], the use of propensity score methods may reduce the bias of estimates, but at the price of a considerably increased variance.

5 Approaches based on data integration

Approaches based on combining data from probabilistic and non-probabilistic samples through data integration techniques are proposed by Rivers [15] and in a series of papers by Kim *et al.* [17], [18], [19], [16].

Suppose that for all units i in the non-probability sample $\mathbf{s}$ the values $(y_i, \mathbf{x}_i)$ of both the variable of interest Y and the covariates $\mathbf{X}$ are observed, whilst in the probability sample $\mathbf{s}^*$ only values $\mathbf{x}_i$ are collected. From now on, observed values y, $\mathbf{x}$ in the non-probability sample $\mathbf{s}$ will be denoted by y_i^{NP}, $\mathbf{x}_i^{NP}$, and observed values

$\mathbf{x}$ in the probabilistic sample $\mathbf{s}^*$ will be denoted by $\mathbf{x}_i^P$; y values in sample $\mathbf{s}^*$, denoted by y_i^P, are missing because unobserved.

In [15] it is proposed to impute missing values y_is in sample $\mathbf{s}^*$ through *kNN* (nearest neighbour) method. For each unit $i \in \mathbf{s}^*$, consider the set $\{j_1, \ldots, j_k\}$ composed by the $k \geq 1$ units $j \in \mathbf{s}$ that are closest to $\mathbf{x}_i^P$ in terms of an appropriate distance $d(\mathbf{x}_i^P, \mathbf{x}_j^{NP})$. Then, the missing value y_i^P is imputed through a value $\widehat{y}_i^P = g(y_{j_1}^{NP}, \ldots, y_{j_k}^{NP})$. Estimation of finite population parameters (such as $\overline{y}_N = \sum_i y_i/N$), or superpopulation parameters, is based on imputed values $\widehat{y}_i^P$, $i \in \mathbf{s}^*$. The method works under the condition that the unknown design that has generated the non-probability sample $\mathbf{s}$ is non-informative, and hence ignorable. Unfortunately this assumption is irremissible and non-testable, and this is a serious limitation of the method.

A similar approach is also considered in [16], where a semi-parametric model $E[Y|\mathbf{X} = \mathbf{x}] = m(\mathbf{x}; \boldsymbol{\beta}_0)$ is considered, $\boldsymbol{\beta}_0$ being un unknown p-dimensional vector and $m(\cdot, \cdot)$ a known function.

The assumptions on which this approach rests are essentially two, both non-testable.

a. The superpopulation model for $(Y_i, \mathbf{X}_i)$ satisfies the relationship $f(y_i|\mathbf{X}_i, D_i = 1) = f(y_i|\mathbf{X}_i)$. This is a form of ignorability of the sampling mechanism that has generated the non-probability sample $\mathbf{s}$, and it is crucial to impute missing y-values in the probability sample $\mathbf{s}^*$.
b. $P(D_i = 1|\mathbf{X}_i = \mathbf{x}) > 0$ for every $\mathbf{x}$. This assumption essentially avoids undercoverage.

A different approach, where data integration (essentially, a variation of record linkage) is used to estimate weights for the non-probability sample, is proposed in [19]. Assume that for all units i in the probability sample $\mathbf{s}^*$ the sample membership indicators D_i of the non-probability sample $\mathbf{s}$ are available (possibly through a simplified version of record linkage techniques). Let π_i^P be the first order inclusion probabilities for units in the probability sample $\mathbf{s}$, and suppose further that

1. the selection mechanism of the non-probability sample is non-informative conditionally on the covariates x_i: $P(D_i = 1|y_i, \mathbf{x}_i) = P(D_i = 1|\mathbf{x}_i)$;
2. $P(D_i = 1|\mathbf{x}_i)$ can be modelled through a parametric model $P(D_i = 1 = 1|\mathbf{x}_i) = p(\mathbf{x}_i^T \boldsymbol{\lambda})$.

In the sequel, the symbol $p_i(\boldsymbol{\lambda})$ will be used to denote $p(\mathbf{x}_i^T \boldsymbol{\lambda})$. The value of $\boldsymbol{\lambda}$ is then estimated by maximizing the pseudo log-likelihood function

$$l(\boldsymbol{\lambda}) = \sum_{i \in \mathbf{s}^*} \frac{1}{\pi_i} \{D_i \log p_i(\boldsymbol{\lambda}) + (1 - D_i) \log(1 - p_i(\boldsymbol{\lambda}))\}.$$

If $\widehat{\boldsymbol{\lambda}}$ denotes the maximum pseudo-likelihood estimator of $\boldsymbol{\lambda}$, the population mean $\overline{y}_N$ is then estimated through the Hájek type estimator (from the non-probability sample)

$$\frac{\sum_{i\in\mathbf{s}} p_i(\widehat{\boldsymbol{\lambda}})^{-1} y_i}{\sum_{i\in\mathbf{s}} p_i(\widehat{\boldsymbol{\lambda}})^{-1}} \tag{10}$$

Additional results on the variance estimation of (10) are in [19].

6 Conclusions

Methodologies and results reviewed in the present paper, as well as many others, make it clear that the naive use of non-probability sample data may be dangerous, and could produce highly erroneous inferential conclusions. Several attempts are made in the literature to propose remedies to this drawdack. They require extra-sample information in the form of a probability reference sample, as well as strong, non-testable assumptions. Their effectivity needs to be more deeply explored, for both theoretical and practical purposes.

References

1. Neyman, J.: On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, **97**, 558–625 (1934)
2. Tillé, Y.: Sampling Algorithms, Springer, New York (2006)
3. Elliot, M R., Valliant, R.: Inference for Nonprobability Samples. Statistical Science, **32**, 249–264 (2017)
4. Baker, R., Brick, J M., Bates, N A., Battaglai, M., Couper, M P., Denver, J A., Gile, K., Tourangeau, R.: Summary report of the AAPOR Task Force on Non-probability Sampling. Journal of Survey Statistics and Methodology, **1**, 90–143 (2013)
5. Smith, T M F.: On the Validity of Inferences from Non-random Sample. Journal of the Royal Statistical Society A, **146**, 394–403 (1983)
6. Pfeffermann, D.: The Role of Sampling Weights When Modeling Survey Data. International Statistical Review , **61**, 317–337 (1993)
7. Meng, X- L.: Statistical Paradises and Paradoxes in Big Data(I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. The Annals of Applied Statistics, **12**, 686–726 (2018)
8. Elliot, M R.: Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. Survey Practice, **2**, https://doi.org/10.29115/SP-2009-0025 (2009)
9. Couper, M P.: Measuring survey quality in a CASIC environment. In: *JSM Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA, 41–49 (1998)
10. Kreuter, F. , Couper, M P., Lyberg, L.: The use of paradata to monitor and manage survey data collection. In: *JSM Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA, 282–296 (2010)
11. Biemer, P P., Chen, P., Wang, K.: Using level-of-effort paradata in non-response adjustments with application to field surveys. Journal of the Royal Statistical Society Series A, **176**, 147–168 (2013)
12. Lee, S.: Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. Journal of Official Statistics, **22**, 329–349 (2006)

13. Lee, S., Valliant, R.: Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociological Methods and Research, **37**, 319–343 (2009)
14. Rosenbaum, P R., Rubin, D B.: The Central Role of the Propensity Score in Observational Studies for Treatment Effects. Biometrika, **70**, 41–55 (1983)
15. Rivers, D.: Sampling for web surveys. In: *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA, 4127–4134 (2007)
16. Kim, J K., Park, T., Vhen, Y., Wu, C.: Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society Series A, **184**, 941–963 (2021)
17. Kim, J K., Berg, E., Park, T.: Statistical matching using fractional imputation. Survey Methodology, **42**, 19–40 (2016)
18. Park, S., Kim, J K., Stukel, D.: A measurement error model for survey data integration: combining information from two samples. Metron, **75**, 345–357 (2017)
19. Kim, J K., Wang, Z.: Sampling Techniques for Big Data Analysis. International Statistical Review, **42**, S177–S191 (2019)

Combining Big Data with probability survey data: a comparison of methodologies for estimation from non-probability surveys

Maria del Mar Rueda, Ramn Ferri-García, Luis Castro-Martín

Abstract The growing adoption of new technologies in our society, along with the restrictions caused by the COVID-19 pandemic, have favored the use of non-probability samples to obtain information from a population of interest. Despite their cost and immediacy, these samples entail a number of drawbacks, specially regarding their selection bias. For this reason, several design-based and model-based methods have been developed to mitigate this selection bias. Design-based methods, such as Propensity Score Adjustment and Kernel Weighting, aim to estimate the probability of an individual of the population of being included in the non-probability sample, while model-based methods, such as Statistical Matching, predict the value of the target variable in a probability sample, where the target variable has not been measured. In this work, we describe the methods, compare them according to their advantages and disadvantages, and explain how Machine Learning techniques could boost these methods. Finally, we give some recommendations on further research lines regarding estimation from nonprobability samples.
Abstract Le restrizioni dovute alla pandemia COVID-19 hanno favorito luso di campioni non probabilistici. Nonostante i vantaggi in termini di costi e di tempi, tali campioni sono affetti da selection bias. Nel lavoro sono discussi approcci alternativi, sia design-based sia model-based, per la riduzione di tale distorsione.

Key words: Non-probability sampling, Kernel Weighting, Propensity Score Adjustment, Statistical Matching

1 Introduction

The global Public Health emergency due to the COVID-19 pandemic has motivated a large number of studies on the magnitude, characteristics and evolution of

Department of Statistics and Operational Research, University of Granada, Spain. e-mail: mrueda@ugr.es, rferri@ugr.es, luiscastro193@gmail.com

its impact, and seroprevalence in the population. Some of these studies are based on probability surveys that have the advantage of making valid inferences about the study population. Probability sampling is the usual method for obtaining a representative sample from a target finite population. Besides, collecting a strict probability sample is almost impossible in certain areas due to unavoidable issues.

New data sources have been considered as alternatives to survey sample data. Examples are big data and web surveys that have the potential of providing estimates in nearly real time, an easier data access and lower data collection costs than traditional probability sampling. Big data and web surveys not only provide an economical means of data collection, they also enable real-time access to statistics, which is of great importance in volatile situations such as the one created by the pandemic. Official agencies and health research centres have adapted their methods and have applied this type of approach in order to obtain rapid information, to monitor the evolution of the pandemic over time and to be in a position to take restrictive measures to alleviate the crisis on the basis of scientific criteria.

The data generating process of such sources is non-probabilistic, given that the probability of being part of the sample is not known and/or is null for some groups of the target population. There are serious issues on the use of non-probability survey samples; the most relevant is that the data generating process is unknown and may have serious coverage, nonresponse and selection biases which may not be ignorable and could deeply affect estimates [11]. These biases tend to be more disrupting as the population size is larger, regardless of the sample size [16].

Different inference procedures are proposed in the literature to correct for selection bias introduced by non-random selection mechanisms. A first class of methods uses model-based approach that relies on the specification of an appropriate superpopulation model. In these methods auxiliary variables must be available for each unit of the observed and the unobserved parts of the population [1] that is complicated in practice. Other studies combine a non-probability sample (or big data) with a reference probability sample to construct models for units in the latter or to adjust selection probabilities. This situation is much more common in practice and has given rise to a great variety of estimators. In this work we are going to review these techniques and we will see the properties, advantages and disadvantages of using each one.

2 Background

Suppose that the finite population U consists of $i = 1,...,N$ different and identifiable units. Let y be a survey variable and y_i be the y-value attached to the i-th unit, $i = 1,...,N$.

Let s_v be a volunteer sample of size n_v, self-selected from U_v a subset of U observing the study variable y.

This sample suffers from two fundamental methodological problems. The first one is self-selection: selection probabilities are unknown. Therefore, no unbiased

estimates can be computed, nor can the accuracy of estimates be established. The second problem is under-coverage. Since data is collected from U_v, people from $U - U_v$ will never be able to participate in this survey. This means research results can not apply to the complete population. We analyse these two problems in more detail.

Without any auxiliary information, the population mean $\bar{Y}$, is usually estimated with the unweighted sample mean

$$\hat{\bar{Y}} = \sum_{k \in s_V} \frac{y_k}{n_v} \tag{1}$$

that makes biased estimates.

Let I_v be an indicator variable of an element being in s_v:

$$I_{vi} = \begin{cases} 1 \; i \in s_v \\ 0 \; i \notin s_v \end{cases} \tag{2}$$

Each element i in the population has unknown probability π_i^v of participating in the survey (or propensity). The expected value of the sample mean is equal to:

$$E(\hat{\bar{Y}}) = \sum_{k \in U_V} \frac{y_k \pi_i^v}{\bar{\pi}^v}$$

where $\bar{\pi}^v$ is the mean of all response propensities. The bias of this estimator is:

$$B(\hat{\bar{Y}}) = E(\hat{\bar{Y}}) - \overline{Y} = \frac{C_{\phi y}}{\bar{\pi}^v} \frac{N_v}{N} (\overline{Y}_v - \overline{Y}_{NV})$$

being $\overline{Y}_v$ the population mean of population U_v, $\overline{Y}_{NV}$ the population mean of $U - U_v$, $C_{\phi y}$ the covariance between the target variable and the response probabilities.

Thus, the size and direction of the bias depend on two factors: the proportion of the population with no chance of inclusion in the sample (coverage) and differences in the inclusion probabilities among the different members of the sample with a non-zero probability of taking part in the survey (selection). This bias cannot be estimated in practice for most survey variables of interest.

Some methods has been developed to treat the non-probability samples. In the next sections we show how to use some of the sampling techniques to reduce the selection bias and make the resulting analysis valid.

3 Propensity score weighting

We consider the situation where there is a probability sample available. Let s_r be a reference probability sample selected from U under a sampling design (s_d, p_d) with $\pi_i = \sum_{s_r \ni i} p_d(s_r)$ the first order inclusion probability and $d_i = 1/\pi_i$ the basic sampling weight for the i-th individual. Let us assume that in s_r we observe some other

study variables which are common to both samples, denoted by x. The available data are denoted by $\{(i,y_i,x_i), i \in s_v\}$ and $\{(i,x_i), i \in s_r\}$. We are interested in estimating a linear parameter $\theta_N = \sum_U a_i y_i$ being a_i known constants.

The propensity score of the individual can be formulated, following notation in [9], as the expected value of I conditional on her/his target variable and covariates' value:

$$\pi_i^v = E[I_{vi}|\mathbf{x}_i, y_i] = P(I_{vi} = 1|\mathbf{x}_i, y_i) \tag{3}$$

The propensity for an individual to take part on the non-probability survey is obtained by training a predictive model (often a logistic regression) on the dichotomous variable, I_{s_V}, which measures whether a respondent from the combination of both samples took part in the volunteer survey or in the reference survey. Covariates used in the model, $\mathbf{x}$, are measured in both samples (in contrast to the target variable which is only measured in the non-probability sample), thus the formula to compute the propensity of taking part in the volunteer survey with a logistic model, π^v, can be displayed as

$$\pi^v(\mathbf{x}) = \frac{1}{e^{-(\gamma^T \mathbf{x})} + 1} \tag{4}$$

for some vector γ, as a function of the model covariates.

Once the pseudo maximum likelihood estimator $\hat{\gamma}$ is obtained, we calculate $\hat{\pi}^v(\mathbf{x}_k)$, the estimated response propensity for the individual k of the volunteer sample using covariates $\mathbf{x}$ and then we can obtain a PS estimator in several ways.

3.1 Inverse of propensity score weighting

When the population size is much greater than the volunteer sample size, one can simply use the inverse of the estimated response propensity as a weight for constructing the estimator [20]:

$$\hat{\theta_{NPSA1}} = \sum_{k \in s_V} a_k y_k / \hat{\pi}^v(\mathbf{x}_k) = \sum_{k \in s_V} a_k y_k w_k^{PSA1}. \tag{5}$$

Alternatively, the approach proposed in [18] can be used regardless of sample size. Weights are defined as

$$w_k^{PSA2} = \frac{1 - \hat{\pi}^v(\mathbf{x}_k)}{\hat{\pi}^v(\mathbf{x}_k)} \tag{6}$$

and resulting estimator for the parameter θ_N is given by

$$\hat{\theta}_{N_{PSA2}} = \sum_{k \in s_V} a_k y_k w_k^{PSA2} \tag{7}$$

3.2 Propensity score adjustment by subclassification

Unlike the inverse of PS method (IPSW), the PS adjustment by subclassification (PSAS) method fits the logistic regression model to the combined volunteer and probabilistic survey sample (Lee and Valliant, 2009) to estimate propensity scores. Instead of estimating the participation probability for each unit, the PSAS method uses the estimated propensity scores to measure the similarity of participants in the volunteer and the survey samples with regard to their covariate values. Specifically, the combined sample is first sorted by the estimated propensity score $\hat{\pi}^v$ and then partitioned into C subclasses. There are multiple ways to form the subclasses. For example the combined sample is sorted and then divided into C classes ([10] recommend the use of five classes) according to each individual's propensity score. The new weights for individuals in the volunteer sample in class c are then calculated as follows:

$$w_j^{PSAS} = \frac{\sum_{k \in s_r^c} / \sum_{k \in s_r}}{\sum_{j \in s_v^c} / \sum_{j \in s_v}} \tag{8}$$

where s_r^c and s_v^c are individuals from the reference sample and the volunteer sample respectively, belonging to the c-th class.

The PSAS estimator of θ_N is is given by

$$\hat{\theta_{NPSAS1}} = \sum_{k \in s_V} a_k y_k w_k^{PSAS} \tag{9}$$

Other method is described in [19]. The process is similar: sort the combined sample by $\hat{\pi}^v$; split the combined sample into g classes, each of which has about the same number of cases in the combined sample; and compute an average propensity, $\bar{\pi}_g$ within subclass g. Use $\bar{\pi}_g$ as the weight adjustment for every person in the subclass. Resulting estimator is:

$$\hat{\theta_{NPSAS2}} = \sum_g \sum_{k \in s_{Vg}} a_k y_k / \bar{\pi}_g. \tag{10}$$

The IPSW method has less bias when the propensity model is correctly specified but can produce extreme weights, which can inflate variances of the weighted estimators. PSAS method is less sensitive to model misspecifications, avoids extreme weights ([21]) and yields less variable estimates. However, the PSAS method is less effective at bias reduction ([19]).

The efficacy of PSA at removing selection bias has been proven when prognostic covariates are chosen [14] and further adjustments, such as calibration, are applied in the estimations [15, 19, 12]. where the reductions in bias were not sufficiently large and consistent in general for estimates to be seen as broadly unbiased.

4 Kernel weighting method

[21] propose a kernel weighting method to create pseudoweights for cohort studies that can be used in our context.

Kernel weighting method (KWM) uses propensity scores to measure the similarity of the covariate distributions between the volunteer and the probabilistic samples. Let be $d(x_i^{(r)}, x_j^{(v)}) = \hat{\pi}^v(\mathbf{x}_i^{(r)}) - \hat{\pi}^v(\mathbf{x}_j^{(v)}$ the distance of the estimated propensity score from $i \in s_r$ and $j \in s_v$. A kernel function is used to smooth the distances:

$$k_{ij} = \frac{K\{d(x_i^{(r)}, x_j^{(v)})\}/h}{\sum_{j \in s_v} K\{d(x_i^{(r)}, x_j^{(v)})\}/h} \tag{11}$$

for $i \in s_r$ and $j \in s_v$, being $K(\cdot)$ a kernel function and h the corresponding bandwidth.

The weight for $j \in s_v$ is given by:

$$w_j^{KW} = \sum_{i \in s_r} k_{ij}/\pi_i \tag{12}$$

and the final estimator is given by:

$$\hat{\theta_{N_{KW}}} = \sum_{k \in s_V} a_k y_k w_k^{KW} \tag{13}$$

This estimator is less sensitive than PSAS estimator to model misspecification while avoiding the extreme weights of the IPSW method ([21]).

5 Statistical Matching

Statistical matching (or mass imputation approach) is a model-based approach introduced by [17] and further developed by [3] for nonresponse in probability samples. The idea in this context is to model the relationship between y_k and x_k using the volunteer sample s_V in order to predict y_k for the reference sample.

Suppose that the finite population $\{(i, y_i, x_i), i \in U\}$ can be viewed as a random sample from the superpopulation model:

$$y_i = m(x_i) + e_i, i = 1, 2, ..., N,$$

where $m(x_i) = E_m(y_i|x_i)$ and the random vector $e = (e_1, ..., e_N)'$ is assumed to have zero mean.

The volunteer sample is used as a training dataset, and imputation is applied to all units in the probability sample. Thus the matching estimator is given by:

$$\hat{\theta}_{SM} = \sum_{s_r} a_k \hat{y}_k d_k$$

being $\hat{y}_k$ the predict value of y_k.

Usually the linear regression model is considered for estimation, $E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta$ that is easy to implement in most of the existent statistical packages, but several drawbacks have to be considered. Parametric models require assumptions regarding variable selection, the functional form and distributions of variables, and specification of interactions. If any of these assumptions are incorrect, the bias reduction could be incomplete or nonexistent.

6 Double robust method

[9] construct a doubly robust estimators of the finite population mean using the estimated propensity scores as well as an outcome linear regression model. Following the idea of these authors, the estimator of a linear parameter for a general regression model is defined as:

$$\hat{\theta}_{DR} = \sum_{s_r} d_k a_k \hat{y}_k + \sum_{s_v} a_k (y_k - \hat{y}_k)/\hat{\pi}^v(\mathbf{x}_k) \tag{14}$$

This estimator is doubly robust in the sense that it is a consistent estimator of θ_N if either the propensity score model or the outcome regression model is correctly specified.

7 The role of Machine Learning in estimation from non-probability samples

The methods introduced in this work are mostly based in prediction techniques. In the case of Propensity Score Adjustment, we must fit a propensity model to predict the probability of an individual to be included in the non-probability sample, while in the case of Statistical Matching, we directly fit a predictive model for the target variable in order to predict the value of that variable for individuals in the probability sample (where the target variable has not been measured). The predictive model often considered in literature for both methods is the generalized linear model: linear regression for Statistical Matching and logistic regression for Propensity Score Adjustment. However, the development of Machine Learning techniques for prediction has enlarged the set of possibilities for this task, offering crucial advantages for the Big Data context such as more flexibility in the specifications of the models (which learn the relationships from the data itself) or more computational efficiency.

In this regard, several Machine Learning techniques have been suggested as promising alternatives to logistic regression for the estimation of propensity scores. [7] presented a simulation study using decision trees, k-nearest neighbors, Naive Bayes, Random Forest and Gradient Boosting Machine that support the view of that machine learning methods can be used at removing selection bias in non-probability

samples. All of those algorithms along with Discriminant Analysis and Model Averaged Neural Networks are used for propensity estimation in the study of [4], which compares the use of linear models and Machine Learning prediction algorithms in propensity estimation. In addition, boosting techniques have also been applied in the Kernel Weighting approach, showing good results overall [13].

In the Statistical Matching context, Machine learning tries to extract the relationship between the target variable and the covariates through a learning algorithm without a priori data model. [4] apply Machine Learning prediction techniques to build Statistical Matching estimators, and compare their performance with PSA estimators. Results show that Statistical Matching provides better results than PSA on bias reduction. Besides, linear models and k-nearest neighbors provided in average better results, in terms of bias reduction, than more complex models such as GBM and Bagged Trees.

8 Advantages and disadvantages of each method

When applying the aforemetioned methods in real-world scenarios, where population parameters are to be estimated using non-probability samples, several features have to be taken into account in order to choose the method that provides the best results. Comparative studies made in [4] and [6] show an advantage (in terms of bias and Mean Squared Error) of model-based methods, such as Statistical Matching, over design-based methods such as Propensity Score Adjustment.

On the other hand, the main advantage of design-based methods relies on the fact that they are able to provide a single weights vector, obtained from a single statistical adjustment, that can be used for the estimation of any population parameter of any target variable that can be estimated from the sample. This is particularly usefult in multipurpose surveys, given that adjusting one model for each target variable (as we would do in Statistical Matching or doubly robust estimators) could increase the risk of model misspecifications, apart from being largely difficult to implement if the number of target variables is large.

However, in those contexts with multiple target variables, it is common that the covariates used in the propensity estimation model do not constitute the optimal subset of variables for the estimation of some target variables. Propensity modelling requires the use of prognostically important covariates which are related to the target variables, and those variables are likely to be different as the target changes. This drawback can be partly mitigated with weight smoothing [2], where adjustment weights (obtained with Propensity Score Adjustment) are substituted by their predictions from models that aim to predict the values of the weights' vector using the target variables themselves. According to the simulation study from [8], the use of weight smoothing in non-probability surveys can increase the efficiency of the estimators.

9 Further research lines

Future research on estimation from non-probability surveys should consider the inclusion of several research lines. The inclusion of design weights in Propensity Score Adjustment should be thoroughly studied. Although [9] developed a consistent estimator that involves design weights under the logistic regression model, other weigthing strategies could be more adequate for other choices that can be considered for estimation of propensities.

Another important issue commonly faced in non-probability surveys is the mitigation of the bias produced by MNAR mechanisms. The treatment of this kind of bias is the most troublesome overall, and it is often not considered in adjustments as they usually work under the assumption of ignorable nonresponse.

Other research lines include the development of theoretical properties, although some advanced have been recently made in this regard [9, 5], and the establishment of a framework of data preprocessing techniques that could be used for modelization, such as dealing with class imbalance (which is particulatly prevalent in PSA for large-scale online surveys) or hyperparameter tuning.

Acknowledgements This study was partially supported by Ministerio de Educación y Ciencia (PID2019-106861RB-I00, Spain), FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (FQM170-UGR20) and IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033

References

1. Buelens, B., J. Burger, and J.A. vanden Brakel: Comparing inference methods for non-probability samples. Int. Stat. Rev., **86**(2), 322–43 (2018)
2. Beaumont, J. F.: A new approach to weighting and inference in sample surveys. Biometrika **95**(3), 539–553 (2008).
3. Beaumont, J.F.; Bissonnette, J.: Variance estimation under composite imputation: The methodology behind SEVANI. Surv. Methodol. **37**,171–179 (2011).
4. Castro-Martn, L., Rueda, M. D. M., and Ferri-Garca, R.: Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. Mathematics **8**(6), 879 (2020).
5. Castro-Martn, L., Rueda, M. D. M., and Ferri-Garca, R.: Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment. Mathematics **8**(11), 2096 (2020).
6. Castro-Martn, L., del Mar Rueda, M., and Ferri-Garca, R.: Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys. J. Comput. Appl. Math. **404**, 113414 (2022).
7. Ferri-Garca, R., and Rueda, M. D. M.: Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS one **15**(4), e0231500 (2020).
8. Ferri-Garca, R., Beaumont, J. F., Bosa, K., Charlebois, J., and Chu, K.: Weight smoothing for nonprobability surveys. TEST, 1–25 (2021).

9. Chen, Y., Li, P., Wu, C.: Doubly Robust Inference With Nonprobability Survey Samples. J. Am. Stat. Assoc. **115**(532), 2011–2021 (2019).
10. Cochran, WG.: The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. Biometrics **24**(2), 295–313 (1968).
11. Elliott, M.R., Valliant, R.: Inference for nonprobability samples. Stat. Sci. **32**, 249–264 (2017)
12. Ferri-Garca, R., Rueda, MM. Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat. Oper. Res. T., **42**(2), 159-182 (2018).
13. Kern, C., Li, Y., and Wang, L.: Boosted kernel weightingusing statistical learning to improve inference from nonprobability samples. J. Surv. Stat. Methodol., **9**(5), 1088–1113 (2021).
14. Lee, S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J. Off. Stat. **22**, 329–349 (2006).
15. Lee, S. Valliant, R.: Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociol. Methods Res. **37**, 319–343, (2009).
16. Meng, X.-L.: Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. Ann. Appl. Stat., **12(2)**, 685–726 (2018).
17. Rivers, D.: Sampling for web surveys. In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA, 1 August 2007.
18. Schonlau, M., Couper, M.: Options for Conducting Web Surveys. Stat. Sci. 32(2), 279–292 (2017).
19. Valliant, R., Dever, JA.: Estimating Propensity Adjustments for Volunteer Web Surveys. Sociol. Method. Res. **40**(1), 105–137 (2011).
20. Valliant, R.: Comparing alternatives for estimation from nonprobability samples. J. Surv. Stat. Methodol., **8**, 2, 231–263 (2020).
21. Wang, GC., Katki, L.: Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. J. R. Stat. Soc. **183**, 1293–1311 (2020).

A Bayesian approach for combining probability and non-probability samples surveys

Un approccio Bayesiano per combinare indagini da campioni probabilistici e non-probabilistici

Salvatore Camilla, Biffignandi Silvia, Sakshaug Joseph, Struminskaya Bella, Wisniowski Arkadiusz

Abstract
Our paper proposes a method of combining probability and non-probability samples to improve analytic inference on logistic regression model parameters. A Bayesian framework is considered where only a small probability sample is available and the information from a parallel non-probability sample is provided naturally through the prior. A simulation study is run applying several informative priors. Comparisons on the performance of the models are studied with reference to their mean-squared error (MSE). In general, the informative priors reduce the MSE or, in the worst-case scenario, perform equivalently to non-informative priors.

Abstract
Si propone di combinare campioni probabilistici e non per migliorare l'inferenza sui parametri del modello di regressione logistica con approccio Bayesiano. Si assume che sia disponibile un piccolo campione probabilistico e le informazioni provenienti da un grande campione non-probabilistico vengono fornite tramite la distribuzione a priori. Viene condotto uno studio tramite simulazione in cui si confrontano varie distribuzioni a priori informative. In generale, l'utilizzo di prior informative riduce l'errore quadratico medio o, nel caso peggiore, la performance è la stessa.

Key words: Selection Bias, Data Integration, Bayesian Inference

[1] Salvatore Camilla, University of Milano-Bicocca, c.salvatore4@campus.unimib.it;
Biffignandi Silvia, CESS, biffisil@teletu.it;
Sakshaug Joseph, German Institute for Employment Research, joe.sakshaug@iab.de;
Struminskaya Bella, Utrecht University, b.struminskaya@uu.nl;
Wisniowski Arkadiusz, University of Manchester, a.wisniowski@manchester.ac.uk;

1 Introduction

Probability-based surveys (PS) are known to have higher data quality but are expensive and subject to relatively small sample sizes. Nonresponse is also becoming a relevant problem both for the sample size and the quality of the data. On the other hand, non probability surveys (NPS) are appealing since they are convenient but suffer from large selection biases. Accuracy of estimates and the inferential framework are still not methodologically defined. Nevertheless, due to large numbers of NPS available, and the problems arising in PS surveys the attention to study methods about how to use NPS and how to improve estimates in PS is growing and the issue more relevant. One natural strand of research is on the integration of PS and NPS. For example, Couper (2013), Miller (2017), Beaumont (2020) talk about exploiting advantages as well as overcoming respective disadvantages of both survey approaches. The most common approach is to adjust for selection bias in NPS estimates using reference PS or census data. A new recent approach proposed is to blend PS surveys with other NPS data sources (see Rao, 2021 for an extensive review).

Integrating both sample types is an ongoing topic of methodological research. We propose a method of combining probability and non-probability samples to improve analytic inference on model parameters. Specifically, we consider a Bayesian framework, where inference is based on a (small) PS and available information from a parallel NPS is provided naturally through the prior.

1.1 Research Aim

Sakshaug et al. (2019) and Wisniowski et al. (2020) proposed a Bayesian data integration approach where inference is based on the PS and the available information from the NPS are supplied through the prior. This framework is studied for the analysis of continuous data. Nevertheless, categorical data, and particularly binary indicators, are of primary interest in surveys, especially in the field of social science, marketing research and psychological analyses. Our original contribution is to develop the abovementioned framework for the analysis of categorical data. In this paper we consider only a binary outcome. The aim is to improve inference about regression coefficients. To evaluate the proposed methodology, we conduct a simulation study assuming different selection scenarios (both missing-at-random MAR and non-missing-at-random MNAR), selection probabilities and sample sizes.

The aim is to compare the performance of some informative priors against a reference non-informative one in terms of mean-squared error (MSE).

The rest of this article is organised as follows. Section 2 introduces the methodological framework. The simulation results are presented and discussed in Section 3. In Section 4 conclusions are drawn.

2 Methodology

We rely on the Bayesian framework which offers a unified approach for integrating multiple data sources of different sizes and quality in a natural way, that is, through the prior structure. We consider a logistic regression to model a binary outcome with covariates. We assume to have a small PS survey and information from a NPS are provided through the prior. Following this approach, biased NP data are incorporated in the estimation process, and posterior estimates are likely to have more bias but possibly less variance than the one obtained using the refence prior.

In the full paper, we also present a real-data analysis study where the potential cost reductions are demonstrated.

2.1 The priors

We propose and test the performance of several informative priors which can be grouped in two categories, distances priors and the power prior.

Distances priors are normally distributed and centred around the maximum likelihood estimates (MLEs) of regression coefficients using only NPS-data, while the scale parameter is a function of the distance between MLEs using the PS and NPS-data only. The smaller is the difference, the more informative is the prior. Hereafter, we refer to the basic Distance prior (Dist) which is representative of this class and its performance is good even in the worst-case scenario. In the full paper, more priors are presented and evaluated. We also consider a mixed version of the distance priors, i.e., only for the intercept the prior is replaced by the reference one.

The Power prior is based on the idea that the prior is proportional to an "initial prior", that we set equal to the reference one, and to the likelihood of the NPS-data. The likelihood is scaled by a parameter which regulates the influence of the NPS data. We set this parameter equal to the p-value resulting from the Hotelling T-test for differences between the two vectors of MLEs from the PS and NPS respectively.

The reference prior is a weakly informative prior proposed by Gelman et al. (2008). It is based on a Student t-distribution with 3 degrees of freedom, centered around 0 and with scale equal to 2.5.

To approximate the posterior distribution, we use the R packages rstan (Stan Development Team, 2021) and rstanarm (Goodrich et al., 2020) based on the No-U-Turn sampler, which is a variant of the Hamiltonian Monte Carlo algorithm.

2.2 The simulation framework

We consider a simulation framework where we take into consideration different models to generate the population, various PS and NPS sizes and several combinations of selection scenarios and selection variables in order draw biased NPS data. Here the results for some selected cases are reported. In our full research study the extensive simulation framework and the complete combination set of the scenarios are considered. We consider the population to be generated from a logistic model with two binary predictors $X_1 \sim Ber(0.5)$ and $X_2 \sim Ber(0.5)$. We assume the coefficients to be $\beta_{MIX} \in (0.5,-1.3,0.9)$ so that the proportion of the outcome variable is almost balanced (0.57). Other results are discussed in the full study.

Under this model, we simulate a population of size N = 1,000,000 from which the PS is drawn with simple random sampling without replacement (srswor) design. We consider different probability sample sizes, from 50 to 1000. We draw a NPS of size 1000 from a simulated NP-panel considering five selection scenarios with different selection probabilities. Here, we present three scenarios and only two selection probabilities which refer to two extreme cases: no bias and high level of bias. The scenarios are the following: (1) p depends on Y (MNAR); (2) p depends on X_1 and X_2 (MAR); and (3) p depends on Y, X_1 and X_2 (MNAR).

3 Results and Discussion

Given the framework presented in the previous section, we repeat the simulation 100 times and to compare the performance of the informative priors against the reference non-informative one, we consider the MSE of the posterior estimates, given by the square of the posterior bias plus the posterior variance.

Figure 1 shows the median MSEs for the selected scenarios and priors. If there is no bias, the reduction in MSEs using informative prior is remarkable, especially when the PS size is small, e.g., lower than 200 cases. When using mixed prior, due to the model formulation, the MSEs for the intercept are always close to the reference prior values.

If the selection mechanism is MAR, using informative priors and controlling for all the selection variables results in an impressive MSE reduction with respect to the reference prior, regardless of the level of selection bias. The power prior performs well when the level of bias is small and especially for small PS sizes (50-100 cases).

In the worst cases, the informative priors perform similarly to the reference prior while for lower levels of selection bias the gain in MSE reduction is evident.

4 Conclusions

The presented framework contributes to the survey data integration literature by proposing a Bayesian data integration approach to improve analytic inference about model parameters by integrating a small PS with a parallel NPS survey.

The current simulation study that, opposite to previous studies, also entails populations with different characteristics and the formulation of various selection mechanisms to account for varying levels of selection bias and selection variables, demonstrates that our approach is robust even in worst-case scenarios. In such a situation, informative priors perform similarly to the reference prior. In the presence of high selection bias, the Distance prior performs better.

In the full research study, we also present a real-case study where potential costs savings are evaluated. We point out that this methodology is not only suitable and profitable for low-budget organisations that can only afford a small PS, but also in the case where a larger PS is available (e.g. greater than 200 units).

In conclusion, in present times when probability samples are suffering from decreasing response rates and high costs and researchers are shifting towards convenient non-probability samples, integrating both samples becomes attractive from both an error and cost perspective.

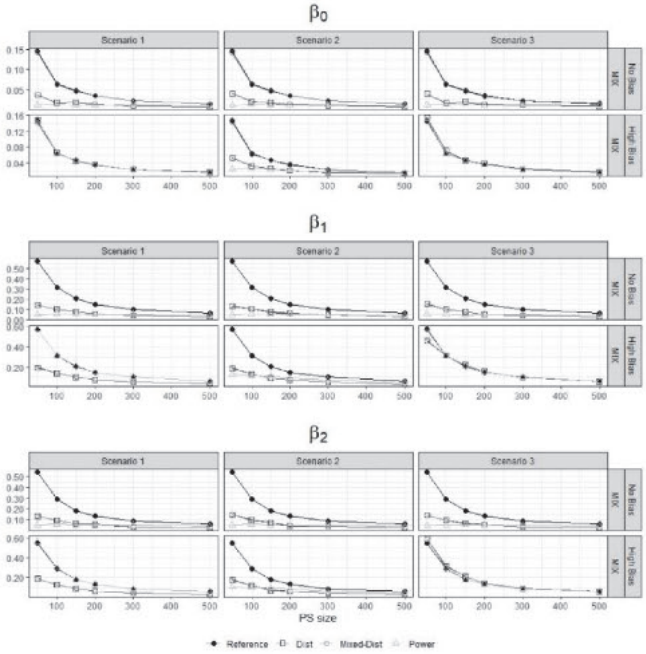


Figure 1: Median MSEs for regression coefficients over 100 iterations in alternative scenarios

References

1. Beaumont, J.-F.: Are probability surveys bound to disappear for the production of official statistics? Survey Methodology 46 (1), 1–29 (2020).
2. Couper, M. P: Is the sky falling? new technology, changing media, and the future of surveys. Survey Research Methods 7 (3), 145–156 (2013).
3. Gelman, A., Jakulin, M. G. Pittau, and Y.-S. Su : A weakly informative default prior distribution for logistic and other regression models. The annals of applied statistics 2 (4), 1360–1383 (2008).
4. Goodrich, B., Gabry J., Ali I., and Brillema S.: rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 (2020).
5. Miller, P. V.: Is There a Future for Surveys? Public Opinion Quarterly 81 (S1), 205–212 (2017).
6. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 28 (2020).
7. Rao, J.: On making valid inferences by integrating data from surveys and other sources. Sankhya B 83 (1), 242–272 (2021).
8. Sakshaug, J. W., A. Wisniowski, D. A. P. Ruiz, and A. G. Blom.: Supplement-ing small probability samples with nonprobability samples: A bayesian approach. Journal of Official Statistics 35 (3), 653–681 (2019).
9. Wisniowski, A., J. W. Sakshaug, D. A. Perez Ruiz, and A. G. Blom.: Integrating probability and nonprobability samples for survey inference. Journal of Survey Statistics and Methodology 8 (1), 120–147 (2020).

Big Data and Official Statistics: some evidences

Big Data e Statistiche Ufficiali: alcune evidenze

Paolo Righi, Natalia Golini, Gianpiero Bianchi

Abstract The paper compares two classes of estimators exploiting a Big Data source. Both classes rely on a probabilistic sampling. Nevertheless, while the first class of estimators uses the Big Data as auxiliary information, the latter uses the probabilistic sample as auxiliary information. We denote this second class as pseudo-calibration estimators, since it applies the calibration to a not random sample. We present an original application of the jackknife method for the variance estimation for the pseudo calibration estimators. Finally, an empirical evaluation on a real survey and Big Data compares several estimators of the two classes with a standard design-based survey estimator.

Abstract *Il lavoro confronta due classi di stimatori che sfruttano una fonte Big Data. Entrambe le classi utilizzano un campione probabilistico. Ma, mentre la prima classe usa i Big Data come informazione ausiliaria, la seconda classe usa il campione probabilistico come informazione ausiliaria. Denotiamo la seconda classe come stimatori di pseudo- calibrazione, poiché si applica la calibrazione a un campione non casuale. Si presenta una applicazione originale del metodo jackknife per la stima della varianza per gli stimatori pseudo-calibrati. Infine, si confrontano empiricamente su dati di indagine e di Big Data reali alcuni stimatori delle due classi con uno stimatore standard design-based.*

Key words: Calibration, Big Data, Official Statistics.

1 Informative context and notation

New sources of data have emerged and are the result of more and more interactions with digital technologies by citizens and business units and the

[1] Paolo Righi, Istat; e-mail: parighi@istat.it

[2] Natalia Golini, Università degli Studi di Torino; e-mail: natalia.golini@unito.it

[3] Gianpiero Bianchi, Istat; e-mail: gianbianchi@istat.it

increasing capability of these technologies to provide digital trails. These sources commonly referred to as Big Data, offer new challenges from the statistical viewpoint in particular generated by a paradigm shift: from designed data for planned statistics to data-oriented or data-driven statistics. Beyond the descriptive statistics, it is necessary to determine under which conditions make valid inference using Big Data. The aim to produce Official Statistics with high-quality standards has stimulated the definition of suitable statistical frameworks (among others: Eurostat, 2018; the American Association for Public Opinion Research (AAPOR) task Force on Big Data, 2015) and quality frameworks (UNECE, 2014).

The paper compares two classes of estimators that use the Big Data source for producing Official Statistics. The two sets of estimators rely on a probabilistic survey but different approach to the inference. The former class concerns a design-based framework, while the latter a model-based framework although an automatic calibration procedure, typical of a model-assisted estimator is carried out. Both classes of estimators apply the calibration techniques and make the estimators appealing to the National Statistical Institutes (NSIs), being these techniques well known by the NSIs. Section 2 introduces the basic notation and the informative context. Section 3 shows the first class of estimators, denoted as data integration estimators (Kim and Tam, 2021). Section 4 illustrates the second class of estimators, referred to this paper as pseudo-calibrated estimator (Righi *et all.*, 2021) or calibration adjustment (Lee and Valliant, 2009). Section 5 shows an empirical evaluation on real survey and Big Data. Finally, Section 6 gives some conclusions.

2 Informative context and notation

Let U be the target population of size N and $U_B \subset U$ be the sub-population of size N_B.

We denote with U_B a Big Data source. In U_B is collected or predicted by a statistical model (with a model error) the random variable $\boldsymbol{y}$. Let us denote with y_k the collected value on the unit $k \in U_B$ and with $\tilde{y}_k$ the predicted value. We use y_k^* notation to indicate either y_k or $\tilde{y}_k$. In case of more than one variable collected or predicted in the Big Data source, we have the $\mathbf{y}_k^* = (y_k^{1*}, \dots, y_k^{h*} \dots, y_k^{H*})'$ vector, being y_k^{h*} the values of hth variable collected or predicted in the Big Data. Furthermore, let $U_{\bar{B}}$ be the set of units without information from the Big Data source being $U_B \cup U_{\bar{B}} = U$ and $U_B \cap U_{\bar{B}} = \emptyset$. Let δ_k indicate the Big Data membership variable, with $\delta_k = 1$ when $k \in U_B$ and $\delta_k = 0$ when $\in U_{\bar{B}}$. Along with U_B, let s be the reference survey sample, in which a probabilistic sample is drawn from U. This is a multi-purpose survey collecting $\mathbf{y}_k$ and a vector $\mathbf{z}_k = (z_k^1, \dots, z_k^q, \dots, z_k^M)'$ of M variables for each $k \in s$. In this setting we assume to know the value of δ_k and we can define $s = s_B \cup s_{\bar{B}}$ with $s_B \cap s_{\bar{B}} = \emptyset$, with $s_B \subset U_B$ and $s_{\bar{B}} \subset U_{\bar{B}}$. Unit nonresponses could affect the reference survey sample. We indicate with r the sample of respondents.

Finally, let $\mathbf{x}_k = (x_k^1, \dots, x_k^p, \dots, x_k^P)'$ be the value vector of the P auxiliary variables known for each $k \in U$. The target parameter is the total

$$Y = \sum_U y_k \,. \tag{2.1}$$

We also consider the total for the domain $U_{(d)} \subset U$ $(d = 1, \dots, D)$,

$$Y_{(d)} = \sum_U y_k \, \lambda_{k(d)}, \tag{2.2}$$

with $\lambda_{k(d)} = 1$ if $k \in U_{(d)}$ and $\lambda_{k(d)} = 0$ otherwise. We indicate with $\boldsymbol{\lambda}_k = (\lambda_{k(1)}, \dots, \lambda_{k(d)}, \dots \lambda_{k(D)})'$ the domain indicator variable vector.

3 Data Integration estimators

We compare two classes of estimators that use in a different way the information coming from the Big Data source. We refer to the first class of estimators as Data Integration (DI) estimators (Kim and Tam, 2021). These estimators define a general tool for making proper use of the Big Data sources for finite population inference by combining the sources with a probabilistic survey.

The DI estimators are design-based, and the Big Data variables are used as auxiliary variables. It is worthy to note that the making design-based inference is an appealing property for the NSIs that usually apply this kind of approach for data production of Official Statistics. The general form of the DI estimators is the Regression DI (RegDI) estimator. By specifying the terms of the RegDI estimators, we can obtain the different DI estimators. Therefore, we focus on the general RegDI estimator. Kim and Tam (2021) give insight on the specific estimators.

The standard survey regression adjusts the survey weights to respect some known totals. In particular the following optimization problem is performed

$$\begin{cases} min \ \sum_s Q(d_k, w_k) \\ \sum_s \mathbf{x}_k \, w_k = \mathbf{X} \end{cases}, \tag{3.1}$$

where d_k is the base sampling weight, w_k is the weight of calibration, $\mathbf{X} = \sum_U \mathbf{x}_k$ is a vector of totals, that we assume as known or estimated by a large and accurate survey (e.g., Dever and Valliant, 2010, 2016) with $\mathbf{x}_k$ known for each $k \in s$.
and,

$$Q(d_k, w_k) = \sum\nolimits_s d_k (\frac{w_k}{d_k} - 1)^2.$$

The RegDI estimator augments the number of auxiliary variables with δ_k and $\delta_k y_k^*$. The estimator is

$$\hat{Y}_{RegDI} = \sum_s y_k w_k \tag{3.2}$$

with $\sum_s \delta_k w_k = N_B$ and $\sum_s \delta_k y_k^* = \sum_{U_B} y_k^*$.

The domain estimator is given by

$$\hat{Y}_{RegDI(d)} = \sum_s \delta_k y_k^* w_k \, \lambda_{k(d)} \tag{3.3}$$

REMARK 4.1: Kim and Tam (2021) deal with the case of unknown δ_k for $k \in s$. We do not analyse this setting in this work.
REMARK 4.2: The $\hat{Y}_{RegDI}$ (3.2)-(3.3) is variable-specific. A more general expression can calibrate the weights on the auxiliary vector $(\mathbf{x}_k, \delta_k, \delta_k \mathbf{y}_k^*)$.

4 Pseudo-calibration estimators

The second class of estimators uses the Big Data source as the large non-probability sample. A critical issue when using a non-probability sample is to dealt with the unknown sample selection mechanism. In particular, since $U_B \subset U$ the no data observations in $U_{\bar{B}}$ (missing data) weakens the representativeness of the Big Data sample with respect to the target population. According to Buelens *et al*. (2014), representativeness is defined as follows: "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective." Meng (2018) underlines that among the different terms generating the selection bias the most important is the correlation between $\boldsymbol{y}$ and δ. When the variable are not correlated, we do not have selection bias. In presence of correlation, there are several approaches for adjusting the selection bias in Big Data. For instance, Kim and Wang (2019), Chen *et all.* (2020), Elliot and Valliant (2017). Here we focus on the estimation process denoted as calibration weighting (Kim, 2022), calibration adjustment (Lee and Valliant, 2009) or pseudo-calibration estimator (Righi *et all.*, 2019).

The estimator calibrates the Big Data distributions on the auxiliary variables related to the target variable so that after this step, these distributions are coherent with the distributions on the target population.

To achieve this objective the calibration process assigns to each unit of the Big Data a final weight acting to satisfy the calibration constraints.

The final weights are obtained by the solution of the following optimization problem:

$$\begin{cases} min \ \sum_{U_B} d(p_k, w_k) \\ \sum_{U_B} \mathbf{x}_k \, w_k = \mathbf{X} \end{cases}, \tag{4.1}$$

where $d(.)$ is a convex function, denoted as a distance function, (Deville and Särndal, 1992), p_k is the initial weight, w_k is the weight of calibration, $\mathbf{X} = \sum_U \mathbf{x}_k$ is a vector of totals, that we assume as known or estimated by a large and accurate survey (e.g., Dever and Valliant, 2010, 2016) with $\mathbf{x}_k$ known for each $k \in U_B$.

We can fix the p_k values in different ways. If we perform a propensity adjustment (Kim, 2022, Elliot and Valliant, 2017, Lee and Valliant, 2009), p_k is the propensity of each unit to be included in the Big Data source. A statistical model estimates this propensity.

In the simplest form $p_k = N/N_B$. When $p_k = p, \forall\, k \in U_B$ and varying p the solution of the optimization problem does not change. So that, with $p_k = 1$ or $p_k = N/N_B$ the calibrated weights, w_k, are the same.

Considering $p_k = 1$ we define a statistical framework where U is a take-all sample (census) with $pr(\delta_k = 1) = 1$ for $\forall\, k \in U$. Nevertheless, U is affected by a kind of unit non-response (alternatively U_B under-covers U). The inclusion probabilities of the respondents, the units in U_B, are adjusted for reducing nonresponse bias by a calibration approach (Little and Rubin, 2007).

4.1 Observe the target variable in the Big Data source

When we collect y_k for $\in U_B$, the pseudo-calibrated estimator is given by

$$\hat{Y}_{PC,B} = \sum_U \delta_k y_k w_k \tag{4.2}$$

being $w_k = 0$ when $\delta_k = 0$. We apply the calibration algorithm (Deville and Särndal, 1992) to solve the optimization problem in (4.1).

The domain estimator is $\lambda_{k(d)}$

$$\hat{Y}_{PC,B(d)} = \sum_U \delta_k y_k w_k \lambda_{k(d)} \tag{4.3}$$

Further discussion on the $\hat{Y}_{PC,B}$ estimator is given in Righi *et all.* (2019).

REMARK 5.1: The proposed estimator has a simple and straight implementation. It leverages well-known and widely used statistical calibration tools in the NSIs.

REMARK 5.2: The proposed estimator is model-based. However, it applies the same process for adjusting unit non-response. The calibration will be generally based on a usual set of auxiliary variables exploited for calibrating or adjusting for non-response the probabilistic sample. The similarity of the process facilitates the consistency of the estimates on the same target population and the estimation computed using either the Big Data or a standard survey based on a probabilistic sample.

4.2 Predict the target variable in the Big Data source

The $\hat{Y}_{PC,B}$ is applicable when we collect y_k in U_B. In some case, we have from a Big Data source a prediction of $\tilde{y}_k$. An example of predicted data is the remote sensing for agricultural statistics (on land use, crop type, crop yield) using the satellite imageries. Another example of predicted data comes from business statistics on the services and functionalities of the enterprise's websites. To count the websites offering specific services (e-commerce, link to social media, job advertisement) we can apply a web-scraping technique by collecting text documents on the website and predicting the presence of the functionalities and services in the website by performing a text analysis and classification by machine learning techniques.

In this case, the estimator (4.2) or (4.3) has to be refined plugging-in the $\tilde{y}_k$ synthetic values for y_k,

$$\hat{Y}^{P}_{PC,B} = \sum_U \delta_k \tilde{y}_k w_k, \tag{4.4}$$

where $\tilde{y}_k$ is null for $\delta_k = 0$. The estimator for the domain $U_{(d)}$ adds the terms $\lambda_{k(d)}$ in the (4.4).

The estimator (4.4) assumes the form of the *projection estimator*. Kim and Rao (2012) define a model-assisted framework of the estimator (4.4) with $\tilde{y}_k = \xi(\mathbf{a}_k \hat{\boldsymbol{\gamma}})$ being ξ a known function, $\mathbf{a}_k$ a vector of auxiliary variable known for $k \in U$ and the $\hat{\boldsymbol{\gamma}}$ vector the estimate of the model parameter vector obtained from a second survey (the reference survey) using the data set $\{(y_k, \mathbf{a}_k): k \in s \subset U\}$ and the survey weights. Kim and Rao (2012) define the conditions to have unbiased estimates. When such conditions are not satisfied, an unbiased estimator is

$$\hat{Y}^{D}_{PC,B} = \hat{Y}^{P}_{PC,B} + \sum_{s \subset U} \delta_k (y_k - \tilde{y}_k) f_k, \tag{4.5}$$

in which the second term of the right-hand side of the (4.5) is the bias correction term, where f_k are the final sampling weights of the reference survey adjusted for the nonresponse in U_B. We assume that y_k and δ_k are observed for $k \in s$. Breidt and Opsomer (2017) denote the (4.5) as a difference estimator and consider the estimator (4.5) based on statistical non-parametric learning techniques such as Kernel methods and regression-tree (Hastie, Tibshirani and Friedman, 2001). In the latter case, the estimation process follows these steps: *i*) the survey-weighted regression tree method is applied to the second survey data $\{(y_k, \mathbf{a}_k): k \in s \subset U\}$ where $\mathbf{a}_k$ represents the auxiliary variable value vector observed in the Big Data source; *ii*) a partition of covariate space in *H strata*, denoted as Endogenous Post Strata (Breidt and Opsomer, 2008), is defined as

$$\tilde{\mathbf{a}}_k = \left[1_{\{\tau_{h-1} < \xi(\mathbf{z}_k) \leq \tau_h\}} \right]_{h=1}^{H}$$

where the $\{\tau_h\}_{h=0}^{H}$ are known breakpoints; *iii*) $\tilde{y}_k$=$\tilde{\mathbf{a}}_k'\widehat{\boldsymbol{B}}$ is computed, where $\widehat{\boldsymbol{B}}' = \left(\frac{N_1}{\widehat{N}_1}, \ldots, \frac{N_h}{\widehat{N}_h}, \ldots, \frac{N_H}{\widehat{N}_H}\right)$ with $\widehat{N}_h = \sum_{k\in h}(1/\pi_k)$. Breidt and Opsomer (2017) introduce in the discussion the use of *random forests* (Breiman, 2001) instead of a tree-based method without a definitive conclusion. Tipton, Opsomer and Moisen (2013) show empirical evaluations of the (4.5) when using the random forest.

5 Variance estimators

DI estimators are design based. For variance estimation, standard linearisation methods (Särndal et all., 1992) or replication methods (Wolter, 2007) for the regression estimator can be applied.

Pseudo-calibration estimator is model-based. We propose to use a replication method. Specifically, we can apply the Delete a Group Jackknife (DAGJK) method (Kott, 2001; Kott, 2006) which is suitable for treating very large sample.

The DAGJK defines G random replication groups drawn from the parent sample, i.e U_B. Then, G estimation processes are carried out using the sample data without the units of one random replication group.

For the difference estimator (4.5) we apply two independent DAGJK variance estimations respectively for the two components of the estimator. Since U_B and s_B are independent samples the variance of the difference estimator is equal to the sum of the variances of its two components.

The estimation process does not consider the re-computation of the $\tilde{y}_k$.

6 Empirical evaluation on European Community survey on ICT usage and e-commerce in enterprises

We implement the above classes of estimator on the real data of the 2018 *European Community Survey on ICT usage and e-commerce in enterprises* (ICT survey) and Internet data scraped from the enterprise websites. The ICT survey's principal aim is to supply users with indicators on Internet connections and usage (website, social media, cloud computing). The target population of the ICT survey is referred to the enterprises with 10 and more persons employed working in the industry and non-financial market services. The frame population is the Italian Business register (Asia) updated to 2 years before the survey reference period. For the 2018 ICT survey, this population has 199,914 units. The design is a stratified sampling. Four classes of the number of persons employed (0-9; 10-19; 20-249; 250 or more), economic activities (24 Nace groups) and geographical breakdown (21 administrative regions at NUTS 2 level) define the strata. The strata including the fourth size class (the enterprises with 250 and more persons employed) are take-all. The number of units in these strata are 3,342. The 2018 sample of respondents is of

22,097 units. The 2018 ICT survey asked the enterprise, among others, if a) *the website gives the possibility to make online ordering or reservation or booking*; *b*) *there are links to social media on the website*. We refer to the two questions as WEBORD and WEBSM variables. The current ICT survey estimator is a calibration estimator. It calibrates on the number of enterprises and persons employed by economic activity, size class and administrative region according to a complex combination of these variables. We uses the Internet data as Big Data sources. We start with the text documents collected by a web-scraping procedure from the enterprises websites. In particular, we have 93,848 ($= N_B$) scraped websites. Note that the total number of websites in target population is unknown. The ICT survey estimate is 134,655.82 enterprises with a relative error of about 1% (Table 6.1). The web-scraping step returns information retrieval for the WEBSM variable. That means we observe the variable with $y_k = 1$ when the website has a link to a social media and with $y_k = 0$ otherwise. Using the text document of each website we predict with a machine learning technique (Random Forest) the WEBORD variable Bianchi *et all.,* 2018; Bianchi and Bruni, 2019). That means we predict the variable with $\tilde{y}_k = 1$ when the website has online ordering or reservation or booking functionalities and with $\tilde{y}_k = 0$ otherwise. The prediction is a value in the interval [0; 1]. Righi *et all.* (2019) give insights on the ICT survey and web-scraping and machine learning procedure.

6.1 Estimators

We compare a simplified version of the ICT survey estimator, denoted as T0, with three different RegDI estimators (T1, T2, T3) and three model-based pseudo calibration estimators (M1, M2, M3). T0 calibrates on the number of enterprises and employed persons for four enterprise size classes (0-9; 10-19; 20-249; +249) and for three macro-regions (aggregation of NUTS 2 regions, Centre, North and South). We have $\mathbf{x}_k = (1, e_k)'$ being e_k the number of employed persons in the unit k. The T1 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)} = \sum_U \mathbf{x}_k\, \lambda_{k(d)}$ and $N_{B(d)} = \sum_U \delta_k\, \lambda_{k(d)}$. The T2 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k\tilde{y}_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$ and $\sum_{U_B} \tilde{y}_k\, \lambda_{k(d)}$. The T3 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k y_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$ and $\sum_{U_B} y_k\lambda_{k(d)}$. The T4 calibration variables are $(\mathbf{x}_k'\boldsymbol{\lambda}_k'; \delta_k\boldsymbol{\lambda}_k'; \delta_k\tilde{y}_k\boldsymbol{\lambda}_k',\ \delta_k y_k\boldsymbol{\lambda}_k')$ and it calibrates on $\mathbf{X}_{(d)}$, $N_{B(d)}$, $\sum_{U_B} \tilde{y}_k\lambda_{k(d)}$ and $\sum_{U_B} y_k\lambda_{k(d)}$. The M1 estimator calibrates the weights on the estimated totals of enterprise and number of employed persons for four size classes and three macro-regions. We use the estimates of T0 of the above totals. The M1 corresponds to the estimator (4.2) for WEBSM and to the estimator (4.4) for WEBORD. The M2 and M3 are difference estimators for WEBORD total. The f_k in M2 is the sampling calibrated weight adjusted by the factor $\sum_r z_k \,/\, \sum_r \delta_k$, with $z_k = 1$ when the enterprise has the website and $z_k = 0$ otherwise. The M3 estimator uses the factor $\sum_r z_k w_k^s \,/\, \sum_r \delta_k w_k^s$ where w_k^s is the calibrated sampling weight of the ICT survey estimator.

6.2 Results

The estimates at the national level (Table 6.1) gives us some preliminary results. The T1 estimator has not effect on the Coefficient of Variation (CV) of the estimates with respect to the T0. The T2 and T3 estimators reduce the CV for the variable involved in the calibration. We have to consider the T4 estimator for improving the standard errors of both WEBORD and WEBSM variables. The M1 estimator gives two main results: *i*) the two estimates are outside the 95% Confidence Interval (CI) of T0. We have to understand if this is a bias evidence or not; *ii*) the CIs of both estimates are very narrow. We apply the difference estimators, M2 and M3, for the WEBORD total estimate. The value is inside the T0 estimator CI. We can assume to have adjusted the bias for the measurement error of the Big Data target variable. Still, the CV increases with respect to M1 but it is smaller than the CV of T0 and the other DI estimators. As far as the bias of WEBSM total is concerned, Table 6.1 shows that M1 is consistent with T3 and T4 estimators that are design-unbiased. We could assume that T0 produces a downward WEBSM estimation.

Table 6.1: Estimates at the national level

Esti-mator	*Variable*	*Total*	*CI(95%) Lower bound*	*CI(95%) Upper bound*	*Estimate not in T0 CI(95%)*	*CV*
T0	WEB	134,655.82	131,831.46	137,480.18		1.07%
	WEBORD	26,451.41	24,473.67	28,429.14		3.81%
	WEBSM	68,221.35	65,157.69	71,285.01		2.29%
M1	WEBORD	30,120.58	29,956.38	30,284.78	**	0.27%
	WEBSM	79,123.88	78,625.52	79,622.24	**	0.31%
M2	WEBORD	26,860.18	25,740.40	28,361.63		2.47%
M3	WEBORD	26,817.45	26,009.59	27,625.31		1.54%
T1	WEBORD	27,150.30	25,092.20	29,208.40		3.87%
	WEBSM	70,520.33	67,388.36	73,652.30		2.27%
T2	WEBORD	27,387.05	25,806.85	28,967.25		2.94%
	WEBSM	70,684.85	67,577.39	73,792.32		2.24%
T3	WEBORD	28,313.23	26,225.65	30,400.82		3.76%
	WEBSM	77,021.37	74,646.39	79,396.34	**	1.57%
T4	WEBORD	27,541.93	25,989.47	29,094.39		2.88%
	WEBSM	77,022.19	74,647.43	79,396.96	**	1.57%

We compare the estimates by size class domains (Figure 6.1) and macro-regions domains (Figure 6.2). The DI estimator CIs always overlap the T0 estimator CI. The length of CIs looks similar even though the DI CIs are a little bit smaller for some domains (size class 0-9 for WEBORD and WEBSM). The pseudo-calibration estimators gives the shortest intervals. For some domains, the WEBSM estimates are significantly different from the T0 (0-9 size class, Center and North macro-regions). The difference estimator adjusts the WEBORD estimates that are within the T0 estimator CI or at least the CIs of the two estimators overlap. Figures 6.1 and 6.2 include the Tb estimator which is a naïve pseudo-calibration estimator defined as $\left(\widehat{N}_W/N_B\right)\sum_{U_B} y_k^*$, where $\widehat{N}_W$ is the survey-based estimate of the number of units with the website. Table 6.3 and 6.4 investigates the sampling errors of the estimators of the cross-classified domains size class by macro-region (12 domains). We

consider two groups of domains: six domains with a sample size between 344 and 547 units (Group 1) and six domains with a sample size between 1,558 and 8,299 sample units (Group 2). Table 6.2 and Table 6.3 show the average CV (%) respectively for WEBORD and WEBSM. The findings point out that the pseudo-calibration estimator are more efficient.

Figure 6.1: Estimator CIs (95%) by size class for WEBORD total (right) and WEBSM total (left).

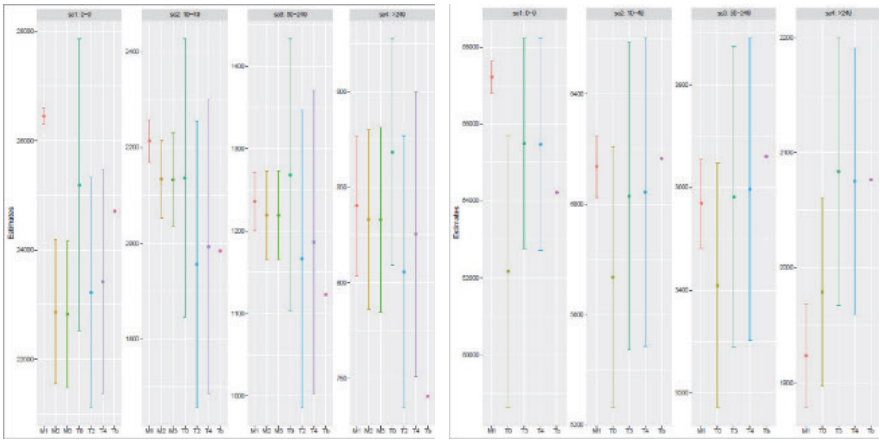


Figure 6.2: Estimator CIs (95%) by macro-regions for WEBORD total (right) and WEBSM total (left).

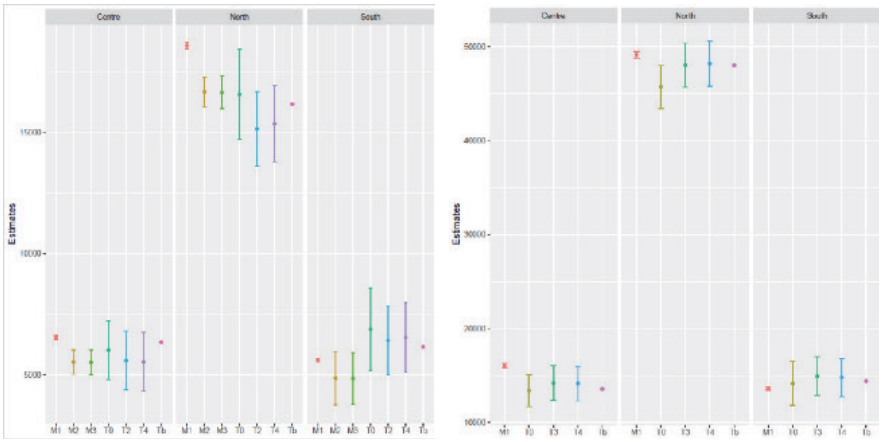


Table 6.2: CV of the estimators for size classes by macro region domain of WEBORD total

Domains	***Average CV(%)***						
	T0	**M2**	**M3**	**T1**	**T2**	**T3**	**T4**
Group 1	12.91	6.18	6.11	13.59	14.36	13.95	14.54
Group 2	7.50	3.73	3.75	7.85	6.26	7.68	6.23

The DI estimators are more efficient than T0 for Group 2 (large domains). Instead, the average of the CV for Group 1 (small domains) is greater than T0.

We explain these findings with the increased number of calibration constraints some units to end up with extreme weights, which will lead to the production of higher variance estimates. This effect is more evident in the small sample size domains.

Table 6.3: CV of the estimators for size classes by macro region domain of WEBSM total

Domains	***Average CV(%)***					
	T0	**M1**	**T1**	**T2**	**T3**	**T4**
Group 1	8.00	3.18	8,47	8.58	9.16	9.26
Group1 2	4.78	1.05	7,02	4.75	3.59	3.59

7 Conclusions

Big Data sources properly used can improve the accuracy of the estimates. In this paper, we introduce, discuss and compare two classes of estimators exploiting the information coming from a Big Data source. The first class takes the Big Data as a source of auxiliary variables into account while a probabilistic survey sample collect the target variables. When the Big Data variables are strictly correlated with survey target variables, the design-based estimates can benefit and the standard errors have a large reduction. The inference approach of these estimators, referred to as Data Integration, is model-assisted. Estimation bias is in the background and depends on the nonresponse issues affecting the survey.

The second class of estimators changes the role of the Big Data. In this case, we directly use the Big Data variables for producing the estimates. Big Data source is a non-probabilistic sample and a probabilistic survey sample focused on the same target population (reference survey) supports the inference. The reference survey needs to: deal with the selection bias of the non-probabilistic survey; adjust the estimates when we have a measurement error on the Big Data target variables. The inference approach of these estimators, referred to this paper as pseudo-calibration estimators, is model-based. Nonetheless, the estimators of this class apply a calibration procedure and the model diagnostic is quite reduced. Variance estimation is computed by means of a replication method. The pseudo-calibration estimators can be biased due to a model failure. On the other hand, the pseudo-calibration estimators increases the real sample size, because they use the non-probabilistic Big Data sample size and the sampling errors can be much smaller than the sampling error of reference survey. The pseudo-calibration estimator sampling errors increase with measurement errors in the Big Data target variables. Both the class of estimators rely on the calibration procedure fostering the practical applicability in the NSIs, in which an automatic estimation process like calibration facilitate the production of the statistics. The experimentation on survey data shows that the sample size make the difference on the sampling errors. The pseudo-calibration estimators based on a large non-probabilistic sample have the best results in terms of precision even though we have to evaluate carefully the risk of bias.

References

1. AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. Public Opinion Quarterly, 79, 839–880.
2. Breidt. F.J., Opsomer. J D. (2008). Endogenous poststratification in surveys: Classifying with a sample-fitted model. Annals of Statistics, 36, 403–427.
3. Breidt. F.J., Opsomer. J.D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. Statistical Science, 32, 190–205.
4. Breiman L. (2001). Random Forests. Machine Learning, 45, 5-32.
5. Bianchi G., Bruni R., Scalfati F. (2018). Identifying e-Commerce in Enterprises by means of Text Mining and Classification algorithms. Mathematical Problems in Engineering, Vol. 2018, n. 7231920.
6. Bianchi G., Bruni R. (2019). Website Categorization: a Formal Approach and Robustness Analysis in the case of E-commerce Detection. Expert Systems with Applications.
7. Buelens B., Daas P., Burger J., Puts M., van den Brakel J. (2014). Selectivity of Big data. Discussion Paper nr. 11. Statistics Netherlands.
8. Elliott M., Valliant. R. (2017). Inference for nonprobability samples. Statistical Science, 32, 249–264.
9. Chang C.-C., Lin C.-J. (2001). Training v-support vector classifiers: Theory and algorithms. Neural Computation, 13(9), 2119-2147.
10. Chen Y., Li P., Wu C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. Journal of the American Statistical Association, 2011-2021,
11. Dever. J., Valliant. R. (2010). A comparison of variance estimators for post-stratification to estimated control totals. Survey Methodology, 36, 45–56.
12. Dever. J., Valliant. R. (2016). GREG estimation with undercoverage and estimated controls. Journal of Survey Statistics and Methodology, 4, 289–318.
13. Deville, J. C., Särndal, C. E., (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 367-382.
14. EUROSTAT (2018). Report describing the quality aspects of Big Data for Official Statistics. Work Package 8 Quality Deliverable 8.2, ESSnet Big Data.
15. Hastie T., Tibshirani R., Friedman J. (2001). The Elements of Statistical Learning: Data Mining. Inference and Prediction. Springer. New York.
16. Kim J.K. (2022). A gentle introduction to data integration in survey sampling. The Survey Statistician, 19–29.
17. Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. International Statistical Review, 382–401.
18. Kim J.K., Rao J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, 85–100.
19. Kim J.K., Wang Z. (2019). Sampling techniques for big data analysis in finite population inference. International Statistical Review, 177-191.
20. Kott, P. (2006). Delete-a-group variance estimation for the general regression estimator under poisson sampling, Journal Official Statistics, 759–767.
21. Kott, P. (2001). Delete-a-group jackknife. Journal Official Statistics, 521–526.
22. Little. R.J.A., Rubin. D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.
23. Meng X-L. (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. The Annals of Applied Statistics, 12, 685–726.
24. Righi P., Bianchi G., Nurra A., Rinaldi M. (2019). Integration Of Survey Data And Big Data For Finite Population Inference In Official Statistics: Statistical Challenges and Practical Applications. Statistica & Applicazioni, 135-158
25. Särndal C.-E., Swensson B., Wretman J. (1992). Model Assisted Survey Sampling. Springer. New York.
26. Tipton J., Opsomer J., Moisen G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. Remote Sensing of Environment, 139, 130–137.
27. UNECE (2014). A Suggested Framework for the Quality of Big Data. Deliverables of the UNECE Big Data Quality Task Team, December 2014.
28. Wolter, K. (2007) Introduction to Variance Estimation. Springer, London.

The analysis of students performance and behaviour based on large databases

Students enrolled in STEM disciplines in Italy: patterns of retention, dropout and switch

Studenti iscritti nelle discipline STEM in Italia: i percorsi di chi prosegue gli studi, abbandona o cambia corso

Valentina Tocchioni, Carla Galluccio, Maria Francesca Morabito, Alessandra Petrucci

Abstract Ongoing technological change has led to a steadily growing demand for Science, Technology, Engineering and Mathematics (STEM) graduates worldwide. But not only do STEM disciplines have a low attractiveness in some contexts, such as in the US and Italy; it is also a matter of persistence of pursuing STEM studies, affected by high rates of dropout and course switches in several countries. Using administrative microdata from the Italian Ministry for Universities and Research and selecting students enrolled in a STEM discipline between 2010 and 2015, our objective is twofold: understanding which distinct patterns characterise students towards retention, dropout, or switch; investigating to what extent each individual and contextual characteristic predict students' outcomes. Identifying at-risk STEM students to dropout/switch is an essential and challenging issue for the delivery of university interventions aiming to reduce failure and dropout rates.

Abstract *Il cambiamento tecnologico ha portato a una domanda sempre crescente di laureati in scienze, tecnologia, ingegneria e matematica (STEM) in tutto il mondo. Nonostante ciò, le discipline STEM sono scarsamente attrattive in vari contesti, come negli Stati Uniti e in Italia; inoltre, sono contraddistinte da alti tassi di abbandono e cambi di corso. Utilizzando i microdati amministrativi del Ministero dell'Università e della Ricerca, abbiamo selezionato gli studenti iscritti a una disciplina STEM tra il 2010 e il 2015 con l'intento di perseguire due obiettivi: individuare quali pattern caratterizzano gli studenti che proseguono gli studi, li abbandonano o cambiano corso; comprendere in che misura ciascuna caratteristica individuale e di contesto*

[1] Valentina Tocchioni, University of Florence; email: valentina.tocchioni@unifi.it
Carla Galluccio, University of Florence; email: carla.galluccio@unifi.it
Maria Francesca Morabito, University of Florence; email: mariafrancesca.morabito@unifi.it
Alessandra Petrucci, University of Florence; email: alessandra.petrucci@unifi.it

predice i tre percorsi degli studenti. Riuscire a identificare gli studenti STEM a rischio rappresenta un elemento cruciale per la definizione di interventi volti a ridurre i tassi di fallimento e abbandono.

Key words: university students, STEM, graduation, dropout, course switch, Italy

Introduction

Ongoing technological change has led to a steadily growing demand for graduates from Science, Technology, Engineering and Mathematics (STEM) worldwide, due to their prominence in the development, productivity and growth of contemporary economies. Skills in STEM disciplines are thus becoming an increasingly important part of basic literacy in today's economy. Several actions have been put in place to increase the attractiveness of degree programs in science and technology, and thus satisfy the growing demand for future scientists for engineers: as stated by the European Schoolnet, "to keep Europe growing, we will need one million additional researchers by 2020". Also, in the US, increasing the number of undergraduate STEM majors has recently emerged as a national priority (Kuenzi, 2008); thus, in the last years, they have concentrated on expanding existing STEM education programs, but also on implementing new programs to increase the number of students entering STEM disciplines (Thompson and Bolin, 2011).

Despite it, the STEM disciplines have different attractiveness in diverse contexts. In this respect, Ireland, with 37 people aged 20-29 over 1,000 of the same age class in 2019, is at the forefront of the number of highly-talented graduates in these fields, followed by France, the UK and Germany, with around 24-28 people aged 20-29 graduated in STEM fields over 1,000 of the same age class in 2019 (Eurostat, 2022). In Italy, the interest in STEM majors is not very pronounced: as a consequence, the number of people who graduated in these fields is below the mean number of graduates in the European Union, equal to 21.6 people aged 20-29 who graduated in STEM fields over 1,000 of the same age class in 2019 (Eurostat, 2022). Despite it, the annual rate of graduates slightly increased over the last years, passing from 13.8‰ in 2013 to 16.4‰ of people aged 20-29 in 2019 (Eurostat, 2022).

Not only do STEM disciplines have an issue of attraction, but also of retention. Indeed, many students who decide to enrol in a STEM discipline then change their minds, thus switching to another course or, even worse, dropping out of university studies. The first few years of enrolment are crucial in this respect, and this is a particular concern for the STEM disciplines, which are the most affected by both practices with respect to other disciplines such as Business or Education fields (Chen et al., 2018; Thompson and Bolin, 2011). Moreover, among those who switch to another course, most students switch to a non-STEM field (Isphording and Qendrai, 2019). Consequently, despite the increase in enrolment in STEM fields in some countries such as the US, this rise has not been followed by a higher number of graduates. Against these premises, the high number of undergraduate students leaving STEM courses represents an issue of societal concern (Seymour and Hewitt, 2000).

Various individual and contextual characteristics may influence this unsuccessful academic outcome. At the individual level, students' prior math achievement and quantitative skills have been identified as the most important predictor of STEM study success (De Winter and Dodou, 2011). As for gender, an association has been identified in previous studies between students' gender and dropout, with female students more likely to dropout than males (Thompson and Bolin, 2011; Isphording and Qendrai, 2019); conversely, a lack of association seems to occur between student's gender and course switch (Thompson and Bolin, 2011). On the other hand, no association has been identified between ethnicity and students' dropping out or switching.

At the contextual level, previous studies have identified a negative association between high school ranking and students' dropout or switch, with students from schools with a higher ranking less prone to dropout or switch the course (Thompson and Bolin, 2011). The social context and the peer effect also have a role, with female students retaining their STEM preferences when other females in their classroom do so (Raabe et al., 2019).

1.1 Objective of the work

Our work aims at investigating the academic outcomes of university students who decided to enrol in a STEM course for the first time in Italy. More specifically, we are interested in understanding which micro-, meso- and macro-level characteristics play a role in predicting students' graduation, dropout or course switch among those enrolled in a STEM course, such as gender, the number of credits attained during the first year of enrolment, and the type of high school. Moreover, we intend to verify if there is a relationship between the athenaeum of enrolment and students' performances in terms of graduation, dropout or course switch. In doing so, we rely on some characteristics of the athenaeum where the student is enrolled, such as admission requirements and rates, service offered, and so on.

Identifying at-risk STEM students is an essential and challenging issue at the individual, university, and societal levels. At the individual level, a successful academic career is undoubtedly beneficial for the students themselves. At the university level, from 2014 funds and economic incentives for universities are related to their success in providing degrees within the prescribed time (Viesti, 2018). Finally, society has both direct and indirect interests in university students' success, given that public universities in Italy receive funds from the government (deriving from taxes) and that the prosperity of a country is strongly affected by its citizens' education and skills, as well as its quality of human resources (Becker, 2009; Schultz, 1971). In this respect, understanding which factors could predict the early failure of undergraduate students in those disciplines, and creating a series of student performance indicators could provide opportunities for the timely delivery of educational interventions, aiming to reduce the high failure and dropout rates.

In the light of these premises, we wonder if there are factors attributed to STEM students who graduated that might serve as predictors or indicators of successful

navigation in STEM majors. If factors can be identified, they may be used as tools by high school counsellors and college advisors in the recruitment and, possibly more importantly, the retention of future STEM students. Conversely, we intend to verify if there are characteristics among STEM students who dropped out or switched the course that might act as warnings for identifying students with weaker paths at the beginning of their academic careers. University advisors may use these signals of poor performance to address specific educational interventions for those students and thus mitigate the significant dropout rates observed in undergraduate STEM education.

2 Data and sample selection

In the present contribution, we used data coming from the National Archive of Students and Graduates (i.e., *Archivio Nazionale degli Studenti e dei Laureati*, ANS), an administrative database that was created[1] with the aim of recording and monitoring the careers of all university students enrolled in a degree program at an Italian university. The database is provided by the Ministry of Education, University, and Research (MIUR) with the involvement of Italian universities.

In the following, we describe the features of the database used and the criteria we considered to select the sample.

2.1 Data description

The ANS database concerns university students enrolled in a degree program at an Italian university since 2010. More specifically, it contains individual longitudinal data, with information about students' demographic characteristics (i.e., gender, region of residence, citizenship) and information on both high school careers (i.e., type of high school attended, final mark) and university careers (i.e., degree program chosen, number of formative university credits achieved per year, type of degree, year in which they get the degree and final grades).

2.2 Sample selection criteria

In this contribution, we decided to focus on the cohorts of students enrolled from the academic years 2010–2011 through 2015–2016.

[1] This database has been realised thanks to the Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide" (PI: Massimo Attanasio).

Moreover, considering the aim of our study, to obtain a set of students consistent with our research goals, we selected students who chose a STEM bachelor's degree at their first university enrolment, and removed from our sample students who did not pay university fees in the year of enrolment. Regarding the definition of a STEM bachelor's degree, according to the ISCED classification of fields of education, we kept only students enrolled in the following three categories: Natural sciences, mathematics, and statistics (ISCED code 5); Information and Communication Technologies (ISCED code 6); Engineering, manufacturing, and construction (ISCED code 7).

The final sample comprises 364,608 students (36.2% females, 63.8% males). Overall, 46,629 students dropped out during the second year or after (until the fifth year after their academic enrolment), whereas 82,532 students switched the course.

Among all the students, about 41.4% came from Northern Italy, 37.6% from South Italy and the Islands, and 21.0% from Central Italy. In relation to high school, the majority of the students attended scientific high school (55.5%), followed by technical institute (18.1%), professional institute and classical high school (both 8.6%), other institutes (5.7%), and foreign language high school (2.1%). There were missing values for 1.4% of students. Finally, regarding the field of study chosen (ISCED code) in the first year, 35.2% enrolled in Natural sciences, mathematics, and statistics courses, 6.6% in Information and Communication Technologies courses, whereas 58.2% of students were enrolled in Engineering, manufacturing, and construction courses. Most of the students were Italian (96.6%).

The number of students per cohort included in the analysis with some descriptive statistics is shown in Table 1.

Table 1: *Demographic information of the selected sample per cohort. Absolute and column percentage values.*

Variables	***Cohort 2010***	***Cohort 2011***	***Cohort 2012***	***Cohort 2013***	***Cohort 2014***	***Cohort 2015***
Number of students	62,258	60,640	59,688	57,980	59,989	64,053
Gender						
Female	23,332	22,591	21,972	20,399	21,046	22,603
	(37.5%)	*(37.3%)*	*(36.8%)*	*(35.2%)*	*(35.1%)*	*(35.3%)*
Male	38,926	38,049	37,716	37,581	38,943	41,450
	(62.5%)	*(62.7%)*	*(63.2%)*	*(64.8%)*	*(64.9%)*	*(64.7%)*
Citizenship						
Italian	60,310	58,568	57,652	55,971	57,992	61,906
	(96.9%)	*(96.6%)*	*(96.6%)*	*(95.3%)*	*(96.5%)*	*(96.6%)*
Foreign	1,948	2,072	2,036	2,009	2,826	2,147
	(3.1%)	*(3.4%)*	*(3.4%)*	*(4.7%)*	*(3.5%)*	*(3.4%)*
Region of residence						
North-West	14,583	14,086	13,981	13,636	14,205	15,165
	(23.4%)	*(23.2%)*	*(23.4%)*	*(23.5%)*	*(23.7%)*	*(23.7%)*
North-East						

	10,962	10,793	10,392	10,500	11,014	11,629
Centre	*(17.6%)*	*(17.8%)*	*(17.4%)*	*(18.1%)*	*(18.3%)*	*(18.2%)*
	12,746	12,975	12,682	12,015	12,685	13,429
South	*(20.5%)*	*(21.4%)*	*(21.2%)*	*(20.7%)*	*(21.1%)*	*(20.9%)*
	17,659	16,993	16,557	16,010	16,043	17,062
	(28.4%)	*(28%)*	*(27.8%)*	*(27.6%)*	*(26.8%)*	*(26.6%)*
Island	6,308		6,076	5,819	6,042	6,768
	(10.1%)	5,793	*(10.2%)*	*(10.1%)*	*(10.1%)*	*(10.6%)*
		(9.6%)				
High School						
Classical	5,527	5,463	5,318	4,779	5,138	5,117
	(8.9%)	*(9%)*	*(9%)*	*(8.2%)*	*(8.6%)*	*(8%)*
Scientific	33,659	33,710	33,622	31,936	33,183	36,230
	(54%)	*(55.6%)*	*(56.3%)*	*(55.1%)*	*(55.3%)*	*(56.6%)*
Foreign Language	1,233	1,182	1,148	1,077	1,207	1,677
	(2%)	*(1.9%)*	*(1.9%)*	*(1.9%)*	*(2%)*	*(2.6%)*
Technical Institute	11,036	10,325	10,319	10,897	11,459	12,016
	(17.7%)	*(17%)*	*(17.3%)*	*(18.8%)*	*(19.1%)*	*(18.8%)*
Professional	5,893	5,332	4,877	5,136	5,289	5,001
Institute	*(9.5%)*	*(8.8%)*	*(8.2%)*	*(8.8%)*	*(8.8%)*	*(7.8%)*
Other	3,922	3,607	3,420	3,353	3,185	3,384
	(6.3%)	*(6%)*	*(5.7%)*	*(5.8%)*	*(5.3%)*	*(5.3%)*
Missing	988	1,021	984	802	528	628
	(1.6%)	*(1.7%)*	*(1.6%)*	*(1.4%)*	*(0.9%)*	*(0.9%)*
ISCED classification						
Natural sciences, mathematics, and statistics	23,123 *(37.1%)*	21,556 *(35.5%)*	21,114 *(35.4%)*	19,804 *(34.2%)*	20,083 *(33.5%)*	22,594 *(35.3%)*
Information and Communication Technologies	3,297 *(5.3%)*	3,504 (5.8%)	3,740 *(6.3%)*	4,181 *(7.2%)*	4,513 *(7.5%)*	4,912 *(7.7%)*
Engineering, manufacturing, and construction	35,838 *(57.6%)*	35,580 *(58.7%)*	34,834 *(58.3%)*	33,995 *(58.6%)*	35,393 *(59%)*	36,547 *(57%)*

3 Methodology

We run a multinomial logistic model to investigate the role of micro-, meso- and macro-level characteristics in predicting individual academic outcomes.

The response variable *Outcome* has four categories, distinguishing between students who graduated (*Graduated on time*), who dropped out of the course (*Dropped out*), who have changed the course (*Course switch*), and those who were still enrolled

at the same course (*Still enrolled*). The students' outcome is observed four years after enrolment.

As micro-level covariates, we include gender, citizenship (Italian or foreign), the high school final mark (in classes of width 10), the number of credits attained during the first year of enrolment (in classes: 0-24; 25-40; 41-56; 57-72), if the student resides outside the region where the athenaeum is located. Furthermore, regarding meso-variables, we look at the type of high school (scientific high school, technical institute, professional institute, classical high school, other institutes, and foreign language high school), and the ISCED code of the course (5, 6, or 7). Finally, among macro-characteristics, we control for the macro area of athenaeum (North-West, North-East, Centre, South, Islands).

4 Results

Figure 1 shows the number of students enrolled in a STEM discipline who graduated, dropped out or switched each academic year by cohort. As for graduates, the highest number of graduates for the cohorts 2010-2012 is during the fifth year of enrolment, whereas for the cohorts 2013-2014 the highest number of graduates is in the fourth year of enrolment (namely, within the legal duration of the course). Instead, most students dropped out or switched the course during the second year of the course and, to a lesser extent, during the third year; starting from the fourth year of enrolment, drops out or switches are considerably less. Moreover, the patterns of those students who dropped out or switched the course are very similar, with a replicating pattern over the different cohorts, too; finally, the number of students who switched is always higher than those who dropped out.

Figure 1: *Students graduated, dropped out and switched by cohort of enrolment. 2010-2014.*

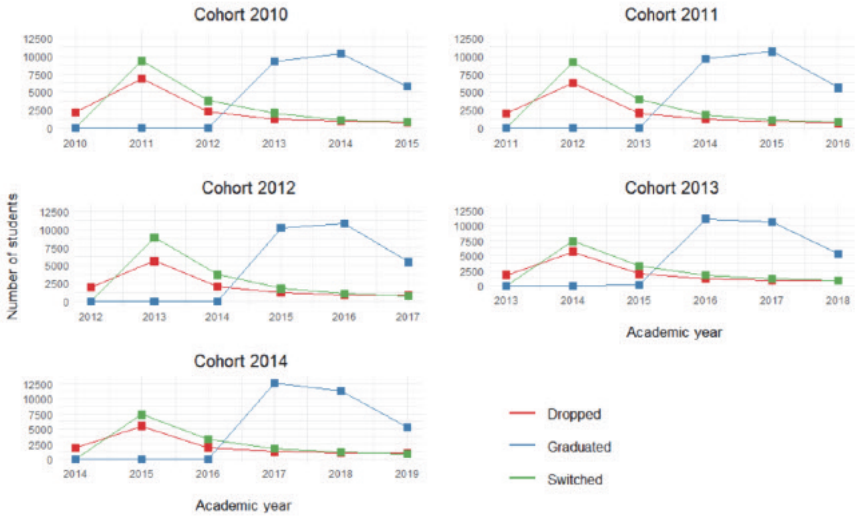


Table 2 shows the relative risk ratios for the estimation of the multinomial logistic regression. As for individual characteristics, the relative risk of dropping out over graduation is lower for women than for men, whereas the opposite is true for course switches. Italian students have a relative risk of dropping out, course switching and not graduating within 4 years over graduation on time lower than their foreign counterparts. Moreover, a high final mark at high school implies a lower relative risk of dropping out, course switching and not being graduated within 4 years over graduation with respect to those students with a low grade. The same holds for the number of credits attained during the first year of enrolment, which is the variable showing the lowest relative risk ratios, thus seeming to be the most relevant in explaining students' academic successful or unsuccessful outcomes. Finally, students who live outside the region where the athenaeum is located have a slightly smaller relative risk of dropping out with respect to those who live in the same region, whereas their relative risk of course switching or not graduating within 4 years over graduation is higher.

Looking at the meso and macro characteristics, students enrolled in ICT courses or in Engineering, manufacturing, and construction have a lower relative risk of course switch over graduation on time with respect to students enrolled in Natural sciences, mathematics, and statistics. Conversely, they all have a higher relative risk of not being graduated within 4 years over graduation on time with respect to students enrolled in Natural sciences, mathematics, and statistics. Also, only students enrolled in Engineering, manufacturing, and construction have a lower relative risk of dropout over graduation on time with respect to students enrolled in Natural sciences, mathematics, and statistics. Furthermore, students enrolled in an athenaeum outside the South of Italy have a lower relative risk of dropping out, course switching and not graduating within 4 years over graduation on time with respect to students enrolled in an athenaeum in the South of Italy. As regards the type of high school, students who attended a scientific lyceum have a lower relative risk of course switching and not graduating within 4 years over graduation on time compared to students who attended a classical lyceum; other students, such as those who attended a linguistic lyceum have a higher relative risk of dropping out, course switch and not being graduated within 4 years over graduation on time with respect to students who attended a classical lyceum; finally, students who attended a technical or a professional institute have a higher relative risk of dropping out and not being graduated within 4 years, but a lower relative risk of course switch over graduation on time with respect to students who attended a classical lyceum.

Table 2: *Model results of the multinomial logistic regression model. Relative risk ratios and standard errors (in brackets).*

	Ref: Graduated in 4 years		
Variable	**Dropped**	**Switched**	**Enrolled not graduated after 4 years**
Constant	4.540***	46.998***	70.927***
	(0.303)	(2.333)	(3.544)

Female	0.862*** (0.017)	1.050*** (0.013)	0.994 (0.012)
Italian citizenship	0.629*** (0.028)	0.667*** (0.024)	0.546*** (0.020)
High school final mark (ref: 60 – 69)			
70 – 79	0.514*** (0.011)	0.620*** (0. 011)	0.650*** (0.012)
80 – 89	0.275*** (0.007)	0.386*** (0. 007)	0.436*** (0.008)
90 – 99	0.166*** (0.005)	0.251*** (0. 005)	0.313*** (0.006)
100 and 100 cum laude	0.091*** (0.004)	0.167*** (0. 004)	0.215*** (0.005)
Missing	0.283*** (0.018)	0.189*** (0. 010)	0.276*** (0.014)
Credits (ref: 0 – 24)			
25 – 40	0.016*** (0.001)	0.115*** (0. 003)	0.169*** (0.004)
41 – 56	0.002*** (0.0002)	0.052*** (0.001)	0.031*** (0.001)
57 – 72	0.001*** (0.0001)	0.027*** (0. 001)	0.005*** (0.0001)
Missing	5.399*** (0.170)	0.578*** (0. 015)	0.090*** (0.002)
Student outside region	0.948** (0.024)	1.333*** (0.021)	1.259*** (0.020)
ISCED code (ref: Natural sciences, mathematics and statistics)			
Information and Communication Technologies	0.977 (0.032)	0.542*** (0.014)	1.373*** (0.034)
Engineering, manufacturing and construction	0.698*** (0.013)	0.796*** (0.009)	1.194*** (0.015)
High School (ref: Classical lyceum)			
Scientific lyceum	0.946 (0.033)	0.661*** (0.012)	0.874*** (0.017)
Technical Institute	4.506*** (0.169)	0.924*** (0.021)	1.448*** (0.034)
Foreign Language lyceum	2.520*** (0.161)	1.344*** (0.054)	1.235*** (0.052)
Professional Institute	5.168*** (0.214)	0.854*** (0.023)	1.555*** (0.043)
Other	4.501***	1.508***	1.607***

	(0.208)	(0.045)	(0.050)
Missing	3.010***	0.502***	1.905***
	(0.208)	(0.031)	(0.098)
Geographical area of University (ref: South)			
Island	0.873***	0.841***	0.762***
	(0.028)	(0.019)	(0.017)
Centre	0.820***	0.854***	0.706***
	(0.020)	(0.014)	(0.011)
North-East	0.430***	0.429***	0.349***
	(0.011)	(0.008)	(0.006)
North-West	0.556***	0.634***	0.426***
	(0.013)	(0.010)	(0.007)

Note: *p<.10; **p<.05; ***p<.01

5 Preliminary conclusions and further steps

In this paper we investigate the determinants of the academic outcomes of university students who decided to enrol in a STEM course for the first time in Italy. Our preliminary analyses show that several micro, meso and macro characteristics play a role in predicting students' graduation, dropout or course switch among those enrolled in a STEM course, such as the high school final mark, the number of credits attained during the first year of enrolment and the type of high school.

In order to fully understand the relationship between micro, meso and macro characteristics and students' academic outcomes, we will estimate separate models by ISCED code and test the inclusion of interaction terms in the model. Moreover, we will estimate a multilevel version of the model, with students nested within the athenaeum where they are enrolled, to assess the overall performance of the athenaeum in terms of graduates and dropout.

Further research is required to deepen STEM students' academic outcomes in Italian public universities and to investigate if there exists a relationship between the athenaeum of enrolment and students' performances, and if so, which features of the athenaeum play a major role.

References

1. Becker, G. S. (2009). *Human capital: A theoretical and empirical analysis, with special reference to education*. Chicago: University of Chicago press.
2. Chen, Y., Johri, A., Rangwala, H. (2018). Running out of STEM: A comparative study across STEM majors of college students At-Risk of dropping out early. In LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia. ACM, New York.
3. De Winter, J.C.F., Dodou, D. (2011). Predicting academic performances in engineering using high school exam scores. *International Journal of Engineering Education* 27(6): 1343-1351

4. EUROSTAT (2022). Graduates in tertiary education, in science, math., computing, engineering, manufacturing, construction - per 1000 of population aged 20-29. Dataset. (Accessed on 31 March 2022).
5. Isphording, I., Qendrai, P. (2019). Gender differences in student dropout in STEM. IZA Research Reports, 87.
6. Kuenzi, J. J. (2008). Science, technology, engineering, and mathematics (STEM) education: Background, federal policy, and legislative action. Congressional Research Service Reports 35.
7. Raabe, I. J., Boda, Z., Stadtfeld, C. (2019). The social pipeline: How friend influence and peer exposure widen the STEM gender gap. *Sociology of Education* 92(2): 105-123.
8. Schultz, T. W. (1971). *Investment in human capital. The role of education and of research*. New York: The Free Press.
9. Seymour, E, Hewitt, N.M. (2000). *Talking about leaving: Why undergraduates leave the sciences*. Boulder: Westview Press.
10. Thompson R., Bolin, G. (2011). Indicators of success in STEM Majors: A cohort study. Journal of College Admission.
11. Viesti, G. (2018). *La laurea negata: le politiche contro l'istruzione universitaria*. Bari: Gius. Laterza & Figli Spa.

The routes of Southern Italy university students: an explorative analysis

I percorsi di mobilità degli studenti meridionali in Italia: un'analisi esplorativa

Gabriele Ruiu[1] and Vincenzo Giuseppe Genova[2]

Abstract The neoclassical migration approach postulates that different conditions in labour markets among territories are the driving forces behind migration. On the other hand, exponents of the new economics of migration argue that the decision to move is not made at the individual level. Considering migration as the result of a decision taken within a social network helps to explain the so-called chain migration. In this paper, we pay attention to the migratory chain of university students. This work introduces a statistical technique to "classify" migratory chains of students living in Sicily, Sardinia or Apulia, and enrolled in some centre-north regions from 2008 to 2017.

Abstract *In accordo agli economisti neoclassici le differenti condizioni del mercato del lavoro sono alla base dei movimenti migratori. D'altro canto, gli esponenti della new economics of migration ritengono che la decisione di migrare non sia solo un processo individuale. Considerare le migrazioni come delle scelte effettuate all'interno di un network aiuta dunque a capire l'esistenza delle catene migratotie. Questo lavoro si concentra sulle catene migratorie studentesche. In particolare, si propone una tecnica statistica per individuare le catenr generate nei percorsi di mobilità degli studenti universitari di Sicilia, Sardegna e Puglia iscritti in un ateneo del Centro-Nord dal 2008 al 2017.*

Key words: Student mobility, chain migration, cluster analysis, university students

[1]Gabriele Ruiu, Department of Economics and Business, University of Sassari; email: gruiu@uniss.it

[2]Vincenzo Giuseppe Genova, Department of Economic, Business, and Statistics, University of Palermo; email: vincenzogiuseppe.genova@unipa.it

This work has been realized in the ambit of the PRIN 2017HBTK5P: From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide; P.I. Massimo Attanasio.

1 Introduction

In Italy, there have been significant changes in terms of student flows and mobility in the last twenty years. Student enrolment has decreased from 2008 to 2015 with consistent recovery in the last five to six years. Student mobility is unidirectional, from the South to the Centre and North of the country. It has been continuously increasing since 2008, with a slight reduction in 2017. Attanasio and Enea [2] analysed the mobility flows of university students in Italy, highlighting that movements are unidirectional from the South and Islands to the Centre-North. They report that 10.4% (15.9%) of students from the South (Islands) were enrolled in a Northern university and that 11% (8.8%) of them moved towards the Centre in 2014. This behaviour also is also observed in the transition from the bachelor's to the master's degree and Ph.D. programs (Ruiu et al. [12], Tocchioni and Petrucci [14], Genova et al. [8])

Among Italian scholars, Boscaino et al. [4] argued that one of the determinants of student mobility from the South to the North is related to better job-market opportunities in Centre-North. Furthermore, Santelli et al. [13] showed how the southern regions are affected by the increasing rate of students moving to Centre-North—especially from Sicily. Furthermore, D'Agostino et al. [6] and Impicciatore and Tosi [9] note how this mobility is also affected by contextual factors such as students' social class and family background.

These mechanisms involve public information and contextual factors that cannot explain specific South-to-North mobility patterns. Since our primary goal is to study the presence of preferential mobility patterns, we invoke the paradigm of chain migration in demography. Specifically, we look at "student chain migration" within the broader context of chain migration in demography.

The literature on students' mobility has advanced the idea that mechanisms similar to chain migration may explain the choice of a university. However, this hypothesis has been tested only using qualitative surveys (Brooks and Waters [5], Pérez and McDonough [11]). This paper investigates the presence of chain migration processes in students' mobility through a quantitative method.

We focus exclusively on the flows from three Southern regions (Sardinia, Sicily, and Apulia) towards Central-Northern Italy. Our main research question can be formulated as follows: what is the role of the chain process in determining the unidirectional patterns, as it has been highlighted in the literature.

The paper is organised as follows: in section 2, we illustrate the data and statistical methods that were employed to analyse the chain migration; in section 3, we present the results and offer some conclusions.

2 Data and Method

The data used in this work were obtained from MOBYSU.IT [9], which contains longitudinal information on the careers of university students from 2008 to 2017.

The analysis is focused on students who received a high school diploma in Sicily, Sardinia and Apulia, and enrolled in a degree course in Piedmont, Lombardy, Emilia-Romagna, Veneto, Tuscany or Latium from 2008 to 2017. The analysis excluded students who enrolled in an online degree course or healthcare area. The former were excluded because there is no real mobility, while the latter were excluded because the rules strongly condition their mobility for admission to those particular degree courses. Given the small size of each cohort, we decided to aggregate the cohorts into two five-year groups, those enrolled from 2008 to 2012 and from 2013 to 2017.

The methodological approach employed in this work consists of two parts. In the first part, we propose a technique to determine areas in which communities of students can directly communicate and eventually trigger a mechanism of chain migration (AreaOri). In the second part, we perform the complete linkage cluster analysis (CLINK) on the residuals of the origin-destination matrixes—under the hypothesis of independence—where the origins are the AreaOri and the destinations are some Central-Northern regions (RegDest) mentioned in section 1 for the two five-year periods under analysis.

The assumption is that student communities in one AreaOri directly communicate within the AreaOri and not directly communicate with other AreaOris. The AreaOris are formed by several municipalities situated around a hub municipality, home to at least one secondary school. Moreover, we assume that the part of student mobility reflecting chain migration only depends on the area of origin, neglecting in this way the mobility deriving from family and/or friendship ties in the destination region of historical type. Finally, the AreaOris were constructed based on students who enrolled in 2008/09 and were kept constant until 2017/18, in order to make feasible comparisons. Furthermore, such AreaOris avoid issues related to the size of municipalities/provinces (too small to capture mobility phenomena/too large and heterogeneous in terms of mobility). For the sake of brevity, the schematic procedure for the determination of the AreaOri is the following: *i)* we construct an origin-destination matrix M(i,j) where each row *i* is the municipality of residence and each column *j* represents municipalities of a high-school. The co-occurrences of each cell are the number of students, who originate from *i* and attended a high school in the municipality *j*; *ii)* from M(i,j) we choose J destination municipalities home to at least one high school with at least 200 students. These municipalities are the hubs (say j_{hub}) used as starting points for the determination of the AreaOri; *iii)* each origin *i* is assigned to the j_{hub} that is able to attract the majority of students from *i*.

As a final step of our analysis and in order to better comprehend the presence of mobility patterns with respect to time, a hierarchical clustering algorithm was applied to the modalities of the pairs (AreeOri, RegDest) to two different sets of covariates: *i)* five-year periods and disciplinary field (STEM/non-STEM), and *ii)* five-year periods and gender. Indeed, our interest is not strictly constrained to the composition of resulting clusters; instead, it focuses on the ordering of variables implicitly provided by the hierarchical clustering that allows highlighting pattern specificities in the data.

This choice of covariates allows us to reveal the persistence or not in time of attraction/repulsion of the origin-destination pairs by isolating the pairs (AreeOri, RegDest) with respect to the five-year period. This choice also permit us to investigate if such "persistence" eventually depends on gender and/or disciplinary field. The interest in gender and disciplinary area derives from the fact that these two variables are associated with university mobility, mostly due to the greater propensity to the mobility of male students interested in the STEM disciplinary areas (see Attanasio et al. [3]). The hierarchical clustering procedure, as applied to the Euclidean distance matrix of residuals, enable us to elicit from data the similarity structure of origin-destination patterns. Indeed, hierarchical clustering constructs a dendrogram of origin-destination units by iteratively grouping them at reduced levels of similarity until they merge into only one cluster—the root. In other words, increasing levels of root depth correspond to larger levels of aggregation. In particular, the clustering algorithm used is the Complete Linkage Cluster Analysis that, compared to methods such as the Single Linkage, or the Average Linkage (weighted and unweighted), guarantees a better separation of the groups (Anderberg [1]) by avoiding the so-called chaining phenomenon, which is typical of clustering algorithms based on the nearest-neighbour distance. Complete Linkage effectively reduces the chaining phenomenon with respect to the other clustering methods since the iterative aggregation is based on the farthest-neighbour distance, thus creating more compact and homogeneous clusters (Anderberg [1], Everitt et al. [7]).

3 Results and discussion

Figures 1, 2 and 3 depict the cluster analysis results for Sicily, Sardinia and Apulia, respectively. The red colour indicates mobility patterns with flows greater than would be expected under the hypothesis of random flows, i.e., the ones expected based on push and pull factors; the blue colour refers to mobility patterns with flows less than expected under the hypothesis of random flows.

For Sicily, the cluster analysis confirms the idea that student migration chains play an important role in explaining mobility patterns. In particular, if we consider the non-STEM field of study, the AreaOri-RegDest are remarkable for both five-year periods in patterns such as: Palermo-Latium, Messina-Lombardy, Trapani-Tuscany, Vittoria-Tuscany, *etc*. It is worth mentioning also that different mobility patterns emerged between large Sicilian cities (Palermo, Catania and Messina) and smaller ones (the others). In particular, students from larger cities are more likely to study in universities located in big cities such as Rome, Milan and Turin. In contrast, students from smaller cities would move to study in smaller university cities.

Note that all mobility patterns linked with Piedmont – that are powered by chain migration – refer to STEM degrees. In Piedmont, the Polytechnic University of Turin (a university mainly devoted to engineering study programs) represents an important basin of attraction, especially for students from small Sicilian cities. Thus, the popularity of this university is due to both the quality of the engineering programs and the social connections between students from the same area of origin.

For the non-STEM programs, in Piedmont, there is a number of students lower than expected, and this result is also confirmed when disaggregated by gender[1]. Unfortunately, due to persistent problems of gender stereotypes, STEM subjects continue to be more prevalent among male students. Therefore, our results depict that Piedmont was chosen predominantly by male students.

As for Sardinia, the strong link between Sassari and Piedmont for scientific disciplines is evident over time. Similar significant connection with Piedmont area is observed for students originated from Nuoro and Oristano territories. In scientific fields, Tuscany is also linked—by a chain effect—to the territories of Sassari, Nuoro, and Olbia-Tempio. Looking at the non-STEM degree for Sassari, it is worth noticing that the network effect at the destination triggers a kind of repulsion effect from Tuscany.

Among the non-STEM, the chain effect seems to power Cagliari's links with Latium, Lombardy and Emilia-Romagna. Emilia-Romagna is also strongly linked to Oristano and Nuoro, while students from Olbia-Tempio seem to have stronger ties with Lombardy. When gender is also considered for Sardinia, mobility patterns related to scientific subjects are mainly composed of male students. Note that also in the case of Cagliari (the biggest city in Sardinia) emerges the same "from city to city" model as showed in Sicily.

Finally, considering Apulia, a partly different picture emerges with respect to the two islands: the mobility patterns formed for scientific subjects largely correspond to those for non-STEM subjects. However, this difference is driven by geographical and logistical reasons rather than a different behaviour model. Sardinia and Sicily, for obvious reasons, have no neighbours and perhaps excluding Milan and Rome (for which the highest number of flights are available), it is equally difficult/expensive to reach any other destination in the Centre-North. In contrast, for students from Apulia, it is relatively easy to reach Emilia-Romagna which seems to be the region that attracts the majority of students from both sectors of knowledge. Therefore, the chain effect could be somewhat amplified by the relative ease of connection between Apulia and the areas of destination. This could also explain the lack of a gender gap both for scientific subjects and mobility patterns in Apulia. Thus, the easiness of moving lowers the monetary and psychological cost of moving. This, in turn, makes students more likely to go outside their regions of origin regardless of subject area and gender.

[1] Due to space constraints, gender and five-years heatmaps are not reported here. However, these are available upon request to the authors.

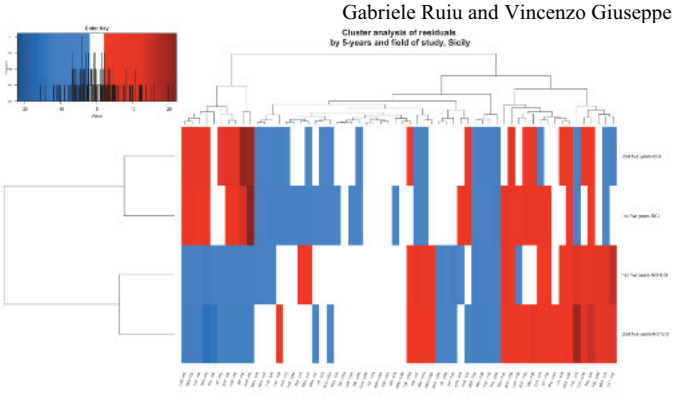


Figure 1: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Sicily

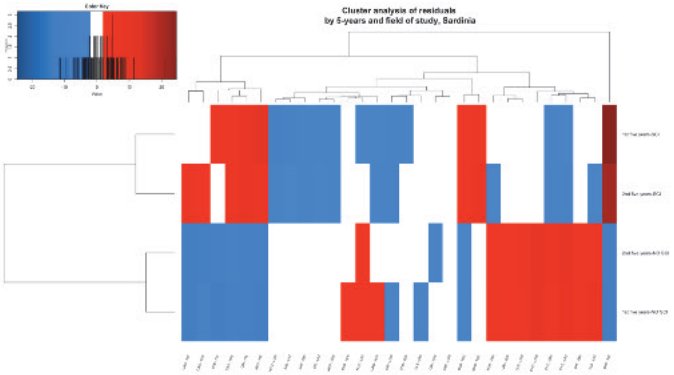


Figure 2: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Sardinia

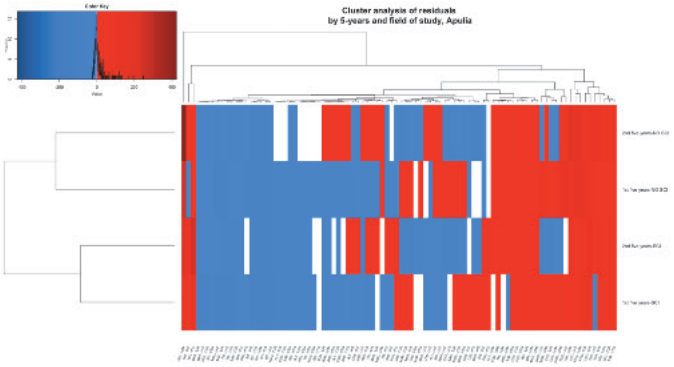


Figure 3: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Apulia

References

1. Anderberg, M. R.: Cluster Analysis for Applications, Academic Press, New York (1973).
2. Attanasio, M., Enea, M.: La mobilità degli studenti universitari nell'ultimo decennio in Italia. In: De Santis, G., Pirani, E., Porcu, M. (eds.), Rapporto sulla popolazione. L'istruzione in Italia, 43-58. Il Mulino, Bologna (2019).
3. Attanasio, M., Giambalvo, O., Porcu, M., Ragozini, G.: Verso Nord. Le nuove e vecchie rotte delle migrazioni universitarie. FrancoAngeli, Milano (2020).
4. Boscaino, G., Sottile, G., Adelfio, G.: Migration and students' performance: detecting geographical differences following a curves clustering approach. J Appl Stat (2020) doi: 10.1080/02664763.2020.1845624
5. Brooks, R., Waters, J.: International higher education and the mobility of UK students. J Res Int Educ 8(2):191–209 (2009).
6. D'Agostino, A., Ghellini, G., Longobardi, S.: Out-migration of university enrolment: the mobility behaviour of Italian students. Int J Manpower 40(1):56–72 (2019).
7. Everitt, B.S. and Landau, S. and Leese, M. Stahl, D.: Cluster Analysis. Wiley, Chichester (2011).
8. Genova, V.G., Tumminello, M., Aiello, F.et al.: A network analysis of student mobility patterns from high school to master's. Stat Methods Appl 30,1445–1464 (2021).
9. MOBYSU.IT Database MOBYSU.IT, Mobilità degli studi universitari italiani, Protocollo di ricerca MIUR—Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II, Fonte dei dati ANS-MIUR/CINECA (2016).
10. Impicciatore R, Tosi F: Student mobility in Italy: the increasing role of family background during the expansion of higher education supply. Res Social Stratif Mobil doi: 10.1016/j.rssm.2019.100409 (2019)
11. Pérez, P.A., McDonough, P.M.: Understanding Latina and Latino College Choice. A Social Capital and Chain Migration Analysis. J Higher Educ 7(3): 249-265 (2008).
12. Ruiu G., Fadda N., Ezza A., Esposito M.: Exploring mobility of Italian Ph.Ds over the last decades. Electron J of Appl Stat Anal 12(4),748-773 (2019).
13. Santelli, F., Scolorato, C., Ragozini, G.: On the determinants of student mobility in an interregional perspective: a focus on Campania region. Italian J Appl Stat 1:119–142 (2019).
14. Tocchioni, V., Petrucci, A.: Italian PhD students at the borders: the relationship between family background and international mobility. Genus doi: 10.1186/s41118-021-00127-5 (2021)

A new bipartite matching approach for record linkage: the case of two big Italian databases

Un nuovo approccio per il record linkage basato sul matching bipartito: il caso di due grandi database italiani

Martina Vittorietti, Andrea Priulla, Vincenzo Giuseppe Genova, Giovanni Boscaino, Ornella Giambalvo

Abstract In recent years, university student mobility in Italy has worsened the north-south economic divide. Therefore, studying this phenomenon and its determinants is necessary to provide helpful information to support socio-economic policies. Thus, this paper aims at integrating two big databases about university students in Italy: the first one is provided by the Ministry of University and Research, concerning the university careers of student cohorts; the second one is provided by the AlmaLaurea consortium, concerning the university experiences of graduates and their success in the labour market. Both databases contain socio-demographic information that complements each other. The proposed method is a modification of the Fellegi-Sunter method, which, from the preliminary outcomes, seems to achieve very satisfactory results.

Abstract La crescente mobilità studentesca universitaria in Italia, negli ultimi anni, ha esarcebato il divario economico nord-sud. Pertanto, è necessario studiare tale fenomeno e le sue determinanti, in modo da fornire informazioni utili a supporto delle politiche socio-economiche. Questo paper mira a integrare due grandi database relativi agli studenti universitari in Italia: il primo è fornito dal Ministero della Università e Ricerca, e riporta le informazioni sulle carriere universitarie di coorti di studenti; il secondo è fornito dal consorzio AlmaLaurea, relativo alle esperienze universitarie dei laureati e al loro successo nel mondo del lavoro. Entrambi i database contengono informazioni socio-demografiche che si completano a vicenda. Il metodo proposto è una modifica del metodo di Fellegi–Sunter e i primi risultati sembrano essere molto soddisfacenti.

Key words: Matching, Record linkage, University students, AlmaLaurea

---

Martina Vittorietti, Andrea Priulla, Vincenzo Giuseppe Genova, Giovanni Boscaino (corresponding author), Ornella Giambalvo
Department of Economics, Business and Statistics – University of Palermo, Italy.
e-mail: martina.vittorietti@unipa.it, andrea.priulla@unipa.it, vincenzogiuseppe.genova@unipa.it, giovanni.boscaino@unipa.it, ornella.giambalvo@unipa.it

1 Introduction

In recent years, the attention towards student mobility flows has increased in Italy [2, 3, 8]. In fact, student mobility has grown, typically characterised by one-way flows, and from the south to the central north. Often, young people who leave never return home. That has considerable repercussions on the country's socio-economic structure. The poorer southern regions become even poorer to the advantage of the wealthier regions, becoming more attractive to young people. Whereas in the past, migration was characterised by workers moving in search of their fortune, the mobility of young people is now anticipated at the university level, especially when they enrol in a master's degree course. Young people seem to prefer studying in the centre-north because they are attracted by the so-called "brand universities", i.e. famous and renowned universities [4], and the more favourable economic context in which they are set. Students perceive these features as winning in the labour market.

Thanks to an agreement between eight Italian universities and the Ministry of Universities and Research (MUR), we can now access the Ministerial database (MOBYSU.IT [5]) that collects information about the careers of all the freshmen in all the Italian universities since 2008. In particular, the longitudinal micro-data allows us to track students' trajectories in terms of career events (dropout, degree course changes) and their regional mobility. At the university level, information about students' mobility could be of interest to gain knowledge about the characteristics of those graduates who moved to another university after bachelor's degree completion.

Another helpful source of information comes from the AlmaLaurea surveys. AlmaLaurea is an Interuniversity Consortium that currently counts 78 Universities as members. It is mainly supported and funded by the Universities that are part of it and by funds from MUR. Among its aims, AlmaLaurea performs several surveys on graduates about their social-economic background, academic experience, and theur occupational status 1, 3, and 5 years after graduation.

The two databases contain both socio-demographic and career information that, in some cases, overlap and, in others, complement each other. Therefore, merging the two databases would allow tracing in detail the trajectory of each student from the first university enrolment in a first-level degree course, up to five years after graduation, for a total of 10 years of information per record.

So, in future, we could answer questions such as i) "Is there a difference between the occupational success of southern students that studied in a southern university and that one of the southern students that studied in a northern university?"; ii) "is the social level class influencing the decision to enrol in a master degree course in a different university after bachelor completion?"; iii) "is the academic performance a predictor of the occupational success of the graduates?".

Hence, the integration of the two databases appears crucial, and this paper aims to merge them following a new approach based on bipartite record linkage [14].

We have available AlmaLaurea data about University of Palermo (Italy) graduates. Usually, a key column that matches the records is necessary to merge two or more datasets. For privacy reasons, that column (i.e. the student's registration num-

ber) is not available. Therefore, we needed a merging approach to match the records, and we used the bipartite matching one.

2 Data

As mentioned, the databases used in this paper come from two distinct sources. In detail:

MOBYSU.IT longitudinal micro-data coming from MUR containing information about the university careers of all the students enrolled in every Italian university. It contains information about first and second level Degree Courses of enrolment, Field of Study, High School Diploma, High School Grade, and some social and demographic information on students like Gender and date and place of birth.

AlmaLaurea survey data about the population of University of Palermo (Italy) graduates. In particular, two different type of surveys are administered:

- *profile survey*: it collects information about students' experience (e.g. satisfaction about the facility, satisfaction for the relationship with teachers), information about their university career (i.e. graduation delay, willingness to carry on studying), and their socio-demographic characteristics (e.g. gender, parents' socio-economic status). These data are enriched and adjusted with students' personal information directly provided by the partner universities.
- *additional post-graduation survey*: administered 1, 3, and 5 years after the degree, to obtain information about graduates' job conditions, job description, earning, study-job coherence and job satisfaction.

3 Methodology

The central assumption for bipartite record linkage is that one unit (i.e. the student) is recorded at most once in each database, so a record from one can be linked with just one record from the other database. Therefore, consider two databases X_1 and X_2 that record information from two overlapping sets of individuals. These databases contain n_1 and n_2 records respectively, with $n_1 \geq n_2$. In both files, there could be errors due to the record-generating process or missing values. We assume that there are no duplicates in both files.

Let n_{12} be the number of individuals simultaneously recorded in both databases, $0 \leq n_{12} \leq n_2$. A bipartite matching can be represented in different ways [14]. The aim is to create a matching matrix Δ of size $n_1 \times n_2$ whose (i, j)th entry is $\Delta_{ij} = 1$, if $i \in X_1$ and $j \in X_2$ identify the same individual, $\Delta_{ij} = 0$ otherwise. One of the most popular approaches to record linkage is the Fellegi–Sunter approach [7], later

modified and re-adapted by many authors (see a review in [6]). The main idea behind the Fellegi-Sunter approach is to perform pairwise comparisons of the records to estimate the matrix Δ. In this paper, we propose a variation of the classical Fellegi-Sunter approach. This allows us to directly employ the distance between the chosen fields of comparison and prevent us from assuming specific probability distributions for matching and using computationally expensive algorithms for obtaining those probabilities.

Let Γ^k be a matrix of dissimilarities, whose element γ^k_{ij} in $X_1 \times X_2$ indicates the dissimilarity measure for the pair (i, j) with respect to the k-th field. Using comparable distance metrics, we can compute

$$\Gamma = \Gamma^1 + \Gamma^2 + \cdots + \Gamma^k, \tag{1}$$

where the element γ_{ij} represents the overall dissimilarity of the record $i \in X_1$ to the record $j \in X_2$. Multiple similarity measures have been used in record linkage approaches. The similarity is easier to measure for numeric data, as reasonable options are Manhattan, Euclidean or Mahalanobis distances. For text-based fields, the similarity measures are more complex [1]. The Levenshtein distance, the Damerau-Levenstein distance, and the Longest Common Substring (LCS) distance are commonly used methods of comparison of two text strings [13]. In the classical Fellegi-Sunter approach, the comparison based on the matrices/vectors of dissimilarities is considered insufficient to determine the matches since the variables being compared usually contain random errors and missing values [14]. In the reference literature, researchers often treat missing data as disagreements, performing imputations or ignoring them because they assume a missing data scheme such as missing at random [9]. In this paper, we propose dealing with them in the computation of the dissimilarity matrix, adding a correction factor that considers missing values and random error. Let Λ be a matrix whose element $\lambda_{ij} = \frac{K}{\sum_{k=1}^{K} \mathbb{1}(\gamma^k_{ij}=0)}$ indicates the inverted proportion of exactly equal fields between the pair (i, j), with respect to the total number of fields K. To enforce the one-to-one constraint of the bipartite matching, we use the optimal assignment record pairs procedure proposed by [11], obtained from the linear sum assignment problem:

$$\min_{\Delta} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \gamma_{ij} \lambda_{ij} \Delta_{ij}$$

$$\text{subject to } \Delta_{ij} \in \{0,1\}; \sum_{i=1}^{n_1} \Delta_{ij} \leq 1, j = 1,2,\ldots,n_1; \tag{2}$$

$$\sum_{i=1}^{n_1} \Delta_{ij} \leq 1, j = 1,2,\ldots,n_2. \tag{3}$$

The constraints ensure that Δ represents a bipartite matching. We employed the Hungarian algorithm [12] to solve the optimization problem.

It is worth noticing that the matching procedure becomes too computationally expensive for extensive databases. A usual solution is to partition the databases into blocks of records determined by the information that is thought to be accurately recorded in both databases, and then solve the task only within blocks [10].

In paragraph 1 we have pointed out that the two databases do not have a key column to match the records. The unit of the University of Palermo dealing with student enrolment and career data (called SIA) has access to the key column (hidden from us for privacy reasons) and provided us with the merged big database of MOBYSU.IT and AlmaLaurea, but only for the data of Palermo. SIA database is provided without the key column (to ensure the students' privacy) and with just the MOBYSU.IT data aligned with those of AlmaLaurea. So, since they have columns in common, these are repeated. Unfortunately, SIA procedure is hardly reproducible for all Italian universities, so obtaining a procedure that merges the databases without needing the key column is essential.

We decided to work with SIA dataset to verify the "quality" of our proposed merging procedure. In brief, we split the SIA database into the part coming from MOBYSU.IT and the part coming from AlmaLaurea. We then merged these two parts following our proposed method and verified that the matches of the record were correct by comparing them with the SIA dataset.

In detail, our merging procedure can be described as follows:

1. Identify the common fields in the two datasets. For this purpose, we distinguished fields into two categories:
   - *socio-demographic context*: Gender, Place of Residence, High School Track, Final Grade, Year Date of High School Completion, and Age at Degree Completion;
   - *university*: University Identification Code, Place and Identification of the Degree Course, Year Date of Degree Completion, and Final Graduation Grade.
2. Data cleaning and homogenization of the information of the two datasets. For instance, the High School Track was encoded differently into the two databases; hence before running the matching procedure, the same categorization was applied to the variable in the two databases.
3. Select blocking variable/s. We selected the *Gender* as the only blocking variable. In fact, it is the variable with fewer missing values and possible data entry errors, and, in addition, we cannot allow matching observations that have this field not equal.
4. Compute matrices Γ^ks:
   - The LCS dissimilarity measure is used for the character fields. This metric returns the number of unmatched characters; therefore, higher values of it correspond to less similar records.
   - For the numeric fields, the absolute difference between the two fields is used;
5. Compute overall dissimilarity matrix Γ (Eq. 1) and its correction matrix Λ;
6. Use $\Gamma \odot \Lambda$ as cost matrix for the Hungarian algorithm.

4 First Results and remarks

Table 1 reports the results of the proposed matching procedure, performed on University of Palermo and Almalaurea data. It is worth noticing that the procedure – regardless of the cohort of students, produces a rate of correct match greater than 98%. Furthermore, we studied the effect of the correction factor Λ on the "quality" of the matching. Results show that using the combined dissimilarity matrix corrected for the proportion of exactly equal fields as cost matrix for the Hungarian algorithm we can obtain more than 98% correct matches without making heavy theoretical assumptions. In addition, our method tackles common problems such as entry errors and missing values, including them in a correction factor.

Finally, these preliminary results, obtained using the dummy dataset provided by SIA, are promising. Therefore, we expected that when AlmaLaurea databases from other universities are available, the matching procedure will produce results as good as these. In such a way, we will enrich the whole MOBYSU.IT database with crucial and helpful information from more universities to investigate deeper, for example, the determinants of students' performance, mobility, job success, gender inequalities in salary and job position.

	Gender			
	Male		**Female**	
Cohort of enrolment	**% (without Λ)**	**% (with Λ)**	**% (without Λ)**	**% (with Λ)**
2010	99,77	99,85	98,88	98,88
2011	99,05	99,31	98,39	98,72
2012	99,58	99,83	99,14	99,43
2013	99,42	99,33	98,63	99,17
2014	99,28	99,82	99,09	99,58
2015	99,19	99,59	98,71	99,39
2016	99,26	99,38	99,03	99,63
2017	99,48	99,48	99,66	99,83

Table 1 Percentages of correct matches (with and without the correction factor Λ) between SIA and AlmaLaurea databases, by Cohort of enrolment and Gender.

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P.

References

1. Asher, J., Resnick, D., Brite, J., Brackbill, R., Cone, J.: An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. INT J ENV RES PUB HE 17.18 (2020): 6937.
2. Attanasio, M., Enea, M., Albano, A.: Dalla triennale alla magistrale: continua la "fuga dei cervelli" dal Mezzogiorno d'Italia. Neodemos, ISSN: 2421-3209 (2019)
3. Boscaino, G, Genova, V.G.: Exploring drivers for Italian university students' mobility: first evidence from AlmaLaurea data.. In C. Perna, N. Salvati, Schirripa Spagnolo F. (Eds.), Book of Short Papers SIS 2021 (pp. 1394-1399). Pearson. (2021)
4. Columbu S., Porcu M., Sulis I.: University choice and the attractiveness of the study area: Insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models. SOCIO ECON PLAN SCI DOI:10.1016/j.seps.2020.100926 (2021)
5. Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia], research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II. Scientific Coordinator Massimo Attanasio (UNIPA), Data Source ANS-MUR/CINECA
6. Enamorado, T., Fifield, B., Imai, K.: Using a probabilistic model to assist merging of large-scale administrative records. AM POLIT SCI REV 113.2 (2019): 353-371.
7. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J AM STAT ASSOC 64.328 (1969): 1183-1210.
8. Genova, V.G., Tumminello, M., Enea, M., Aiello, F.: Student mobility in higher education: Sicilian outflow network and chain migrations". Electron J Appl Stat Anal: DOI:10.1285/i20705948v12n4p774 (2019)
9. Harron, K., Goldstein, H., and Dibben, C.: Methodological developments in data linkage. John Wiley & Sons, (2015).
10. Herzog, T. N., Scheuren, F.J., and Winkler, E. W.: Data quality and record linkage techniques. Vol. 1. New York: Springer, (2007).
11. Jaro, M. A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J AM STAT ASSOC 84.406 (1989): 414-420.
12. Kuhn, H. W.: The Hungarian method for the assignment problem. NAV RES LOGIST Q 2.1-2 (1955): 83-97.
13. Navarro, G.: A guided tour to approximate string matching. ACM COMPUT SURV 33.1 (2001): 31-88.
14. Sadinle, M.: Bayesian estimation of bipartite matchings for record linkage. J AM STAT ASSOC 112.518 (2017): 600-612.

Statistical Methods for Science Mapping

A word embedding strategy to study the thematic evolution of ageing and healthcare expenditure growth literature

Una strategia di word embedding per lo studio dell'evoluzione tematica della letteratura su invecchiamento e crescita della spesa sanitaria

Milena Lopreite, Michelangelo Misuraca and Michelangelo Puliga

Abstract The impact that ageing has in terms of growing investments in long-term care and the reflection of increasing shares of older people in every sector of activity is a primary concern for who governs and for policymakers. As a consequence, the relation between ageing and healthcare expenditure growing is more and more studied by scholars interested in social and economic issues and scholars interested in health issues. A bibliometric analysis of publications related to this domain may highlight the drivers of this relation and tracks the evolution of the debate about this matter. Here a strategy based on word embedding is proposed, showing how this natural language processing approach can unveil the knowledge base embodied in the reference literature and offer the scientific community valuable insights.

Abstract *L'impatto che l'invecchiamento ha in termini di crescita negli investimenti per l'assistenza a lungo termine e il riflesso di maggiori quote di anziani in ogni settore di attività è una preoccupazione primaria per chi governa e per i responsabili politici. Di conseguenza, il rapporto tra invecchiamento e crescita della spesa sanitaria è sempre più approfondito da studiosi interessati alle problematiche sociali ed economiche e da studiosi interessati alle problematiche sanitarie. Un'analisi bibliometrica delle pubblicazioni relative a questo dominio può evidenziare i fattori caratterizzanti questa relazione e tracciare l'evoluzione del dibattito sull'argomento. In questo lavoro è proposta una strategia basata sul word embedding, mostrando come questo approccio di elaborazione del linguaggio naturale possa svelare la base di conoscenza incorporata nella letteratura di riferimento e offrire alla comunità scientifica spunti preziosi.*

Key words: science mapping, thematic analysis, natural language processing

Milena Lopreite
DESF - University of Calabria, Arcavacata di Rende, e-mail: milena.lopreite@unical

Michelangelo Misuraca
DiScAG - University of Calabria, Arcavacata di Rende, e-mail: michelangelo.misuraca@unical.it

Michelangelo Puliga
CNC Laboratory - Linkalab, Cagliari, e-mail: michelangelo.puliga@linkalab.it

1 Introduction

In modern societies, the mechanisms relating to health and the investments in healthcare are affected by an increasing share of older people. Ageing has a substantial impact on every sector of a country, including economic growth, labour market, housing, migration and health [17, 22, 6]. People 65 years and over can be considered a sort of "burden" for a society in economic terms, as they benefit from welfare systems but do not actively contribute to creating wealth with their activities. A change in the demographic structure leads, for example, to a higher incidence of chronic-degenerative diseases (e.g., heart disease, cancer, Alzheimer's disease) and a greater demand for long-term care. The expected effect is a higher per-capita health cost, undermining over time the financial sustainability of the healthcare systems in terms of systemic performances and healthcare supplies. Governments have to face the problem of increasing demand for public health services combined with a strain on the available resources. Many studies [9, 15, 41, 4] underlined the relevance of the relationship between ageing and healthcare expenditure and how it represents a primary concern for those addressing health policy interventions, both at an international and a national level.

This study aims to better understand the consequences of population ageing on health-expenditure growth, highlighting the potential and less-known key drivers influencing this relationship over time. The recent development of telemedicine and e-health and the introduction of devices that control and encourage physical activity may be relevant in this sense. The technological progress can help reduce chronic illness in older people, limit expenditures on home nursing, home health care, personal care, adult daycare and multiple visits to the physician and hospital, with significant gains in long-term health care. This aspect is still poorly inspected. The literature on medical devices is scarce and related to the revolution of the "Internet of things" (IoT) that is changing the playground in many fields, from industry to healthcare, to remote monitoring of patients. We aim to improve the debate on the effects of ageing on health spending analysing the scientific literature related to this domain, considering the emerging Covid-19 pandemics and the challenges that the epidemic posed to the healthcare system, particularly to the diagnostics for the elders.

Reviewing the literature related to this issue has several limitations due to the different results obtained when micro or macro-level data are used [18, 37, 27, 29]. Instead of getting data from prior hypotheses on the analysed phenomenon, a statistical text analysis of the scientific publications is proposed and implemented – in the framework of the so-called *science mapping* [24, 8] – trying to systematise the knowledge emerging from the ageing literature and identify the main discussed topics and their evolution across time. Topic analyses on bibliographic records typically use keywords attributed by publications' authors to categorise them and make their indexation and retrieval from citation databases easier. To overcome the problem induced by the higher variability of these keywords, often also the keywords automatically attributed by the databases are used. The informative power of the two sets of keywords is limited by the absence of their context of use, allowing the detection of topics that are sometimes hard to understand and connect to the debate

concerning the research under investigation. To accomplish this task, here we developed a strategy based on *word embedding* [21], a natural language processing approach that captures the semantic structure of a text, allowing us to encompass the context surrounding the different terms used in a textual collection.

After providing an overview of the theoretical background of science mapping and introducing the proposed strategy in Sec. 2, the main findings of a study on the recent 12-years literature related to ageing and healthcare expenditure are presented in Sec. 3. Some final remarks and future developments conclude the paper in Sec. 4.

2 Theoretical background and proposed strategy

Literature reviews are commonly used to assess the state-of-the-art and the primary trends of a given scientific sphere or a research issue. [19] tried to classify the approaches typically followed by scholars, taking into account the different scopes and techniques used to explore the reference literature of a domain systematically. The majority of these approaches rely on qualitative techniques, but more and more, a quantitative viewpoint is considered for this purpose. Literature reviewing tasks can be accomplished using statistical techniques, thanks to the availability of online databases collecting the diverse publications and software tools able to perform automatic analyses on massive amounts of data.

A bibliographic record retrieved from an indexing database − e.g., Web of Science, Scopus or PubMed − lists different information about a publication, such as papers published in journals or conference proceedings or book chapters included in edited volumes. Numerical and categorical data concerning the documents themselves and their references, the authors and their corresponding countries/institutions can be used to evaluate the relevance and the impact of the publications as well as the social and intellectual structure of the involved actors [35, 42]. Textual data concerning the documents − typically referring to their titles, abstracts and keywords − can be used instead to depict the conceptual structure of the analysed domain, visualising the main themes discussed from a theoretical and empirical viewpoint with a synchronic [34], or a diachronic overview [16, 40].

The typical data structure used to detect and map the cognitive frames of a domain or an issue of interest derives from the *vector space model* (VSM) [38]. VSM is an algebraic representation of texts that allows transforming unstructured data included in a textual chunk into a set of structured data (in the form of vectors) that can be quantitatively treated. The encoding scheme beneath this model is the so-called *bag of words* (BoW), in which each textual chunk is represented as a multi-set of its basic components (i.e., the terms used in the text), keeping their multiplicity but disregarding the grammatical roles. The different text-vectors can be arranged in a lexical table $\mathbf{F}$ cross-tabulating the set of n texts belonging to the analysed collection and the different p used terms. The adoption of BoW simplifies the computational treatment of extensive textual collections, allowing skipping terms order. However, at the same time, it atomises the text limiting the possibility of considering the con-

text of the use of the different terms. To cope with this shortcoming, it is possible to build a p-dimensional co-occurrence matrix $\mathbf{A}$, counting for the joint use of the terms in the collection, with a granularity following the level of the textual chunk from the whole text to the single clauses. The latter data structure also has the advantage that it may be viewed as an adjacency matrix and depicted as a graph, allowing network analysis tools to detect the main topics from the terms and their semantic relations. Several techniques have been used to map the conceptual structure in a bibliometric framework, considering the two different data structures as a basis. Starting from the early works of M. Callon's research group [10, 11], concerning the use of *co-word analysis* as a "systematic content analysis of publications", it has been developed the so-called *thematic analysis* [12]. This technique allows visualising the topical patterns in a given temporal horizon or tracking their evolution across different time slices [14, 3]. Despite the popularity of this approach, also due to the diffusion of bibliometric tools and libraries such as *SciMAT* [13] or *bibliometrix* [2] that allow analysing the conceptual structure in a relatively simple way, thematic analysis has some drawbacks.

Firstly, because of the typical data structure used in the analysis, the publication dimension is not considered. The co-occurrence matrix only expresses how many times a couple of terms is jointly used in the collection or at least considers another kind of similarity measure to define the term network and reconstruct the context. This means that all the publications are considered simultaneously without using any meta-data that can discriminate the content with respect to the different research categories or domains. Secondly, topics are built only on the basis of the keywords (chosen by the authors or attributed by the indexing database), or other content-bearing words that can be derived from the abstracts, making challenging the labelling of each topic and its interpretation for the investigated issue. Some authors proposed to use *topic modelling* to overcome these limits since this approach allows considering at the same time the topical prevalence (i.e., the per-document topic distributions) and the topical content (i.e., the per-topic word distributions) [20, 39]. The advantage is that it is possible to characterise the different topics with respect to the keywords and also compute the topic similarity [30]. Moreover, it is possible to include meta-data in the analysis to explore the conceptual structure conditioned to some covariates of interest [36]. On the other hand, topic modelling is based on VSM and BoW as thematic analyses, sharing the same problems concerning the context of the use of the different terms. Moreover, even if both approaches are unsupervised, topic models require to prior set the number of topics and performing model selection − without a shared strategy and with alternative proposals that are often counter-intuitive and hard to carry out − whereas thematic analysis automatically determine the optimal number of topics to consider.

An alternative representation of texts that allows quantitative processing of textual collections is the so-called *word embedding* [21]. Word embedding refers to language models and feature extraction methods whose primary goal is to map terms or phrases into a low-dimensional continuous space. Differently from BoW, both semantic and syntactic information of terms are encoded. Semantic information mainly correlates with the meaning of terms, while syntactic information refers

to their structural roles in the texts. The different models can be classified as either paradigmatics or syntagmatics, looking at the term distribution. The given text region where terms co-occur is the core of the syntagmatic models, whereas exploring similar contexts is the key to the paradigmatic models.

The use of word embedding in a bibliometric framework has been recently explored, for example, to extract and visualise topics from bibliographic data [25, 43], or to detect topics and analyse correlations between publications [23]. Following this new emerging frontier, here we propose a strategy based on word embedding that aims at exploring the conceptual structure of a scientific domain and tracking its temporal evolution. In particular, we jointly use *word2vec* [31] and *doc2vec* [26] algorithms to encode the terms and then represent the set of abstracts on which the topic extraction is performed. The language model used in word2vec is based on a two-layer neural network [5] (namely, a shallow neural network) trained to reconstruct the contexts of use of the different terms. It processes a text collection to produce a vector space, usually of several hundred dimensions, in which each unique term is represented as a distinct point. Term vectors are positioned in the vector space such that terms that share common contexts in the text are located close to one another. Once terms are vectorised, they are processed in a subsequent stage by the doc2vec to derive a set of text vectors. This text representation is used in the analysis. Given a collection of publications retrieved from an indexing database and focusing on the abstract of each publication, the different stages of the strategy can be summarised as follows:

1. a t-years rolling window is created on the textual collection, according to the analysed time horizon;
2. after pre-treating the texts, a word2vec algorithm is employed to vectorise the terms belonging to the abstracts within each window of t years;
3. the t-years vectors for each term belonging to the abstracts are used to represent each abstract through a doc2vec algorithm;
4. from the set of vectors representing the abstracts, a similarity matrix $\mathbf{M}$ based on cosine similarity is built and depicted as a network;
5. a community detection procedure based on the *Louvain* algorithm [7] is employed to group the abstract into k sub-networks;
6. after fixing a threshold on community size, the keywords characterising each community are analysed according to their normalised frequency;
7. the changes in the frequency and composition of keywords belonging to the communities across time are counted and highlighted to track the evolution of the topical structure.

The rationale underpinning the strategy is that we first consider clusters of similar publications (according to the content of their abstracts), then we explore the most important keywords describing each cluster – highlighting the main topics discussed in the included publications – and how they change across the time. In the following, the strategy is applied to a set of publications concerning the literature related to ageing and healthcare expenditure.

3 Empirical evaluation and main findings

To determine which topics have been mostly discussed in the literature about the potential link between ageing and health expenditure, in March 2022, we accessed the Web of Science (WoS) database to build a bibliographic dataset. WoS – early developed by the Institute for Scientific Information and now maintained by Clarivate Analytics – is one of the leading databases to explore the literature of a research domain. It incorporates several citation databases specialised in given scientific fields (e.g., the Social Science Citation Index for Social Science), covering more than 20,000 journals, conference proceedings and books. We used the query ("aging" OR "ageing") AND ("health*") AND ("expen*" OR "cost$" OR "spending") in the WoS field "topic" (including title, abstract, authors' keywords and WoS keywords) to retrieve the publications related to this research area. The number of records downloaded initially was 8,336. Then, we considered only original articles published in scientific journals in the analysis. A careful review led to the inclusion of articles with an abstract written in English. Subsequently, we selected the last 11-years publications (from 2011 to 2021) to consider the most recent literature, obtaining at the end of the process 4,171 complete records. An additional check led to the exclusion of 7 publications without an abstract. Fig. 1 depicts the flow of the different searching steps, mapping out the number of identified publications, the included and excluded ones, and the reasons for the exclusions [33].

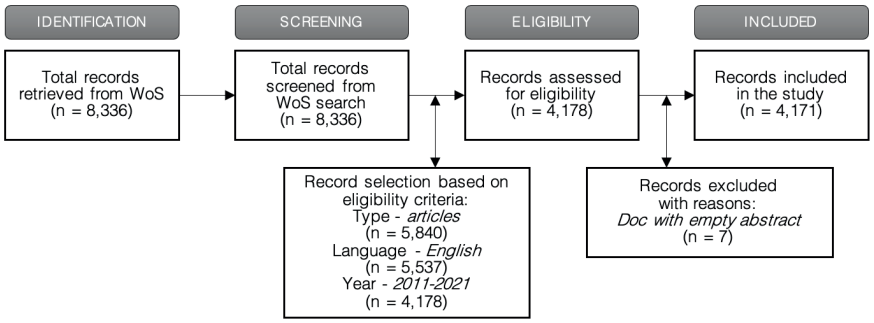


Fig. 1 PRISMA diagram related to the present study.

The 4,171 publications included in the study were written by 20,659 different authors, with a share of single-authored articles of just 7.24% and an average number of co-authors per article of 6.4, showing a high degree of collaboration among scholars in the research domain. Concerning the expansion of the scientific production across the 11-years time horizon, we observed an annual growth rate of 10.9%.

To implement the strategy described in Sec. 2, we decided to set a 5-years rolling window, from 2011 to 2021. The word2vec algorithm employed to vectorise the terms was set to scan segments of 4 adjacent terms through a *skip-gram* [32] procedure, predicting the source context terms (the surrounding terms) for each given

target (the centre term). The similarity matrix built from the doc2vec representation of abstracts was filtered considering a cosine similarity greater than 0.8 to consider only the core publications sharing a common knowledge base. Finally, we performed the community detection saving only the communities containing more than 15 abstracts. Fig. 2 shows the communities detected through the latter procedure (only the biggest 8 communities are highlighted by way of example).

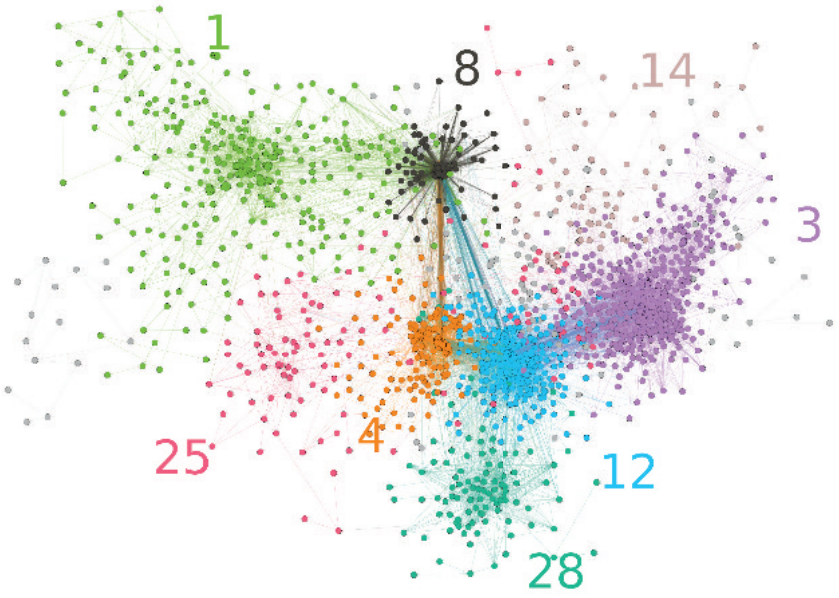


Fig. 2 Community detection on the abstract network (2011–2021).

The analysis pointed out different kinds of publications. Community 1, for example, includes publications on the development of e-health and telemedicine to support older people, especially during the coronavirus outbreak. Community 12 includes medical publications on the impact of exercise and physical activity of older people. Community 14 includes studies concerning the policies associated with the incidence of several chronic degenerative diseases the burden that economies have to sustain. Fig. 3 depicts the temporal evolution of the topics discussed by publications included in the last described community.

The growing increase of chronic degenerative diseases affects the health status of the elders generating a greater use of health resources and more pressure on health spending. In fact, in many studies several diseases related to cardiovascular diseases (atrial fibrillation and strokes) are analysed looking at their impact on the healthcare systems, their costs and their relationship with the population ageing. From a dynamical perspective, the studies widened from cardiovascular diseases to

Fig. 3 Temporal evolution of publications included in community 14.

osteoporosis and diabetes. The term "medicare" (the Obama programme for public health) appeared in the literature starting from 2013–2017 together with several studies analysing the impact of disabilities related to the diseases of elders. From 2013–2017, other studies addressed the problem of HIV and the burden that this disease induced in Western countries (e.g., Canada and the US). The interest in HIV for the old people of such rich countries is somehow unexpected, as usually HIV – and the frequent co-morbidity with Tuberculosis – is often associated with developing countries [28]. In the period 2015–2019, the analysis is enriched by studies forecasting the future evolution of the healthcare costs, with the simulation of specific

scenarios related to the principal chronic diseases. Forecasts about the economic impact are made for example in countries such as Japan where the population is ageing with a sustained rhythm. In more recent years, the literature extended these analyses to developing countries and to cost-effectiveness of prevention programmes.

4 Conclusion and final remarks

In recent years, the average age of many countries increased, transforming the demographic structure for the presence of a more significant share of older people. The impact of ageing on healthcare expenditure may be influenced by new core driver variables such as the new technologies that create opportunities and challenges. The result of our literature analysis, with the help of the semantic abilities of the word embedding techniques, confirms these transformations, putting in evidence how the improvements in medical care technologies are strictly related to the changes in the demographic structure. The introduction of new technologies such as artificial intelligence, and the Internet of things revolution, empowered the healthcare systems, creating, for example, emergency telemedicine that during the Covid-19 pandemics was of incredible importance for the most fragile population.

Literature clearly shows the impact in terms of the economic burden of an ageing population and the enormous potentialities of telemedicine. This shift from sociological and health economics studies to medical experiments on the ageing effects is an essential enrichment for this field of study. Reducing the distance between medical experiments, with their rigorous randomised trial methodologies, and the social studies is also important to bring the most evident results to policymakers.

According to our bibliometric analysis, we foresee the profound influence of the most recent trends: artificial intelligence applied to telemedicine devices, the ability to remotely and privately monitor health status proactively to improve health care values for the elderly in the long-term. The use of word embedding offered valuable insights and a more interpretable knowledge base with respect to other techniques used to automatically explore scientific literature, such as thematic analysis or topic modelling. The advantage of grouping publications with a similar abstract allows for better discrimination among studies developed in different research domains, distinguishing the different topics debated by scholars. Future developments of this study will be directed to better visualise the topics and their evolution and the inclusion of other covariates of interest that can enhance the understanding of the research domain under investigation.

References

1. Aisa, R., Clemente, J., Pueyo, F.: The influence of (public) health expenditure on longevity. Int. J. Public Health **59**, 869–875 (2014)

2. Aria, M., Cuccurullo, C.: bibliometrix. An R-tool for comprehensive science mapping analysis, J. Informetr., **11**, 959–975 (2017)
3. Aria, M., Misuraca, M., Spano, M.: Mapping the evolution of social research and data science on 30 years of Social Indicators Research. Soc. Indic. Res. **149**, 803–831 (2020)
4. Baltagi, B.H., Moscone, F.: Health care expenditure and income in the OECD reconsidered: evidence from panel data. Econ. Model. **27**, 804–11 (2010)
5. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, New York, NY (1995)
6. Blanco-Moreno, Á., Urbanos-Garrido R.M., Thuissard-Vasallo I.J.: Public healthcare expenditure in Spain: measuring the impact of driving factors. Health Policy **111**, 34–42 (2013)
7. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**, P10008 (2008)
8. B´orner, K., Chen, C., Boyack, K.: Visualizing knowledge domains. Annu. Rev. Inf. Sci. Technol. **37**, 179–255 (2003)
9. Breyer, F., Felder, S.: Life expectancy and health care expenditures: a new calculation for Germany using the costs of dying. Health Policy **75**, 178–186 (2006)
10. Callon, M., Courtial, J.P., Turner, W.A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Soc. Sci. Inf. **22**, 191–235 (1983)
11. Callon, M., Courtial, J.P., Laville, F.: Co-word analysis as a tool for describing the network of interactions between basic and technological research. The case of polymer chemsitry. Scientometrics **22**, 155–205 (1991)
12. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualising the evolution of a research field. A practical application to the Fuzzy Sets Theory field. J Informetr. **5**, 146–166 (2011)
13. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: SciMAT. A new science mapping analysis software tool. J. Am. Soc. Inf. Sci. Technol. **63**, 1609–1630 (2012)
14. Cobo, M.J., Chiclana, F., Collop, A., de Ona, J., Herrera-Viedma, E.: A Bibliometric Analysis of the Intelligent Transportation Systems Research Based on Science Mapping. IEEE Trans. Intell. Transp. Syst. **15**, 901–908 (2014)
15. Crivelli, L., Filippini, M., Mosca, I.: Federalism and regional health care expenditures: an empirical analysis for the Swiss cantons. Health Econ. **15**, 535–41 (2006)
16. Garfield, E.: Scientography. Mapping the tracks of science. Curr. Contents Soc. Behav. Sci., **7**, 5–10 (1994)
17. Gerdtham, U.G., Søgaard, J., Andersson, F., Jönsson, B.: An econometric analysis of healthcare expenditure: a cross-sectional study of the OECD countries. J. Health Econ. **11**, 63–84 (1992)
18. Getzen, T.E.: Population aging and the growth of health expenditure. J. Gerontol. **47**, S98–S104 (1992)
19. Grant, M.J., Booth, A.: A typology of reviews: an analysis of 14 review types and associated methodologies. Health Inf. Libr. J. **26**, 91–108 (2009)
20. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. USA **101**, 5228–5235 (2004)
21. Hinton, G.E.: Learning distributed representations of concepts. In: Morris, R.G.M. (ed.), Parallel distributed processing: Implications for psychology and neurobiology, pp. 46–61. Clarendon Press, London (1989)
22. Hitiris, T., Posnett, J.: The determinants and effects of health expenditure in developed countries. J. Health Econ. **11**, 173-181 (1992)
23. Hitha, K.C., Kiran, V.K.: Topic Recognition and Correlation Analysis of Articles in Computer Science. In: 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 1115–1118. IEEE, Palladam (2021)
24. He, Q.: Knowledge discovery through co-word analysis. Libr. Trends **48**, 133–159 (1999)
25. Hu, K., Qi, K., Yang, S., Shen, S., Cheng, X., Wu, H., Zheng, J., McClure, S., Yu, T.: Identifying the "Ghost City" of domain topics in a keyword semantic space combining citations. Scientometrics **114**, 1141–1157 (2018)

26. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Xing, E.P, Jebara, T. (eds.) Proceedings of the 31st International Conference on International Conference on Machine Learning, pp. 1188–1196. JMLR.org (2014)
27. Lopreite, M., Mauro, M.: The effects of population ageing on health care expenditure. A Bayesian VAR analysis using data from Italy. Health Policy **121**, 663–674 (2017)
28. Lopreite, M., Puliga, M., Riccaboni, M., De Rosis, S.: A social network analysis of the organizations focusing on tuberculosis, malaria and pneumonia. Soc. Sci. Med. **278**, 113940 (2021)
29. Lopreite, M., Zhu, Z.: The Effects of Ageing Population on Health Expenditure and Economic Growth in China. A Bayesian-VAR Approach. Soc. Sci. Med., Volume **265**, 113513 (2020)
30. Maiya, A.S., Rolfe, R.M.: Topic similarity networks. Visual analytics for large document sets. In: 2014 IEEE International Conference on Big Data, pp. 364–372. IEEE, Washington, DC (2014)
31. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting Similarities among Languages for Machine Translation. arXiv (2013). Available via https://arxiv.org/abs/1309.4168
32. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Proceedings of the 26th International Conference on Neural Information Processing Systems **2**, pp. 3111–3119. Curran Associates Inc., Red Hook, NY (2013)
33. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. **6**, e1000097 (2009)
34. Noyons, E.C.M., van Raan, A.F.J.: Advanced mapping of science and technology. Scientometrics **41**, 61–67 (1998)
35. Peters, H., van Raan, A.F.J.: Structuring scientific activities by co-author analysis: An exercise on a university faculty level. Scientometrics **20**, 235–255 (1991)
36. Rehs, A.: A structural topic model approach to scientific reorientation of economics and chemistry after German reunification. Scientometrics **125**, 1229–1251 (2020)
37. Richardson, J., Robertson, I.: Ageing and the cost of health services. In: Policy implication of the aging of Australia's population, pp. 329–355. Productivity Commission: Melbourne Institute of Applied Economic and Social Research, Melbourne (1999)
38. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
39. Suominen, A., Toivanen, H.: Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. J. Am. Soc. Inf. Sci. Technol. **67**, 2464–2476 (2016)
40. Trevisani, M., Tuzzi, A.: Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories, Knowl. Based Syst. **146**, 129–141 (2018)
41. Wang, Z.: The determinants of health expenditures: evidence from US state-level data. Appl. Econ. **41**, 429–435 (2009)
42. White, D., McCain, K.: Visualising a discipline: An author co-citation analysis of information science, 1972–1995. J. Am. Soc. Inf. Sci. **49**, 327–355 (1998)
43. Zhang, Y., Lu, J., Liu, F., Liu, Q., Porteer, A., Chen, H., Zhang, G.: Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. J. Informetr. **12**, 1099–1117 (2018)

An automatic approach for bibliographical co-words networks labelling

Un approccio automatico per etichettare le co-words networks bibliografiche

Manuel J. Cobo and Maria Spano

Abstract Different measures have been proposed in the science mapping software tools to identify the most representative keywords for bibliographical co-words networks, identified by means of community detection procedures. However, the latter take into account only a single aspect, be it linked to the structure of the network rather than to the frequency of keywords in the bibliographical collection. In this work we propose an automatic approach for labeling the clusters derived from co-words networks, considering three different aspects (topological, quantitative and qualitative). In that way, our method aggregate these three measure into a global measure that indicates more robustly which is the most representative label for each topic.

Abstract *Nei software di science mapping sono state proposte diverse misure per individuare le keywords più rappresentative per le co-words networks, individuate con tecniche di community detection. Tuttavia, queste ultime tengono conto di un unico aspetto, sia esso legato alla struttura della rete piuttosto che alla frequenza delle keywords nella collection bibliografica. In questo lavoro proponiamo un approccio automatico per il labelling delle co-words networks, che non solo considera tre aspetti differenti (topologico, quantitativo e qualitativo), ma consente di sintetizzarli in una misura globale che indichi in maniera più robusta quale sia l'etichetta più rappresentativa per ciascun topic.*

Key words: sciencec mapping analysis, co-word networks, automatic labelling, community detection

Manuel J. Cobo
Deparment of Computer Science and Artificial Intelligence, University of Granada, Calle Periodista Daniel Saucedo Aranda s/n, E-18071, e-mail: mjcobo@decsai.ugr.es

Maria Spano
University of Naples Federico II, Corso Umberto I, 40, 80138 Napoli e-mail: maria.spano@unina.it

1 Introduction

With the increasing availability of scientific information through bibliographic databases, such as Web of Science, Scopus or Dimensions, researchers have great difficulties in analyzing and processing such information. Thus, the science of science [10] provides us with algorithms, methods and software tools from computer science, artificial intelligence, sociology, bibliometrics and statistics, in order to be able to process the huge amount of scientific information, extracting the underlying knowledge. In fact, in the last years, a great variety of bibliometric software tools are available [2, 7, 9, 13].

In particular, one of the main techniques employed is the science mapping analysis, which allows us to summarize large volumes of information in a map showing the social, intellectual and conceptual aspects of a scientific field. To do that, the corpus is converted into a bibliographical network [3, 8] where the nodes are the unit of analysis, and the edges represent a similarity or relation among them. In that sense, using authors as unit of analysis and co-occurrence as relation measure, a co-author network [11, 14] could be made. Similarly, using the set of keywords provided in the papers, a co-words network [5] could be built. After applying a community detection algorithm over the whole network, a set of clusters of sub-communities (i.e. a set of units of analysis strong related) are detected. Thus, if keywords have been the unit of analysis, the clusters will represent the themes covered in the research field.

In order to represent the clusters in a science map [8], a label or a name should be given for each one. That is, among the set of nodes inside the clusters, the most representative one should be selected as the representing node. But, this key node could be selected in different ways, and therefore, the existing science mapping software tools compute it in their own way, being based on a specific aspect of the clusters or corpus.

Thus, the main objective of this contribution is to develop a new method that allows labeling the clusters detected after the community discovery process, and that takes into account the structural aspects of the network, and the quantitative and quantitative aspects of the units of analysis within the global corpus. Moreover, to test the method, a case study was carried out using the dataset *management* included in the bibliometrix package [2].

2 Proposed method

The results of a community detection procedure are usually subgroups of strongly linked terms. To deeply investigate each community it is possible to plot it as a subgraph of the whole analysed network or to look at the list of terms that it includes. When the bibliographic collection is huge, the number of communities could increase and the analysis of single entity could be difficult. In this sense, plotting the

results on the thematic diagram allow us to obtain graphical representations that automatically summarise the main topics of a research field.

Therefore, the real issue is how to label the topics in a synthetic way on the diagram, choosing the most representative word for each of them. As we said above, there are different alternatives to label the topics, each of that consider a single aspect.

The main idea is that the labelling could be done based on different characteristics of the discovered communities, such us, topological (network structure), quantitative (keyword frequency), or qualitative (citations achieved).

Firstly, we take into the account the topological aspect of each single community by identifying the most central keyword. In graph theory, centrality is a very important concept in identifying important nodes in a network. It is used to measure the importance (or "centrality" as in how "central" a node is in the graph) of various nodes in a graph. Obviously, each node could be important with respect to how "importance" is defined. In literature, different centrality measures have been proposed (e.g. closeness centrality, betweenness centrality, eigen vector centrality) [1] that provide relevant analytical information about the network and its nodes. Here, we consider *Degree Centrality* as proposed by Callon [6] that is the most suitable centrality measure for being calculated on a relative small network as a community. Degree Centrality defines importance of a node in a network on the basis of its degree, where degree, in a non-directed graph, is the number of direct connections a node has with other nodes. Obviously, the higher the degree of a node, the more important it is in a network. Formally, we can define Degree Centrality of each keyword i as in 1

$$DC_i = \frac{\sum_j m(i,j)}{n-1} \tag{1}$$

where $m(i,j)$ is equal to 1 if there is a link between node i and j, and n is the number of vertices in the network.

Secondly, considering the quantitative aspect, we measure how much a keyword is present in the whole collection of documents. In the most simplest form it would be to calculate its frequency i.e. how many times a keyword appear in the collection. In this way, we assume that the higher is the frequency f_i of a keyword i, the more that keyword is important for defining the research field (or for describing the content of the collection). To this aim we calculate the frequency distribution of keywords in the whole collection.

Finally, regarding the qualitative aspect, we introduce a third measure, devoted to quantify if a keyword is used in the documents that have a major impact in the analysed research field. In bibliometrics, the impact of documents is usually measured in terms of achieved citations since their publication. To reflect this information on a generic keyword i, we calculate the total number of citations received by the documents in which that keyword appears TC_i.

Using the above described measures, we could compute a global measure to detect in an automatic way the community labelling using all of them. Starting from a bibliographic dataset, retrieved from indexing databases like Web of Science, Sco-

pus or Google Scholar [12], for each publication, a set of data concerning the document itself, the author(s) and the corresponding affiliation(s), as well as the references, are available. Among the different textual metadata reflecting the content of each publication, we focus our attention on authors' keywords (AK).

At the beginning of the algorithm we calculate the AK' frequency distribution on the whole collection, disregarding to which community each keyword will belong to. In this way, we obtain for each AK i how many times it appears as f_i. In the same way, we compute the third of the aforementioned measures TC_i, as the total number of citations achieved by the documents in which AK i appears. Then, we perform a community detection by using the Louvain algorithm [4], obtaining a set of communities C_k ($k = 1, \ldots, K$) that reflects the main topics of the analysed research field. For each community C_k, we have the list of words that it includes and its related sub-network. At this point, for all nodes/AK of the community we compute the topological measure as the degree centrality DC_i and to the latter we place the corresponding values of frequency f_i and the total citations TC_i side by side.

For each of the considered measures, we code their value in terms of *rank*. With that transformation numerical values are replaced by their rank when the data are sorted. For instance, if the observed frequencies of a set of words are described by the vector $(10, 12, 15, 17, 12)$, the ranks of words would be the corresponding vector $(4, 3, 2, 1, 3)$. The ranks are assigned to values in ascending order and as in standard competition ranking, words that have the same frequency value, as in the example, receive the same ranking number.

Finally, we compute the global rank, obtaining the most representative AK for each community, by counting how many times a AK is ranked as the first (rank equal to 1) for the three measures. If there in not a consensus among the measures (each of them identifies a different candidate for labelling the community) we look at the second position and, then, to the third and so on. The use of a ranking transformation allow us to compare metrics with different scales and ranges and to identify a single ordered list of candidates for labelling the communities.

3 A case study

In order to test the effectiveness of our proposal, we consider as a case study the dataset *management* included in the bibliometrix package [2]. The collection consists of 449 articles about the use of bibliometric approaches in business and management disciplines from 1985 to 2018, retrieved from *Web of Science*. By performing the community detection we identify 9 communities describing the main bibliometry topics adressed in the field of business and management.

Table 1 shows the results about some of the discovered communities, highlighting the values for the three considered measures, their transformation in rank and the resulting labelling of the proposed *global rank* (GR).

Table 1 Scores and rankings for AK candidates

Community	Word	f_i	DC_i	TC_i	r_{f_i}	r_{DC_i}	r_{TC_i}	GR
1	bibliometrics	122	0.863	4568	1	1	1	$\star$
1	citation analysis	35	0.425	1853	2	2	2	
1	scientometrics	16	0.247	228	3	3	20	
1	patent analysis	13	0.164	288	4	4	14	
1	knowledge management	8	0.151	504	5	7	6	
1	social network analysis	8	0.164	305	5	4	12	
...	...	...	...	...	...	...	...	
2	nanotechnology	18	0.500	668	1	1	1	$\star$
2	patents	8	0.375	378	2	2	3	
2	technology forecasting	8	0.250	77	2	6	7	
2	data mining	5	0.375	101	4	2	6	
2	productivity	5	0.313	578	4	5	2	
...	...	...	...	...	...	...	...	
3	lotka's law	4	1.000	22	1	1	3	$\star$
3	bibliometric distributions	3	1.000	20	2	1	4	
3	business ethics	3	0.714	47	2	6	1	
3	empirical regularity	3	1.000	20	2	1	4	
3	human resource management	3	0.571	24	2	7	2	
...	...	...	...	...	...	...	...	
4	bibliometric study	11	0.389	299	1	3	6	
4	research	10	0.444	3746	2	1	1	$\star$
4	citations	8	0.278	325	3	4	5	
4	co-word analysis	7	0.167	93	4	6	9	
4	impact	5	0.167	580	5	6	4	
...	...	...	...	...	...	...	...	

We noted that in some cases (e.g. Community 1 and 2) a complete consensus among the three measures is achieved and this is reflected on the GR. In the others (e.g. Community 3 and 4) the accordance between two of three metrics (r_{f_i} and r_{DC_i}, r_{DC_i} and r_{TC_i} respectively) is also sufficient to compute the GR. To present in a synthetic way all the results of our strategy in 1 the thematic diagram with the different AK candidates for labelling is shown.

4 Conclusions and future research

In this work we propose a strategy for automatic labelling of co-words networks. The idea of combining different metrics, devoted to consider aspects of the network together with the AK frequency distribution and their citations pattern, seems to produce promising results. Nevertheless, future work will be addressed to evaluate if the total citations is the most suitable measure to reveal the impact of AK in the analysed field of research and also if it could be necessary to weight the metrics

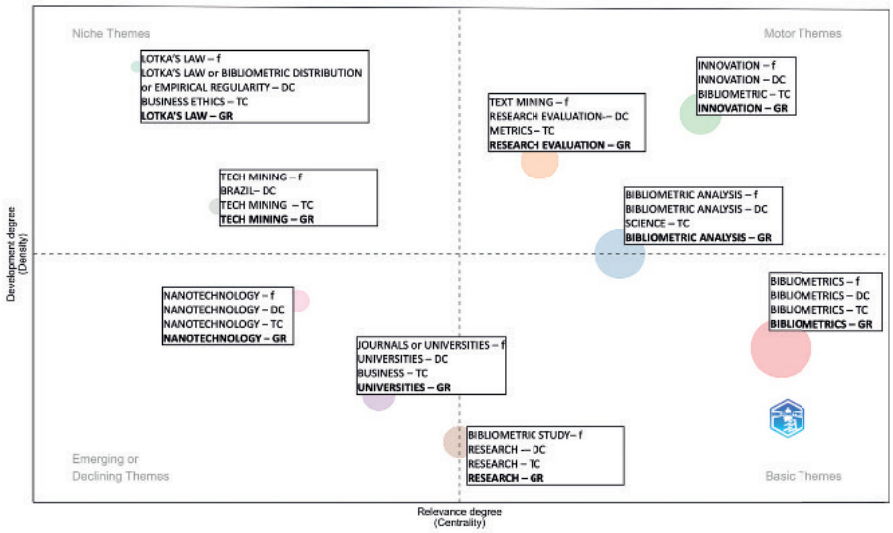


Fig. 1 Thematic diagram with the different AK candidates for labelling

in a different way to compute the global rank, giving different importance to the considered measures.

Acknowledgment

This work has been supported by the project PID2019-105381GA-I00 (iScience) funded by MCIN / AEI / 10.13039/501100011033.

References

1. Aggarwal, C.C.: Social network data analytics. Springer, Boston, (2011)
2. Aria, M., Cuccurullo, C.: Bibliometrix: An R-tool for comprehensive science mapping analysis. J. Informetr. **11**, 959–975 (2017)
3. Batagelj, V., Cerinšek, M.: On bibliographic networks. Scientometrics, **96:3**, 845–864 (2013)
4. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. P10008 (2008)
5. Callon, M., Courtial, J. P., Turner, W. A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Social Science Information, **22(2)**, 191–235 (1983)
6. Callon, M., Courtial, J.P., Laville, F.: Co-word analysis as a tool for describing the network of interactions between basic and technological research - The case of polymer chemistry. Scientometrics **22**, 155–205 (1991)

7. Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology, **62:7**, 1382–1402 (2011)
8. Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of Informetrics, **5:1**, 146–166 (2011)
9. Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology, **63:8**, 1609–1630 (2012)
10. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A.M, Radicchi, F., Sinatra, R., Uzzi, B. Vespognani, A., Waltman, L., Wang, D., Barabási, A.L.: Science of science. Science, **359:6379**, eaao0185 (2018)
11. Glänzel, W.: National characteristics in international scientific co-authorship relations. Scientometrics, **51:1**, 69–115 (2001).
12. Harzing, A.W., Alakangas, S.: Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. Scientometrics, **106**, 787–804 (2016)
13. Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., Cobo, M. J.: Software tools for conducting bibliometric analysis in science: An up-to-date review. El Profesional de La Información, **29:1**, (2020).
14. Peters, H. P. F., Van Raan, A. F. J.: Structuring scientific activities by co-author analysis. Scientometrics, **20:1**, 235–255. (1991).

Characterising Research Areas in the field of AI

Temi di ricerca caratterizzanti nel campo dell'IA

Alessandra Belfiore[1], Angelo Salatino[2], Francesco Osborne[3]

Abstract Interest in Artificial Intelligence (AI) continues to grow rapidly, hence it is crucial to support researchers and organisations in understanding where AI research is heading. In this study, we conducted a bibliometric analysis on 257K articles in AI, retrieved from OpenAlex. We identified the main conceptual themes by performing clustering analysis on the co-occurrence network of topics. Finally, we observed how such themes evolved over time. The results highlight the growing academic interest in research themes like deep learning, machine learning, and internet of things.

Abstract *L'interesse nell'intelligenza artificiale (AI) continua a crescere rapidamente, per questo è importante aiutare ricercatori e organizzazioni nel comprendere dove si sta dirigendo la ricerca in AI. In questo studio, abbiamo eseguito un'analisi bibliometrica su 275 mila articoli di ricerca in AI, scaricati da OpenAlex. Abbiamo identificato i principali temi concettuali eseguendo un'analisi dei gruppi sulla rete delle co-occorrenze dei topic. Infine, abbiamo osservato come questi temi si sviluppano nel tempo. I risultati mostrano un crescente interesse accademico nei temi di ricerca come deep learning, machine learning, e internet of things.*

Key words: Thematic evolution, Science of Science, Bibliometric Analysis, Scholarly Data, Topic Detection, Research Trends

1 Introduction

Interest in Artificial Intelligence (AI) continues to grow rapidly, hence it is crucial to support researchers and organisations with novel ways of exploring the scientific landscape as they can take informed decisions.

In this paper, we present a bibliometric analysis on the recent trends in AI. In particular, we initially downloaded 257K papers in the field of AI from OpenAlex, from the 1990 to February 2022, and we associated them with research topics in the Computer Science Ontology (CSO), the largest ontology of research topics in the field

[1] Alessandra Belfiore, Università della Campania Luigi Vanvitelli; email: alessandra.belfiore@unicampania.it

[2] Angelo Salatino, The Open University; email: angelo.salatino@open.ac.uk

[3] Francesco Osborne, The Open University; email: francesco.osborne@open.ac.uk

of Computer Science. Then, we organised all the documents in 7 periods based on the publishing year. In each time period, we first identified conceptual themes (i.e., clusters of topics) representing research areas and then we computed the Callon's indices of density and centrality. These indices allowed us to determine whether the themes are motor, niche, basic, and emerging (or declining). Finally, we mapped the similar themes across the different timeframes and analysed how they developed over time: e.g., before they started being niche and after became motor.

In this analysis, we identified eight themes experiencing a significant shift, which we explained with actual events happened in the field of Artificial Intelligence, such as the Deep Learning revolution and the emergence of IoT.

The remainder of the paper is organised as follows. In Section 2, we present our dataset and methodology. In Section 3, we present our results. Finally, Section 4 concludes the paper, outlining future directions.

2 Material and Methods

To perform this analysis, we first downloaded the research papers in the field of AI, from OpenAlex[1], a recently launched scholarly dataset. Then, we run the CSO Classifier on the papers metadata (title and abstracts) to extract their relevant research topics, and finally we run the thematic analysis to assess how the various themes evolved over time.

We used openalexR (Aria, (2022)) to retrieve all papers having "artificial intelligence", "machine learning", "deep learning" and "data science" either in titles or abstracts, published during the period 1990 to February 2022 inclusive, resulting in 257K research papers.

We extracted the relevant topics from all the research documents with the CSO Classifier[2] (Salatino et al., (2019)), a tool that takes in input the text of a research paper (title, abstract and keyword) and returns a selection of research topics drawn from the Computer Science Ontology (Salatino et al., (2018)).

After associating each document with its relevant research topics, we split the corpus in 7 timeframes. The first six timeframes are of 5 years each (1990-94, 1995-99, up to 2015-19), the last timeframe goes from 2020 to 2022. In each timeframe, we identified and characterised the different conceptual themes and then we observed how they evolved over time. For instance, we may want to detect when and whether a theme became highly relevant and well developed.

As a first step, in each timeframe, we created the topic co-occurrence network using the topics returned by the CSO Classifier. The topic co-occurrence network is a fully weighted graph describing the interaction between topics. In this graph, nodes (i.e., topics) are linked together by undirected arcs to describe the extent of their co-occurrence. The node weight represents the number of publications that a topic has

[1] OpenAlex - https://openalex.org

[2]The CSO Classifier - https://github.com/angelosalatino/cso-classifier

received in such timeframe, and the link weight is equal to the number of papers the two topics appeared together in the same period.

We applied the Louvain community detection algorithm (Blondel et al., (2008)) on these networks to extract clusters of topics (i.e., conceptual themes). For this, we leveraged Bibliometrix (Aria and Cuccurullo, (2017)), which is an open-source tool developed in R for quantitative research in scientometrics and bibliometrics. Specifically, as parameters we set 1000 topics, with the minimum cluster frequency set to 5.

We computed the Callon's centrality and density indices on the resulting clusters to respectively measure the relevance and the degree of development of the theme (Callon et al., (1991); Aria et al., (2020)).

Based on the values of both centrality and density, each cluster has been classified according to four themes: i) motor, ii) basic, iii) emerging or declining, and iv) niche (He, (1999); Cahlik, (2000)). The *motor themes* are highly relevant and well developed in research, as they have levels of centrality and density above average. The *basic themes* are low developed in research but relevant themes, displaying low levels of density and high levels of centrality. The *emerging or declining themes* have the lowest levels of density and centrality. This occurs in two moments of their life: when they either emerge or decline. The distinction between emerging or declining themes can be understood only by comparing their evolution over time. Finally, the *niche themes* are highly developed, but they are developed by a small niche of researchers, displaying density above average and centrality below average.

After determining the class of all the extracted conceptual themes in the 7 timeframes, we mapped the similar ones appearing in multiple timeframes. Two themes in consecutive time periods have been mapped if they had the same top-3 topics. We also mapped together two themes that differed of just one topic, but at the condition that the unmatched topic was among the top-5 topics of the other theme.

We used this mapping to analyse the most significant shifts in this space.

3 Results

This section presents the main results of our analysis on the evolution of conceptual themes in the considered seven periods.

The Bibliometrix tool extracted from the topic co-occurrence networks 6-9 conceptual themes (7.42 ± 0.97) in each period. We observed that some themes appeared just once or in two consecutive timeframes, whereas some others appeared in multiple time periods. To this end, we only mapped the clusters that recurred in three or more consecutive timeframes so to attain a better understanding of their life trajectory. Altogether, we identified eight recurring clusters portraying interesting dynamics. In Table 2, we report the eight themes, identified by their highly representative topic, and their classification over the seven time periods based on Callon's centrality and density.

It is worth pointing out that the four-themed classification and the life trajectory is contextualised to the whole cluster and not just its highly representative topic. The context surrounding the first conceptual theme is about expert systems, intelligent systems, and more in general symbolic AI which has experienced a steady decline in the past decades as an increasing number of researchers have shifted their focus to probabilistic AI. The second conceptual theme is related to machine learning and includes relevant research areas such as supervised machine learning and neural networks. According to the data, initially it was highly relevant for the community (motor), but in the following it lost some momentum, until its resurgence in recent years, also thanks to the availability of more powerful machines that can handle large machine learning models.

Table 1. Recurring conceptual themes identified in the seven temporal times. In the theme column, we report only the most representative topic.

Theme	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19	2020-22
expert systems	decline	decline	decline	decline	decline	decline	decline
machine learning	motor	basic	basic	decline	basic	basic	motor
reasoning	basic	motor	motor	-	-	-	-
data mining	-	-	niche	motor	motor	-	-
genetic algorithms	-	motor	motor	motor	decline	-	-
sensors	-	-	-	emerging	motor	motor	motor
robots	niche	basic	decline	motor	-	-	-
deep learning	-	-	-	-	niche	motor	motor

The third theme includes reasoning, multiagent systems, semantics, logic programming, and intelligent agents, outlining the multi-agent era. The theme went off the radar from the 2005 onward. This finding is confirmed by previous bibliometric analysis (Osborne, et al., (2014)) and discussed by some articles at that time. For instance, a 2007 editorial titled “Where are all the Intelligent Agents?” (Hendler, (2007)) suggested that the role of agent research in Semantic Web community was not as strong as envisaged in the original 2001 vision.

The fourth theme, mostly associated with data mining, shows the emergence of the big data era and the applications of knowledge discovery. We do not have data in the last two time periods due to the sensitivity of the algorithm in returning the relevant clusters. In the future, we plan to run a deeper analysis investigating a high number of clusters per period.

The fifth theme includes topics like genetic algorithms, adaptive algorithms, optimization, and optimization problems, which had their culmination in the decade 2000-10.

The sixth theme identified by sensors shows the life trajectory of the internet of things, emerged in 2005-09 and currently highly relevant and well developed.

The seventh theme shows the application of robots in control systems, including sensors, switching control, and process control, which starts being niche and goes through a decline up to 2004. From 2005 onwards there is a paradigm shift in which the theme of robots includes neural network architectures, making it highly relevant.

Finally, the eighth theme is centred around deep learning including pattern recognition, neural networks, convolutional neural networks, and deep belief network, which was niche in 2010, but from 2015 onwards it started being highly relevant and well developed.

4 Conclusions and Future Work

In this paper, we performed a bibliometric analysis in the domain of Artificial Intelligence, and we observed how conceptual themes evolved over time.

We identified eight conceptual themes experiencing significant shift, signalling how the AI field is highly dynamic, and we are confident this will continue in the future as the attention to AI is growing. Specifically, we observed the development over time of specific themes like Deep Learning, IoT, Robotics, Machine Learning, Genetic Algorithms and then we provided an explanation for such shift based on actual events happened in the field of AI.

The findings of this study have to be seen in light of some limitations. Both the thematic evolution and the conceptual themes can be affected by the selection of papers we retrieved from OpenAlex.

For the future, we plan to work on multiple fronts. First, we would like to increase the sensitivity of the Bibliometrix tool to return more than 9 clusters of topics in each timeframe. This would enable us to perform a more comprehensive analysis. Second, we plan to analyse the whole Computer Science, so to be able to characterise a larger pool of conceptual themes. Third, we plan to perform a qualitative analysis to understanding how the four-themed classification relates to the Khun's phases of scientific revolution (Kuhn, (1970)). Finally, we plan to analyse the patterns of Callon's centrality and density to understand whether they have predictive power to forecast how novel topics will develop in the upcoming years.

References

1. Aria, M. (2022). openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API. URL: https://github.com/massimoaria/ openalexR r package version 0.0.1.
2. Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics, 11, 959-975.
3. Aria, M., Misuraca, M., & Spano, M. (2020). Mapping the evolution of social research and data science on 30 years of Social Indicators Research. Social Indicators Research, 149, 803/831.

4. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008, P10008.
5. Cahlik, T. (2000). Comparison of the maps of science. Scientometrics, 49, 373-387.
6. Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics, 22, 155-205.
7. He, Q. (1999). Knowledge discovery through co-word analysis.
8. Hendler, J. (2007). Where are all the intelligent agents?. IEEE Intelligent Systems, 22, 2-3.
9. Kuhn, T. S. (1970). The structure of scientific revolutions (Vol. 111). University of Chicago Press: Chicago.
10. Osborne, F., Scavo, G., & Motta, E. (2014, November). A hybrid semantic approach to building dynamic maps of research communities. In International Conference on Knowledge Engineering and Knowledge Management (pp. 356-372). Springer, Cham.
11. Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019). The CSO classifier: Ontology-driven detection of research topics in scholarly articles. In International Conference on Theory and Practice of Digital Libraries (pp. 296-311). Springer, Cham.
12. Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018, October). The computer science ontology: a large-scale taxonomy of research areas. In International Semantic Web Conference (pp. 187-205). Springer, Cham.

Mapping evolutionary paths of a society: the longitudinal analysis of the Italian Economia Aziendale

Mappare i percorsi evolutivi di una società: l'analisi longitudinale dell'Economia Aziendale italiana

Corrado Cuccurullo, Luca D'Aniello and Michele Pizzo

Abstract
In the last decades, scientific literature production has been increasing rapidly in several research domains. Researchers are pushing to develop new methods allowing them to analyse the conceptual structure from this huge amount of information. Bibliometrics provides several methods for knowledge extraction. By means of science mapping techniques, namely the co-word network analysis and thematic maps, we analyse the longitudinal strategic positioning of all the Italian scholars affiliated to the discipline "Economia Aziendale", one of the five disciplinary groups of the Italian AIDEA (*Accademia Italiana di Economia Aziendale*). We considered their publications indexed on Scopus. This work aims to understand how some political and cultural measures impact on the research activity of a scientific community.
Abstract *Negli ultimi decenni, la produzione della letteratura scientifica è aumentata rapidamente in diversi domini di ricerca. I ricercatori spingono per sviluppare nuovi*

[1] Corrado Cuccurullo, Department of Economics, University of Campania Luigi Vanvitelli, Caserta, Italy; email: corrado.cuccurullo@unicampania.it

[2] Luca D'Aniello, Department of Social Sciences, University of Naples Federico II, Naples, Italy; email: luca.daniello@unina.it

[3] Michele Pizzo, Department of Economics, University of Campania Luigi Vanvitelli, Caserta, Italy; email: michele.pizzo@unicampania.it

metodi che permettano di analizzare le strutture concettuali di questa enorme quantità di informazioni. La bibliometria fornisce diversi metodi per estrarre conoscenza. Attraverso tecniche di science mapping, ovvero la co-word network analysis e le mappe tematiche, analizziamo il posizionamento strategico longitudinale di tutti gli studiosi italiani affiliati alla "Economia Aziendale", uno dei cinque gruppi disciplinari di AIDEA (Accademia Italiana di Economia Aziendale). Abbiamo considerato le loro pubblicazioni indicizzate su Scopus. Questo lavoro si propone di capire come alcune misure politiche e culturali impattino sull'attività di ricerca di una comunità scientifica.

Key words: economia aziendale, bibliometric analysis, science mapping, knowledge synthesis, thematic map.

1 Introduction

Scientific communities are well-known concepts in the study of science. Some of them are country-based. Studying the development of scientific knowledge in a society provides insight into the structures and dynamics of knowledge production. However, it is true that there are still only a limited number of studies that examine the research front from a community perspective. This work fills the gap.

Bibliometrics introduces transparent and reproducible methods to run a *science mapping analysis* to trace the research front dynamics (Cuccurullo et al., 2016). It carries out a conceptual structure of the extant research activity by synthesizing past research findings, detecting trends and gaps, and identifying the main centres of interest. This process of knowledge extraction is called *science mapping* (Zaho, (2010)).

In this work, we use a science mapping approach to identify and display themes and trends with a synchronic (Callon et al., 1983) and diachronic perspective (Cobo et al., 2011). Through science mapping techniques, namely the co-word network analysis and thematic maps, we analyse the longitudinal strategic positioning of all the Italian scholars affiliated to the discipline "Economia Aziendale" (Alexander et al., 2011; Coronella et al., 2018; Capalbo et al., 2008; Galassi, 2011; Lai et al, 2015; Viganò et al, 2007), one of the five disciplinary groups of the Italian AIDEA (*Accademia Italiana di Economia Aziendale*). It is the scientific Society of Academic Scholars of Accounting, Business Administration, Public Administration, Management, Governance, Organizational studies, Banking, and Finance.

Our study contributes to the sociology of science, providing useful elements for understanding community paths and contingent factors that impact them. Furthermore, our study can be useful in a policy perspective to understand how some political and cultural measures impact on the research activity of a scientific community.

2 Methods

2.1 Data collection

Italian scholars of Economia Aziendale are currently 815. Using their first name and last name (from the list of Ministry of Research and Universities), we retrieved their ID Scopus through *rscopus,* a R package that provides Elsevier API to query Scopus bibliographic database about authors' research production. We identify 657 scholars with an ID Scopus. These IDs were used to massively download the publications of each scholar. We limited our search just to English articles and reviews published in ANVUR Management Journals (source https://www.anvur.it/attivita/vqr/vqr-2015-2019/gev/area-13b-scienze-economico-aziendali/) from January 2000 to December 2021. ANVUR is the Italian Agency for Academic Research Appraisal.

We loaded the data in R and converted it into a data frame using *bibliometrix*, an open-source tool for quantitative research in scientometrics and bibliometrics that includes all the main methods for performance analysis and science mapping (Aria and Cuccurullo, (2017)). The data frame nxp was composed of n=2348 observations and p=48 variables. Each row is a publication of the whole collection, and each variable represents a meta-data, i.e., information about the record (e.g., the title, abstract, keywords, authors name). Then, we classified the publication to a specific Italian geographical area by considering the affiliation of AIDEA researchers associated with the reference work.

To highlight the main themes of the collection and evaluate their evolution over time, we divided our timespan (2000–2021) into two equal time slices: 2000-2010, and 2011-2021.

2.2 Data analysis

We map the longitudinal conceptual structure of Economia Aziendale through (i) term co-occurrence network analysis and (ii) thematic map. The combined use of these methods allows illustrating how terms are linked to each other, to highlight the main research themes and their evolution. We grouped scholars on the basis of their affiliation to Universities located in the following geographical areas: North-West, North-East, Centre, and South/Islands.

The term co-occurrence network analysis (Wang et al., 2019) relates to a set of terms (e.g. keywords, terms extracted from titles, or abstracts) that identify a specific

research field or topic. Network representation aims to understand the themes covered by a research field. It allows the detection of topics that are the most important and the most recent research fronts.

Following the network approach, we compute a term co-occurrence matrix. Each cell outside the principal diagonal counts the co-occurrences, i.e., the number of times that two terms appear together in the articles. Then, the association index as proposed by Van Eck and Waltman (2009) was used to normalize the co-occurrences terms matrix. This measure assumes values in the interval [0,1] and reflects the association strength among terms. Co-occurrence matrices can be seen as undirected weighted graphs; therefore, we can build a network in which each term is a node and the association between linked terms is expressed as an edge, visualizing both single terms and subsets of terms frequently co-occurring together. To detect subgroups of strongly linked terms, where each subgroup identifies a center of interest or a topic extracted from the analyzed collection, we run the Louvain algorithm (Blondel et al., 2008), a community detection algorithm (Fortunato, 2010).

Themes, identified through community detection, were plotted on Strategic or Thematic map (Cobo et al., 2011), a bi-dimensional matrix where axes are functions of the Callon centrality and density, respectively (Callon et al., 1983). Centrality can be read as the importance of the theme in the research field, while density can be read as a measure of the theme's development.

Computing the co-occurrence network and then the thematic map, we carried out the conceptual structure of each Italian geographical area research activity in two reference's timespans. Then, centrality and density values were standardized to compare the research fronts of the different geographical areas by plotting themes on a joint map (Cuccurullo et al, (2021); Cuccurullo et al., (2022); Aria et al. (2022)). The output was a strategic map. It allows defining four typologies of themes (Cahlik, 2000) according to the quadrant in which they are placed. In the upper-right quadrant, there are motor themes, characterized by both high centrality and density. This means that they are both developed and important for the research field. In the upper-left quadrant, there are isolated themes, also called niche themes, characterized by high density and low centrality values. They have well-developed internal links but unimportant external links and so are of only limited importance for the field. In the lower-left quadrant, there are emerging or declining themes. They have both low centrality and density values meaning that are weakly topics developed or marginal ones. There are basic and transversal themes in the lower-right quadrant, characterized by high centrality and low-density values. These themes are important for a research field and concern general topics transversal to the different research areas of the scientific field.

In each temporal interval, we carry out the strategic maps using the Authors' Keywords (DE) as units of analysis.

3 Main findings and conclusions

Figures 1 and 2 show the thematic evolution. Each topic identified with the community detection algorithm is plotted on bi-dimensional maps and labeled with the corresponding most frequent authors' keywords.

In the first time slice (2000 – 2010), topics distributed on the left side of the map are mainly focused on accounting and governance. Topics of scholars affiliated with the universities of Central Italy are all motors and basic themes (first and fourth quadrants). They are mainly focused on *intellectual capital, management accounting, governance, internal auditing, outsourcing, and Italian listed companies*.

Almost all the topics of North-West Italian universities are niche and emerging or declining themes (second and third quadrants). Some topics are *social responsibility, organization identity,* and *focused hospital*. Other terms, such as *performance measurement systems, internazionalization,* and *pharmaeconomics,* show the scholars' attention on emerging topics, suggesting new and innovative research paths.

From the first to the second period, there is an evident growth in the number of topics. Moreover, in the second time slice (2011-2021), we note that main topics on the right side of the map concern accountability and performance and the themes of the North-West Italian universities move to these quadrants. Health issues - labeled with *health technology assessment* and *health-related quality of life* keywords – remain important for the Italian community. This feature concerns also the North-East Universities, that focus their attention on the emerging topic related to *healthcare organizations*. Still in this second time slice, the research activity of South/Islands Universities becomes transversal and focus mainly on *innovation, corporate social responsibility,* and *intellectual capital*. Finally, in the upper-left quadrant, niche themes have increased over time for all Italian universities, meaning a compact movement towards more and more specialized studies from the first to second period.

Our study fits in the stream of sociology of science, that especially deal with the social structures and processes of scientific activity. Universities in some geographical areas following new research trajectories (emerging topics), while others are mainly positioned on mainstream topics.

Some points of contingency that explain the Economia Aziendale scientific field development fall into the political (University reform in 2010) and cultural arenas (research internalization and research appraisal).

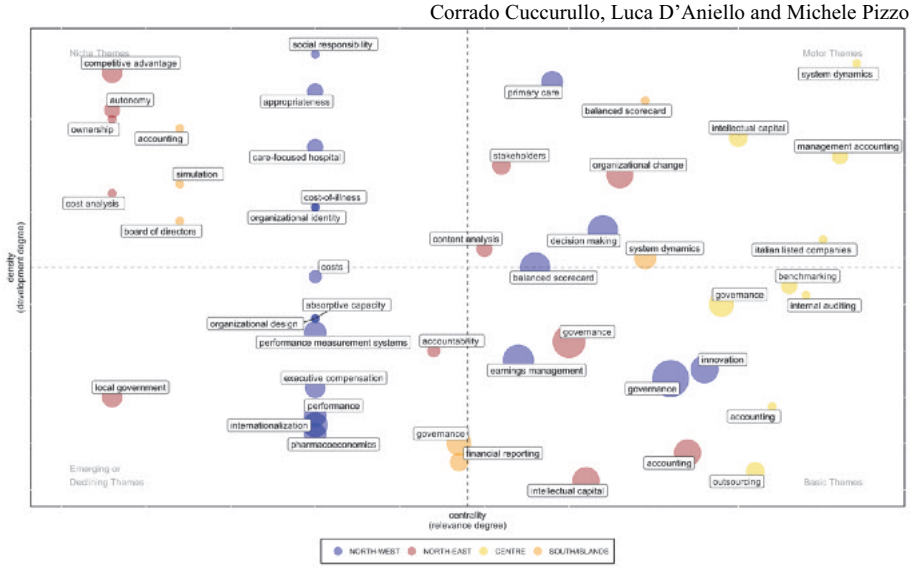


Figure 1: Thematic map. Focus on 2000-2010 timespan

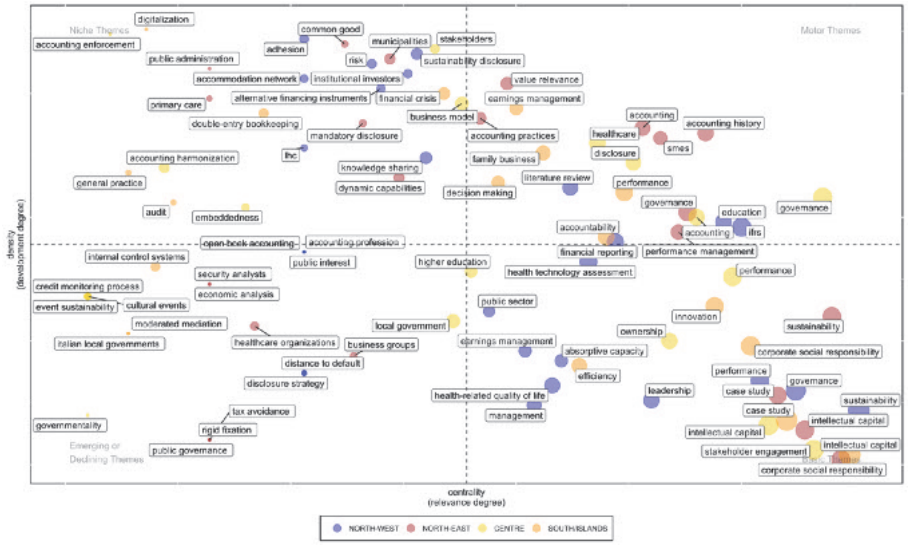


Figure 2: Thematic map. Focus on 2011-2021 timespan

References

1. Alexander, D., Servalli, S.: Economia Aziendale and financial valuations in Italy: Some contradictions and insights. Account. Hist., 16(3), 291-312, https://doi.org/10.1177/1032373211407052, (2011)
2. Aria M., Cuccurullo C., D'Aniello L., Misuraca M, Spano M.: Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy. Sustain.; 14(6):3643. https://doi.org/10.3390/su14063643, (2022)
3. Aria, M., Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Informetr.* 11 (4), pp. 959-975, (2017)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech-Theory E., URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008, (2008)
5. Cahlik, T.: Comparison of the maps of science. Scientometr. 49 (3), pp. 373-387, (2000)
6. Callon, M., Courtial, J.P., Turner, W. A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Soc. Sc. Infor. 22 (2), pp. 191-235, (1983)
7. Capalbo, F., & Clarke, F. The Italian economia aziendale and chambers' CoCoA. Abacus, 42(1), 66-86; https://doi.org/10.1111/j.1467-6281.2006.00191.x. (2006)
8. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. J. Informetr. 5 (1), pp. 146–166, (2011)
9. Coronella, S., Caputo, F., Leopizzi, R., Venturelli, A.: Corporate social responsibility in Economia Aziendale scholars' theories: A taxonomic perspective, Meditari Account. Res., Vol. 26 No. 4, pp. 640-656. https://doi.org/10.1108/MEDAR-03-2017-0124. (2018)
10. Cuccurullo, C., Aria, M., Sarto, F.: Foundations and trends in performance management. A twentyfive years bibliometric analysis in business and public administration domains. Scientometr. 108 (2), pp. 595-611, (2016)
11. Cuccurullo, C., D'Aniello, L., Aria, M., Spano, M.: Thematic evolution of Academic Medical Centers' research: a focus on Italian public owned AOUs in metropolitan areas. In: 10th International Conference IES 2022 (pp. 67-72). PKE-Professional Knowledge Empowerment srl. ISBN 978-88-94593-35-8 (2022)
12. Cuccurullo, C., D'Aniello, L., Spano, M.: Thematic atlas of Italian oncological research: the analysis of public IRCCS. In: ASA 2021 Statistics and Information Systems for Policy Evaluation: Book of short papers of the opening conference 127, pp. 97-103. Firenze University Press (2021)
13. Eck, N.J.V., Waltman, L.: How to normalize cooccurrence data? An analysis of some well-known similarity measures. J. Am. Soc. Inf. Sci. Tec. 60 (8), pp. 1635-1651, (2009)
14. Fortunato, S.: Community detection in graphs. Phys. Rep. 486 (3), pp. 75-174, (2010)
15. Galassi, G.: Is Economia Aziendale Research Programme 'fit for purpose'? A commentary on 'Contextualizing the intermediate financial accounting courses in the global financial crisis'. Acc. Educ., 20(5), 505-509, https://doi.org/10.1080/09639284.2011.614432. (2011)
16. Lai, A., Lionzo, A., Stacchezzini, R.: The interplay of knowledge innovation and academic power: Lessons from "isolation" in twentieth-century Italian accounting studies. Account. Hist., 20(3), 266-287. https://doi.org/10.1177/1032373215595138. (2015)
17. Viganò, E., Mattessich, R.: Accounting research in Italy: second half of the 20th century, Rev. of Account. and Financ., Vol. 6 No. 1, pp. 24-41. https://doi.org/10.1108/14757700710725449. (2007)
18. Wang, H., Zhao, Y., Dang, B., Han, P., Shi, X.: Network centrality and innovation performance: the role of formal and informal institutions in emerging economies, J. Bus. Ind. Mark. 34 (6), pp. 1388-1400, (2019)
19. Zhao, D.: Characteristics and impact of grant-funded research: a case study of the library and information science field. Scientometr. 84 (2), pp. 293-306. DOI= 10.1007/s11192-010-0191-y, (2010)

Modelling complex structures in ecological data

New insights on the ecology and conservation of Mediterranean sharks through the development of Citizen Science networks and new modeling approaches

Nuove conoscenze sull'ecologia e la conservazione degli squali del Mediterraneo attraverso lo sviluppo di reti di Citizen Science e nuovi approcci modellistici

Stefano Moro[1], Francesco Ferretti[2], Francesco Colloca[3]

Abstract Rare Species can be highly complex to model and study from ecological and statistical perspectives. In what follows, we illustrate several years of work trying to increase the knowledge about sharks in the Mediterranean. Our work connected complex statistical and ecological tools, including citizen science observation. Our data are presence-only data and require appropriate tools to be analyzed, such as spatial point processes.

Abstract *Le specie rare possono essere molto complesse da modellare e studiare da un punto di vista ecologico e statistico. Di seguito illustriamo diversi anni di lavoro in cui le conoscenze sugli squali nel Mediterraneo sono state migliorate.. Il nostro lavoro ha collegato strumenti statistici ed ecologici complessi, incluse le osservazioni provenienti da applicativi di citizen science.. I nostri dati sono dati di sola presenza e richiedono strumenti adeguati per essere analizzati, come i processi di punto spaziali.*

Key words: spatio-temporal patterns, Mediterranean elasmobranchs, point process, citizen science, opportunistic data.

1 Motivation and summary

One of the main challenges in conservation studies is the development of effective strategies to mitigate as much as possible the progressive biodiversity loss we are experiencing in our times (Cardinale *et al.* 2012). This goal is particularly harsh to achieve in marine environments where Global Change (i.e., over-exploitation,

[1] Stefano Moro, DBA "Sapienza" university of Rome, Dept. of Integrated Marine Ecology, Stazione Zoologica Anton Dohrn; email: stefano.moro@uniroma1.it,
[2] Francesco Ferretti, Dept. of Fish and Wildlife Conservation, Virginia Tech University; email: ferretti@vt.edu
[3] Francesco Colloca, Dept. of Integrated Marine Ecology, Stazione Zoologica Anton Dohrn, email: francesco.colloca@szn.it

habitat loss, climate change, and pollution) profoundly impacts marine communities' structure (Halpern *et al.* 2008). These drivers of change are particularly strong on apex predators, given their trophic role, with not fully understood effects on marine ecosystems (Hammerschlag *et al.* 2019). Sharks are often at the top of trophic chains. Furthermore, because of their life-history traits (i.e., long life span, late maturity, low fecundity), which are characteristics of low productive species, they are less capable of undergoing a high level of human pressure (Dulvy *et al.* 2021). This situation is reflected by their poor global conservation status, with more than one-third of the species threatened with extinction (Dulvy *et al.* 2021). These numbers make chondrichthyans the second most threatened Class of vertebrates after amphibians (Díaz *et al.* 2019), with extinction rates comparable with terrestrial vertebrates (Dulvy *et al.* 2021). Overfishing has been identified as a major threat to all threatened species and it might have caused the extinction of at least three species (Dulvy *et al.* 2021). These will represent the first cases of extinction due to overfishing in global marine fishes (Dulvy *et al.* 2021).

In this context, the Mediterranean Sea can be considered a major hotspot of biodiversity loss, due to its millenarian history of human exploitation, its high human population density (EEA 2015), and also the actual level of fishing exploitation, which is one of the highest in the world (Kroodsma *et al.* 2018). These conditions brought the worst elasmobranchs' conservation status worldwide, with more than half of the species threatened with extinction and 13 species still considered data deficient (Dulvy *et al.* 2016). However, the most severe wake-up call is that no improvement in elasmobranchs conservation was seen since the first assessment in 2007 (Walls & Dulvy 2021). Mediterranean elasmobranchs in general and pelagic species more in particular are affected by an inbuilt lack of ecological knowledge, caused by the sparse and tenuous nature of the data collected (Cashion *et al.* 2019). In addition, even if long-term time series of fishing surveys exist, they are rarely analyzed at a basin-scale to identify large-scale species-specific patterns (Follesa *et al.* 2019). It results in an overall data paucity, which raises the uncertainty around the regional conservation assessments and reduces the effectiveness of the conservation measures (Moro *et al.* 2020).

In the recent years, citizen science initiatives (CS) have become increasingly frequent in conservation studies (Follett & Strezov 2015), and opportunistic data can represent a valuable alternative source of information when more conventional data are scarce or insubstantial (McPherson & Myers 2009). Given that they are not collected with systematic surveys, they come with different biases that are difficult to handle, such as the spatio-temporal variation of the observation effort associated with their collection (Dickinson *et al.* 2010).

Here we provide new approaches that can handle opportunist data to estimate standardized abundance and distribution patterns. Several proxies of the observation effort had been tested, from human population size to other species opportunistic sightings and AIS (Automatic Identification System) data to track human activities at sea.

Choosing approaches that respect the nature of the analyzed data when estimating distribution models with opportunistic records is one of the main points stressed in this Thesis. In this sense, looking at the presence/absence events from a point-

perspective is the most rigorous way to analyze them since presence-only data represent a partial realization of this process. Consequently, Point Process Models are the natural way to treat presence-only data (Renner *et al.* 2015). Considering a point perspective also allows for performing an Integrated Distribution Modeling (IDM), including systematic and opportunistic data sources (Martino *et al.* 2021). The use of a location-dependent thinned Poisson Process can characterize each data source with a different detection function that describes the observation process generating the data. This approach has been tested in the Thesis with a case study concerning marine mammals, but it will be implemented in the future for the Mediterranean white shark. One of the primary outreaches of this work is that the approaches presented can be potentially applied to any data-poor and highly endangered taxa both in marine and terrestrial ecosystems.

Our results increased the ecological knowledge of the elasmobranch presence in the Mediterranean Sea. The analysis spanned many ecological features, from spatio-temporal patterns of abundance to strongholds of presence for highly endangered species. This information is pivotal to better characterize the elasmobranchs' condition in the Mediterranean Sea and carry out conservation plans based on quantitative results more than perceived patterns. Finally, the results obtained informed fieldwork activities targeting highly endangered Mediterranean sharks. They allowed the collection of high-quality ecological data, proving that opportunistic data can provide baselines to enhance the shark research in the Mediterranean Sea and potentially worldwide.

References

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., *et al.* (2012). Biodiversity loss and its impact on humanity. *Nature*, 486, 59–67.

Cashion, M.S., Bailly, N. & Pauly, D. (2019). Official catch data underrepresent shark and ray taxa caught in Mediterranean and Black Sea fisheries. *Mar. Policy*, 105, 1–9.

Díaz, S.M., Settele, J., Brondízio, E., Ngo, H., Guèze, M., Agard, J., *et al.* (2019). The global assessment report on biodiversity and ecosystem services: Summary for policy makers.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.*, 41, 149–172.

Dulvy, N.K., Allen, D.J., Ralph, G.M. & Walls, R.H.L. (2016). The conservation status of sharks, rays and chimaeras in the Mediterranean Sea [Brochure]. *IUCN Malaga Spain*.

Dulvy, N.K., Pacoureau, N., Rigby, C.L., Pollom, R.A., Jabado, R.W., Ebert, D.A., *et al.* (2021). Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. *Curr. Biol.*

EEA. (2015). *SOER 2015 — The European environment — state and outlook 2015*. European Environmental Agency.

Follesa, M.C., Marongiu, M.F., Zupa, W., Bellodi, A., Cau, A., Cannas, R., *et al.* (2019). Spatial variability of Chondrichthyes in the northern Mediterranean. *Sci. Mar.*, 83, 81–100.

Follett, R. & Strezov, V. (2015). An analysis of citizen science based research: usage and publication patterns. *PloS One*, 10, e0143687.

Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., *et al.* (2008). A global map of human impact on marine ecosystems. *science*, 319, 948–952.

Hammerschlag, N., Schmitz, O.J., Flecker, A.S., Lafferty, K.D., Sih, A., Atwood, T.B., *et al.* (2019). Ecosystem function and services of aquatic predators in the Anthropocene. *Trends Ecol. Evol.*, 34, 369–383.

Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., *et al.* (2018). Tracking the global footprint of fisheries. *Science*, 359, 904–908.
Martino, S., Pace, D.S., Moro, S., Casoli, E., Ventura, D., Frachea, A., *et al.* (2021). Integration of presence-only data from several sources. A case study on dolphins' spatial distribution. *Ecography*, 44, 1533–1543.
McPherson, J.M. & Myers, R.A. (2009). How to infer population trends in sparse data: examples with opportunistic sighting records for great white sharks. *Divers. Distrib.*, 15, 880–890.
Moro, S., Jona-Lasinio, G., Block, B., Micheli, F., De Leo, G., Serena, F., *et al.* (2020). Abundance and distribution of the white shark in the Mediterranean Sea. *Fish Fish.*, 21, 338–349.
Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., *et al.* (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.*, 6, 366–379.
Walls, R.H. & Dulvy, N.K. (2021). Tracking the rising extinction risk of sharks and rays in the Northeast Atlantic Ocean and Mediterranean Sea. *Sci. Rep.*, 11, 1–15.

An overdispersed Poisson model for forest fires occurrences in Southern Italian municipalities

Un modello di Poisson sovradisperso per il numero di incendi nei comuni del sud Italia

Crescenza Calculli and Serena Arima

Abstract In recent years, the number and the magnitude of wildfires are constantly growing in southern EU countries due to extreme climate conditions. This study proposes a modeling approach to investigate the relation between fire occurrence and several potential socio-economic and environmental driven factors considering two neighboring regions in southern Italy (Apulia and Basilicata). Multiple sources of data with different spatial support are used and data were preprocessed in order to reconduct the analysis to the municipality scale. A Bayesian zero-inflated Poisson model with spatial component is proposed to accommodate the excess of zeros in the counts and to account for the neighboring structure between municipalities. Preliminary results suggest the appropriateness of such approach with some insights explaining the dependence relation.

Abstract *Negli ultimi anni, il numero e l'entità degli incendi nei paesi meridionali dell'EU sono in costante crescita a causa delle condizioni climatiche estreme. Questo lavoro propone un approccio modellistico per indagare la relazione tra il numero di incendi e i potenziali fattori socio-economici e ambientali influenti in due regioni limitrofe del sud Italia (Puglia e Basilicata). Diverse fonti di dati caratterizzate da supporti spaziali differenti vengono utilizzate nello studio; i dati considerati sono stati trattati per ricondurre l'analisi alla scala comunale. Viene proposto un modello bayesiano di Poisson con componente spaziale per tener conto sia dell'eccesso di zeri nei conteggi che della struttura di vicinato delle aree. I risultati preliminari mostrano la bontà dell'approccio insieme ad alcune considerazioni riguardanti la relazione di dipendenza.*

Key words: wildfires, poisson models, zero-inflation, satellite data

Crescenza Calculli
Department of Economics and Finance, University of Bari Aldo Moro, Largo Abbazia S. Scolastica - 70124 Bari, Italy e-mail: crescenza.calculli@uniba.it

Serena Arima
Department of History, Society and Human Studies, University of Salento, Piazza Tancredi, n.7 - 73100 Lecce, Italy e-mail: serena.arima@unisalento.it

1 Introduction

In the last decades, the Mediterranean region has become a wildfires *hot spot* [6]. The current trend of climate change, with prolonged drought seasons and severe heatwaves, exposes a vast territory of southern Europe to an increasing wildfire risk. Many areas of Spain, Italy and Greece currently experience, especially in the summer seasons, large- and small-scale wildfires with huge social, economic and environmental costs [5]. Although fire, as a natural process, plays a role in some ecosystems (being a biomass controller and used as a management tool for pastoral and agricultural activities), the frequent occurrence of extreme and severe events affects the ecological stability of extensive areas, neglecting the capability of ecosystems to naturally recover. Furthermore recurrent events also impact negatively on air and water quality, biodiversity, soil, landscape aesthetics and threaten human lives in populated areas. Human-induced fires represent the majority of the total number of wildfires occuring in EU Mediterranean regions every year ($> 85\%$). In most cases, the causes of human induced wildfires can be accidental, intentional (arson), derived from acts of negligence or remain unknown. Thus the study of the main driving factors of ignitions is an essential step towards effective prevention and controlling policies. The socio-economic context (e.g. population trends, deprivation), the agricultural activity, the land use and the topographic and climate characteristics of fire-prone regions can be used to model the *fire occurrence* at the areal-level scale. The local analysis requires a considerable effort in terms of data integration. Different spatial supports are involved to use fire information, specifically satellite images provided by EFFIS (European Forest Fire Information system), together with official statistics concerning socio-economic driving factors provided by the Italian National Central Statistics Institute (ISTAT). Georeferenced fire data are used in this context to obtain local information about the number and magnitude of wildfire events only officially provided aggregated at regional scale.

Poisson regression models are commonly used for counts of rare events and have been proposed in prediction of fire occurrences [1]. Nevertheless, for many real-world phenomena, simple Poisson distribution is oftentimes inappropriate to model data that exhibit overdispersion and excess of zeros. In such a case, a modified version of traditional model's probability distribution, known as the zero-inflated Poisson (ZIP) distribution, results in better fittings [2]. In this work, the fire counts with extra zeros are modeled by means of a Bayesian zero-inflated Poisson regression model (Section 2) that allows to handle the excess of zeros and to take into account spatial dependence between municipalities. The spatial component can be either specified as unstructured or structured in order to explain the fires spatial dynamics. The rest of the work presents the case study and the used data (Section 3) and some preliminary results with some considerations about further developments (Section 4).

2 The model

For $i = 1, \ldots, n$, let Y_i denote the count outcome for area i, taking a non-negative integer value, and let X_i denote the associated covariate vector. We assume that conditional on X_i, the response Y_i is sampled from two sources with certain probabilities, either from the "Poisson" group where the measurements follow a Poisson distribution or from the "zero-excess" group where the measurements are zero. More formally, let E_i be a latent indicator variable showing from which sources Y_i is sampled, so that when $E_i == 1$ then $Y_i \geq 0$ and when $E_i == 0$ then $Y_i = 0$. Let $\theta_i = P(E_i = 1|X_i)$ represent the conditional probability of a sampling area i from the Poisson group and let $\lambda_i = E(Y_i|E_i = 1, X_i)$ denote the conditional mean of Y_i, given being sampled from the zero-excess group and covariates X_i, where $\theta \in [0,1]$ and $\lambda_i > 0$. Then the distribution of Y_i is given by

$$\begin{array}{ll} Y_i \sim P_0 & \text{with probability } 1-\theta_i \\ Y_i \sim Poisson(\lambda_i) & \text{with probability } \theta_i \end{array}$$

where P_0 represents the degenerate distribution for a random variable whose value is always 0. Consequently, given the covariates, the conditional probability mass function for Y_i is:

$$\begin{aligned} P(Y_i = 0|X_i) &= (1-\theta_i) + \theta_i e^{-\lambda_i} \\ P(Y_i = y_i|X_i) &= \theta_i \frac{e^{\lambda_i} \lambda^{y_i}}{y_i!} \end{aligned}$$

where $log(\lambda_i) = \beta_0 + \beta^T X_i$ and $logit(\theta_i) = \gamma_0 + \gamma^T X_i$.
Relying on the Bayesian approach, the model can be written as a hierarchical model as follows

$$P(Y_i = 0|X_i) = (1-\theta_i) + \theta_i e^{-\lambda_i} \tag{1}$$

$$P(Y_i = y_i|X_i) = \theta_i \frac{e^{\lambda_i} \lambda^{y_i}}{y_i!} \tag{2}$$

$$log(\lambda_i) = \beta_0 + \beta^T X_i + u_i \tag{3}$$

$$logit(\theta_i) = \gamma_0 + \gamma^T X_i \tag{4}$$

where $\beta_0, \gamma_0 \sim N(0, 10^2)$ and $\beta, \gamma \overset{\text{i.i.d.}}{\sim} N(0, 10)$. For the random effect u_i, describing the extra within-area variability, we specify a standard normal prior distribution. Notice that the set of explanatory variables in (3) and (4) might also be different according to the available information.
Posterior distributions cannot be obtained in closed form but the Markov Chain Monte Carlo (MCMC) algorithm is necessary in order to obtain samples from the joint posterior distribution. The aforementioned model is implemented in `R` package `NIMBLE` [3] in which an *ad-hoc* function specifying the zero-inflated Poisson distribution is coded. We allow the sampler for 15000 iterations with burn-in 5000

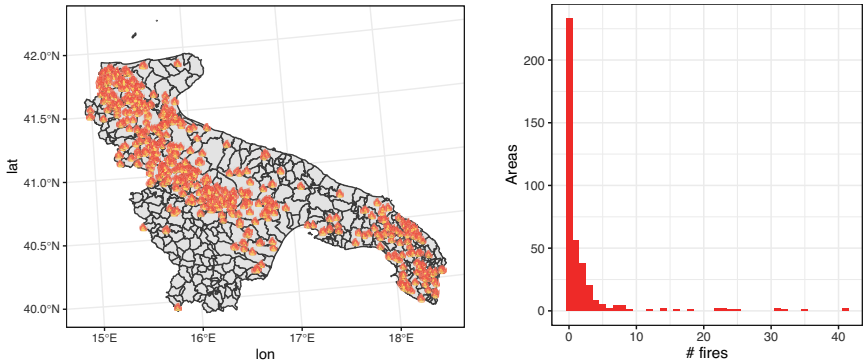


Fig. 1 Areas affected by fires (left) and fire counts (right) for Apulia and Basilicata regions during summer season 2021

and a thinning rate equal to 10. Chain convergences have been inspected visually through the visualization of trace plots and diagnostics measures (i.e. autocorrelation function, Geweke plot, Gelman and Rubin statistics)[1].

3 The case-study

During the summer season, wildfires rage every year across the Italian territory. In 2021, Italy ranked second among countries with the largest burnt area in Europe, doubling the average of 248,000 ha between 2008-2020 with a new record of 464,000 ha of burnt surface (Copernicus Atmosphere Monitoring Service, CAMS). The areas most affected by fires are the southern regions such as Sardinia, Sicily, Apulia and Basilicata. These latter regions present a high fire risk due to the large presence of Mediterranean scrub, wooded areas, intensive agricultural and pasture activities. Furthermore, the territory is characterized by several rural municipalities with low population densities and regressive demographic dynamics. A total number of 388 municipalities are present in the two neighbouring regions. For the summer season of 2021 (01 June - 31 August 2021), the number of fires for each municipality is obtained using remote sensing and satellite imagery. In particular, we retrieved data from MODIS Burned Area Product (MCD45) provided by CAMS. This product allows to detect the location of active fires using burned area algorithms (based on surface temperature anomalies and reflection by sunlight) for mapping directly the spatial extent of the area affected by fires with a 500m resolution.

Figure 1 (left) maps the municipalities that experienced fire events during the summer season of 2021. The figure also shows (right) the high number of areas without fire occurrences. To investigate factors affecting fires in Apulia and Basilicata municipalities, we consider a large set of auxiliary variables, integrating geo-

[1] `NIMBLE` code is available upon request to the authors.

referenced information with data available only on aggregate areal-level scale, synthesized as follows

- *proportion of days with (and without) rain* from meteorological monitoring networks provided by Apulia and Basilicata Civil Protection Departments
- *percentage of land cover use* considering the satellite data from CORINE Land Cover (CLC) inventory at level 5 of surface classification [4]
- number of cattle units (adults) from agricultural census (ISTAT)
- *socio-demographic* indicators and *relevant measurement of well-being* from ISTAT Bes report.

4 Preliminary results and further developments

For the estimation of the model in (1)-(4), a subset of covarites are used as suggested from a preliminary analysis: the municipality surface (*area_km2*), the cattle adults units (*animals*), the % of agricultural and pastures land cover (*land agri*, *pastures*), the population density (*density*), the territory typology (*mountain/flat area*), the deprivation index (*deprivation*), the urbanization rate (*urbanization*) and the proportion of rainy days in the period (*rain days*).

Left panel of Figure 2 shows the posterior distribution of the regression parameters affecting the number of fires at different Apulia and Basilicata locations. Because of the presence of large amounts of missing data in covariates, the final dataset contains 101 areas spread around the two regions and for roughly 60% of areas no fires have been observed.
In agreement with [1], the amount of land devoted to pastures has a significant positive effect on the number of fires while the amount of land devoted to agriculture has a negative effect. Moreover, our results show that mountain areas are less prone to have larger amounts of fire occurrences with respect to coastal or flat zones as well as more urbanized counties are more likely to present fires with respect to less urbanized areas.

With respect to the zero-inflation, the probability of registering excess zero counts is negatively affected by the dimension of each area as well as by the population density. The number of rainy days does not significantly impact either the mean number of fires or the probability of excess zeros.

The proposed model has been compared with a standard Poisson model ignoring the zero-inflation through WAIC index: WAIC of the proposed model is equal to 1549.889 while for the competing model is 2272.861, highlighting the necessity of accounting for the zero inflation component.

The proposed approach can be considered as a first step for the analysis of a complex phenomenon that can be improved in several aspects. Due to the spatial variability of the response variable among areas, the model can be extended by including a spatial component accounting for the fire dynamics and the neighboring structure of the areas. Relying on a very recent paper [7], we propose to extend

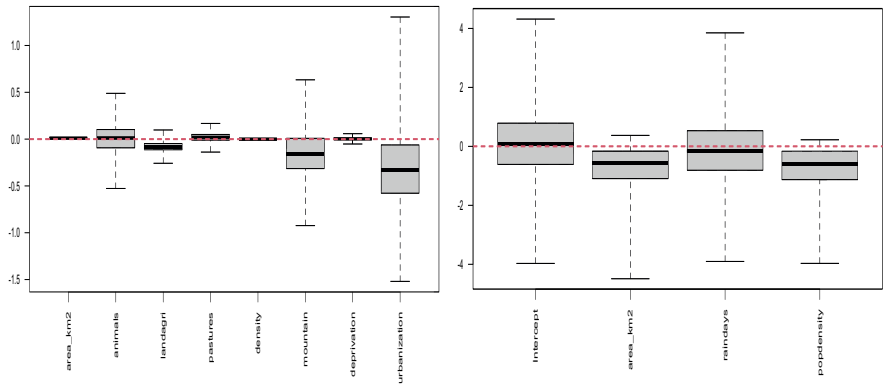


Fig. 2 Left panel: posterior distribution of regression parameters in (3). Right panel: posterior distribution of the regression parameters related to the zero component (4).

our approach for modeling measurement error to describe error contaminated count data.

References

1. Boubeta, M., Lombardia, M.J., Marey-Pérez, M.F., Morales, D.: Prediction of forest fires occurrences with area-level Poisson mixed models. J. Environ. Manage. **154**, 151–8 (2015)
2. Cameron, A.C. and Trivedi, P.K.: Regression Analysis of Count Data, 2nd edition. Cambridge University Press, Cambridge (2013)
3. de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman. C., Temple L.D., Bodik, R.: Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics. **26**, 403–413. doi: 10.1080/10618600.2016.1172487
4. European Union, Copernicus Land Monitoring Service 2022, European Environment Agency (EEA)
5. Johnston, D.W., Önder,Y.K., Rahman M.H., Mehmet A. Ulubasoglu, M.A. Evaluating wildfire exposure: Using wellbeing data to estimate and value the impacts of wildfire, Journal of Economic Behavior and Organization (2021) doi: 10.1016/j.jebo.2021.10.029
6. San-Miguel-Ayanz, J., Durrant, T., Boca, R., Maianti, P. (et al.). Advance report on wildfires in Europe, Middle East and North Africa 2021, EUR 31028 EN, Publications Office of the European Union, Luxembourg (2022) doi:10.2760/039729
7. Zhang, Q and. Yi, G.Y.: Zero-Inflated Poisson Models with Measurement Error in the Response, Biometrics. **64**, 1–25 (2022)

Assessment of the impact of anthropic pressures on the Giglio island meadow of Posidonia oceanica

Valutazione dell'impatto delle pressioni antropiche sulla prateria di Posidonia oceanica dell'isola del Giglio

Gianluca Mastrantonio, Daniele Ventura, Gianluca Mancini, Giandomenico Ardizzone

Abstract We present a Bayesian Beta regression model for the assessment of anthropic pressures on the Posidonia meadows along the Giglio island coasts. The evolution of the meadows was assessed by analysis of aerial photos taken from 1968 until 2013.

Abstract *Il lavoro presenta un modello di regressione Beta Bayesiano per la valutazione dell'impatto antropico sulle praterie di Posidonia presenti lungo la costa dell'isola del Giglio. L'evoluzione delle praterie è studiata attraverso l'analisi di foto aeree scattate annualmente tra il 1968 e il 2013.*

Key words: Beta regression, Posidonia oceanica, Bayesian statistics

1 Introduction

Posidonia oceanica is the most important and widespread endemic seagrass species in the Mediterranean Sea, capable of developing large meadows from the sea surface level up to 40-45 meters depth (Duarte, 1991). It forms one of the most valuable coastal ecosystems on Earth in terms of goods and services for its ecological, physical, economic, and bio-indicator role (Vassallo et al., 2013). Due to its wide distribution and its unique features, P. oceanica is protected by EU legislations and local measures both at species and at habitat levels. Even though the P. oceanica is protected by a legal framework its meadows are rapidly declining during the last century, mainly due to human activities, climate changes, and alien species invasion (Telesca et al., 2015). Effective coastal zone management plans and conservation ef-

Gianluca Mastrantonio
DISMA - Politecnico di Torino, e-mail: gianluca.mastrantonio@polito.it

Daniele Ventura, Gianluca Mancini, Giandomenico Ardizzone
DBA - Università di Roma La Sapienza

forts on P. oceanica could benefit from a more profound knowledge of seagrass spatial distribution. Marine spatial planning and integrated coastal zone management are pivotal in promoting sustainable growth of maritime and coastal activities and using coastal and marine resources sustainably, as also recently highlighted by the European Commission (Schaefer and Barale, 2011). Coastal benthic habitats, such as P. oceanica, can be described through spatial representations of discrete seabed areas associated with particular species, communities, or co-occurrences (Papakonstantinou et al., 2020), known as benthic or bionomic maps. These maps provide baseline information for research activities and maritime activities in coastal areas. Motivated by the above, in this work we analyze human impacts on the P.oceanica's meadow of the Giglio island in the period 1968-2013. The main source of information is the percentage of Posidonia coverage on an area extrapolated from aerial photos. Proportional data, in which response variables are expressed as percentages or fractions of a whole, are analysed in many fields. The scale-independence of proportions makes them appropriate to analyse many biological phenomena, but statistical analyses are not straightforward. Transformations to overcome these problems are often applied, but can lead to biased estimates and difficulties in interpretation. Beta regression overcomes some problems inherent in applying classic statistical approaches to proportional data.

2 Study area and human activities

The study area is represented by the Island of Giglio (Central Tyrrhenian Sea, Italy), one of the seven primary islets, plus several smaller, composing the Tuscany Archipelago National Park (TANP). The aquatic environment of Giglio Island is characterized by the presence of a vast and almost continuous P. oceanica meadow thriving on matte, sand, and rock from few centimeters below the sea surface up to 37 meters depth on a gently sloping seabed. The meadow runs all around the island except for the west-south quadrant characterized by vertical cliffs and steep bottoms, a harsh environment for P. oceanica thriving. The upper and lower edges (i.e., the landward and seaward boundaries defining the meadow) are localized at different depths and distances from the coastline. They follow the seabed slope, the hydrodynamic forces, the photosynthetic process, and the anthropogenic pressures (Montefalcone et al., 2010). The coastal area is divided into 13 zones around the perimeter of the Island (Fig.1) according to the seabed morphologies. The latter determined the visibility, which allowed the identification of Posidonia meadow limits from aerial images. Only shallow coastal areas (up to 12 meters depth) were selected for polygon editing in GIS software aimed at defining the extension of the Posidonia meadow. The same zones were divided into Shallow and Deep. No active protection is undergone on the meadow all over the area. For this reason, P. oceanica has been directly and indirectly threatened by several anthropic pressures such as the i) pleasure boats anchoring, ii) constructions (harbors, public works, urban and rural areas development) and agricultural practices, and iii) mining.

Anchoring: Due to the land proximity and its sheltered bays, Giglio Island represented a popular seaside destination for touristic boating, which anchoring was localized on the P. oceanica meadow close to the coastline. The anchoring is defined as the short-term deployment of a physical device to hold fast to the substrate by a vessel. It has been proved to disturb P. oceanica meadows at different levels (Deter et al., 2017;).
Constructions and agricultural practices: During the last fifty years, the island faced a massive anthropic outbreak in terms of touristic frequentation, leading to increased public works and urban and rural areas development. Coastal constructions involved harbor enlargement and the desalination system development. Many constructions were next to the coastline. Due to the temperate weather and the fertile soil, Giglio Island was characterized by grapevines (Vitis vinifera) and olive trees (Olea europaea) cultivations. To face the mountainous environment, terracing was adopted as agricultural practice all over the island. Terraces were built by constructing dry-stone walls, named 'greppe,' using granite blocks; landward, regolite soil was laid over the bedrock as a substrate for cultivation. Today, few terraces are actively cultivated and maintained, whereas the ones abandoned are deteriorating and collapsing, leading to landslide events and contributing to water runoff and sediment generation moving to the seaside.
Mining: Since the Roman age, mining activities have interested the island with granite, limestone, and gypsum extraction and, more recently, pyrite and further iron minerals exploitation. Granite caves, mainly localized in the eastern side of the Island from Arenella to Caldane Bays, provided monzogranite rocks up to 1950, serving all central Italy. Metamorphic and sedimentary rocks mining interested the northwestern side of the island, in the Frengo Promontory (next to Campese Bay), up to the 1960s. Caves produced limestone and gypsum, whereas pyrite and iron minerals were obtained from the Frengo mine, which closed in 1976. To move the pyrite from the mine to the barges moored in Campese Bay, a cableway was mounted on three pillars built over the P. oceanica meadow at 5 meters depth. Mining activities led to debris production and dump areas, resulting in a high quantity of the reduced size of rocks, from a few centimeters up to one meter. Each impact is recorded as intensity, presence/absence and distance between the zone and the impact source.

2.1 Further available data

Together with the above described human-activities variables, mean depth and mean slope for each zone, errors in the aerial photos, resolution of the photo, sea state when the photo was taken, are available for modelling.

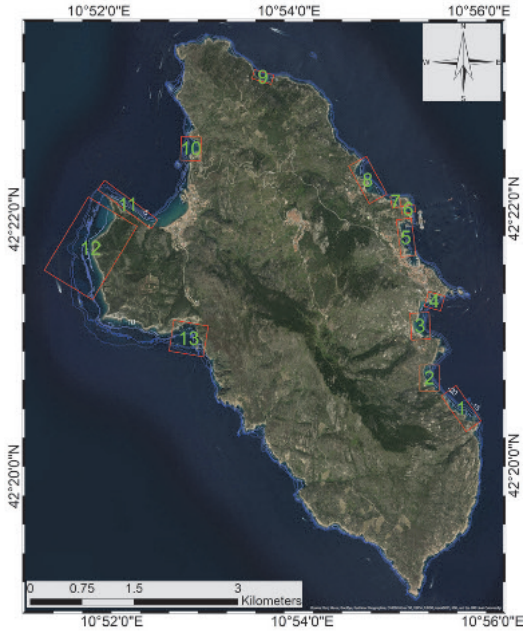


Fig. 1: Study area with the 13 zones highlighted.

3 The model

To understand the relationship of P. oceanica coverage of the Giglio island coastal area with the measured impacts and environmental conditions, we used a Beta regression model(Ferraro, Cribari-Neto, 2004), that is a generalized linear model based on the Beta distribution

$$Y_{it} \sim \text{Beta}(\mu_{it}, \tau_{it})$$

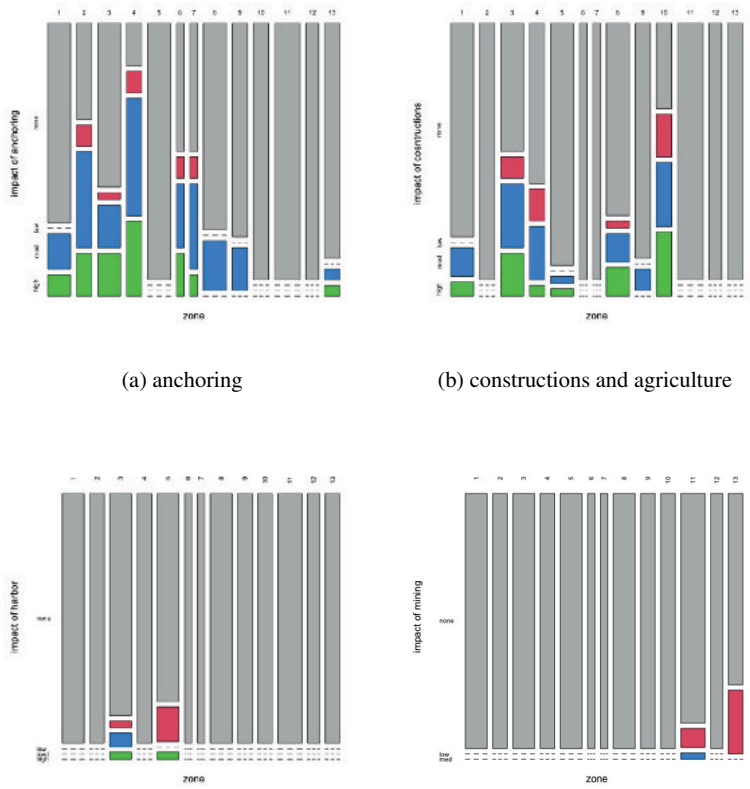
where Y_{it} is the i-th observation of Posidonia coverage at time t, μ_{it} is the mean of the distribution and τ_{it} the precision. Further

$$\text{logit}(\mu_{it}) = \beta_{0\mu}^{z_{k_i}} + \sum_{h=1}^{p} x_{hti}\beta_{h\mu}$$

and

$$\log(\tau_{it}) = \beta_{0\tau} + \sum_{l=1}^{k} x_{lti}\beta_{l\tau}$$

were $\{x_{hti}\}$ and $\{x_{lti}\}$ are the set of available data, and z_{k_i} denotes cluster membership of the zone k_i ($k_i = 1, 2, \ldots, 13$) where observation i occurs. Hence, by the same model we investigate both the influence of anthropic impacts on the Posidonia cov-



(a) anchoring

(b) constructions and agriculture

(c) harbor

(d) mining

Fig. 2: Distribution of impact levels by zone (grey = no impact, blue = low impact, green = moderate impact, red =high impact).

erage and the presence of homogeneous clusters of zones. We perform the model estimation in the Bayesian setting, implementing our code in JAGS. The set of parameters' priors distributions are: for all $\beta_\mu, \beta_\tau \sim N(0, 1000)$, $z_j \sim \text{Multinomial}(\pi)$, with $j = 1, 2, \ldots, 13$, and $\pi \sim \text{Dirichlet}(1, 1, \ldots, 1)$. We run the MCMC sampler for 160000 iterations with a burn-in of 80000, keeping 5000 samples for inference after thinning.

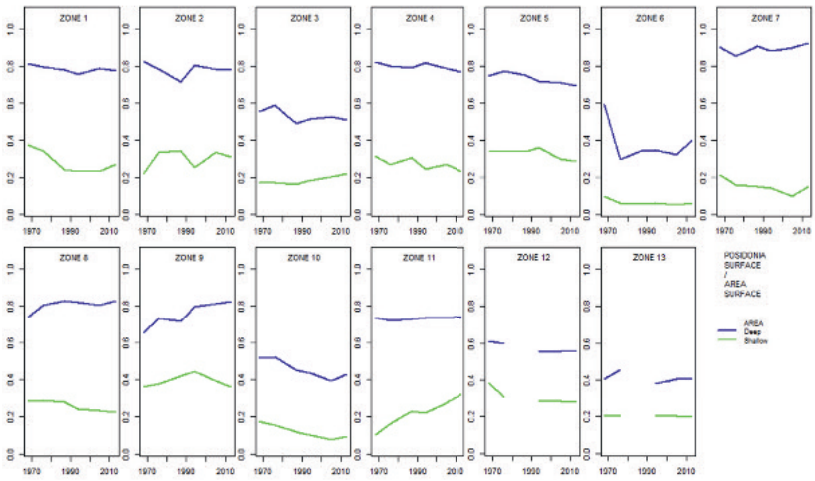


Fig. 3: Posidonia o. meadows coverage: time trend by zone and depth.

4 Exploratory data analysis and preliminary results

In Fig. 2 the mosaicplots of zones and impacts intensities are shown. It appears that only few zones are affected by multiple impacts. However, no zone is free from impacts of some kind. From further explorations of time trends (Fig.3) it appears that for some zones a decrease in the Posidonia coverage is suspected. Preliminary results from model selection based on the DIC criterion, suggest that the distance from the impact source is not influential, while the presence/absence of the impact is important in terms of model fitting. There is evidence of 4 different groups. In Fig. 2 the grouping of zone coefficients is well described. Impacts of harbor, anchorage and mining activities are relevant factors. We are currently refining the model, in particular by adding more specific variables to the description of the precision parameter τ.

Acknowledgements Gianluca Mastrantonio research has been partially supported by MIUR grant Dipartimenti di Eccellenza 2018 – 2022 (E11G18000350001), conferred to the Dipartimento di Scienze Matematiche – DISMA, Politecnico di Torino

References

1. Deter, J., Lozupone, X., Inacio, A., Boissery, P., Holon, F.: Boat anchoring pressure on coastal seabed: Quantification and bias estimation using AIS data. Mar. Pollut. Bull. 123, 175–181 (2017)
2. Douma, J.C., Weedon, J.T.: Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods Ecol Evol. 10, 1412– 1430

(2019)
3. Duarte, C.M.: Seagrass depth limits. Aquat. Bot. 40, 363–377 (1991)
4. Ferrari, S.L.P., Cribari-Neto, F.: Beta Regression for Modeling Rates and Proportions. Journal of Applied Statistics 31(7), 799–815 (2004)
5. Montefalcone, M., Parravicini, V., Vacchi, M., Albertelli, G., Ferrari, M., Morri, C., Bianchi, C.N.: Human influence on seagrass habitat fragmentation in NW Mediterranean Sea. Estuar. Coast. Shelf Sci. 86, 292–298 (2010)
6. Papakonstantinou, A., Stamati, C., Topouzelis, K.: Comparison of true-color and multispectral unmanned aerial systems imagery for marine habitat mapping using object-based image analysis. Remote Sens. 12 (2020)
7. Schaefer, N., Barale, V.: Maritime spatial planning: Opportunities and challenges in the framework of the EU integrated maritime policy. J. Coast. Conserv. 15, 237–245 (2011)
8. Telesca, L., Belluscio, A., Criscoli, A., Ardizzone, G., Apostolaki, E.T., Fraschetti, S., Gristina, M., Knittweis, L., Martin, C.S., Pergent, G., Alagna, A., Badalamenti, F., Garofalo, G., Gerakaris, V., Louise Pace, M., Pergent-Martini, C., Salomidi, M.: Seagrass meadows (Posidonia oceanica) distribution and trajectories of change. Sci. Rep. 5, 1–14 (2015)
9. Vassallo, P., Paoli, C., Rovere, A., Montefalcone, M., Morri, C., Bianchi, C.N.: The value of the seagrass Posidonia oceanica: a natural capital assessment. Mar. Pollut. Bull. 75, 157–167 (2013)
10. Waycott, M., Duarte, C.M., Carruthers, T.J.B., Orth, R.J., Dennison, W.C., Olyarnik, S., Calladine, A., Fourqurean, J.W., Heck, K.L., Hughes, A.R., Kendrick, G.A., Kenworthy, W.J., Short, F.T., Williams, S.L.: Accelerating loss of seagrasses across the globe threatens coastal ecosystems. Proc. Natl. Acad. Sci. U. S. A. 106, 12377–12381 (2009)

Accounting for complex observation processes in spatio-temporal ecological data

Processi osservazionali complessi in modelli spazio-temporali per dati ecologici

Janine Illian

Abstract In ecological research there is a strong interest in understanding how individuals – plants, animals or other organisms – interact with each other and with the environment they live in. The spatial pattern formed by the locations of individuals in space along with their properties can reflect both local interactions among individuals as well as preferences of different species for specific environmental conditions or habitats. log-Gaussian Cox models have proven to be particularly flexible in this context as are able to reflect properties of complex spatio-temporal point patterns while accounting for spatial structures not accounted for by existing covariates. Paired with computationally efficient model fitting methodology such as integrated nested Laplace approximation (INLA), realistically complex spatial and spatio-temporal models may be formulated and fitted to spatial point pattern data within feasible time (Simpson et al., 2016). In what follows we are going to illustrate how the `inlabru` wrapper of `R-INLA` helps to flexibly include complex observation design into LGC models.

Abstract *Nella ricerca ecologica esiste un forte interesse a capire come gli individui - piante, animali o altri organismi - interagiscono tra loro e con l'ambiente in cui vivono. I modelli log-gaussiani di Cox si sono dimostrati particolarmente flessibili in questo contesto, in quanto sono in grado di riflettere le proprietà di processi di punto spazio-temporali complessi e di tenere conto delle strutture spaziali non considerate dalle covariate esistenti. In combinazione con una metodologia di stima efficiente dal punto di vista computazionale, come l'integrated nested Laplace approximation (INLA), è possibile formulare modelli spaziali e spazio-temporali realisticamente complessi e adattarli alle osservazioni di processi di punto spaziali in tempi ragionevoli (Simpson et al., 2016). Di seguito illustreremo*

[1] Janine Illian, School of Mathematics & Statistics, University of Glasgow
email: Janine.Illian@glasgow.ac.uk

come il wrapper `inlabru` *di* `R-INLA` *aiuti a includere in modo flessibile procedure osservazionali complesse nei modelli* LGC.

Key words: point patterns, log-Gaussian Cox process, INLA, inlabru, Bayesian models

1 Motivation and summary

In ecological research there is a strong interest in understanding how individuals – plants, animals or other organisms – interact with each other and with the environment they live in. The spatial pattern formed by the locations of individuals in space along with their properties can reflect both local interactions among individuals as well as preferences of different species for specific environmental conditions or habitats. A statistical analysis based on spatial or spatio-temporal (marked) point process methodology can analyse these patterns and – as a result – reveal, e.g. specific habitat preferences in a changing environment. log-Gaussian Cox models have proven to be particularly flexible in this context as are able to reflect properties of complex spatio-temporal point patterns while accounting for spatial structures not accounted for by existing covariates. Paired with computationally efficient model fitting methodology such as integrated nested Laplace approximation (INLA), realistically complex spatial and spatio-temporal models may be formulated and fitted to spatial point pattern data within feasible time (Simpson et al., 2016).

In many cases, however, it is not necessarily straight forward to collect data on individuals within a study area of interest, for example because the environment a species lives in is hard to access or the general area of interest is very large. This implies that data collection has to be adapted to the to the specific study system and species. As a result, the observation processes vary with the nature of general system a study is interested in (e.g. is it terrestrial or aquatic?) and the specific behavioural patterns of a species data (e.g. are there any detection issues or are we likely to have seen every individual in the areas we surveyed?). In order to provide practically relevant modelling methodology – and software – these different observation processes have to be taken into account. Classical statistical ecology literature has typically developed specific methodology for each type of observation process, associated with a specific software package.

Rather than re-inventing the wheel every time a new observation process comes along, the `R` package `inlabru` provides a more unified approach to accounting for observation processes (Bachl et al., 2019). Here, the observation process is seen as an operation on the ecological process of interest. For example, spatially varying detection probabilities may be regarded as a thinning operation operating on a point process. The software allows us to estimate the parameters of the detection process as well as those of the process of interest simultaneously. Since `inlabru` is a wrapper around the well-known package `R-INLA` it exploits both the computational

efficiency of INLA and the flexibility of the SPDE approach to approximating the Gaussian random fields (Rue et al., 2009; Lindgren et al., 2011). This also implies that the functionality availability within `R-INLA` is also available in `inlabru` and a wide range of different spatio-temporal models that can be fitted with `R-INLA` may be fitted with it as well.

More generally, `inlabru` is not only relevant to spatial point processes and ecological data with complex observation processes. In particular, while facilitating point process modelling for log Gaussian Cox processes and accounting for complex observation processes (Yuan et al., 2017; Williamson et al., 2021), it is also relevant to modelling data without detection issues and with spatial data structures that are not point patterns. In order to make the functionality of `R-INLA` more accessible to users it provides a streamlined interface with the aim of simplifying the user's code. To ease usability the syntax within the software compartmentalises the models and aims to reflects the role of the different model components within the model.

`inlabru` provides a general and flexible, computationally efficient fitting tool for complex statistical models, extending the range of models currently available through `R-INLA`. In addition, it uses an iterative method to estimate parameters in non-linear functional relationships between a response and a covariate affecting either the process of interest or the observation process through an iterative approach. This is again particularly relevant in ecology, where e.g. detection probabilities might depend in a non-linear way on the properties of individuals of interest or the local environmental conditions. A simple example of this is a case where a half-normal detection function is used and detection probabilities depend, e.g. on the size of the animals under study.

In this talk we will discuss and illustrate the capabilities of `inlabru` through a number of examples drawn primarily from ecological applications, mainly in the context of animal conservation and population assessment. In particular, we will look at modelling partially observed point pattern data relating to orangutang conservation in Borneo, Malaysia, spatio-temporal marked point process modelling of nesting cranes in the UK (Soriano-Redondo et al., 2019), as well as data on endangered birds in Hawai'i derived from point transect sampling.

References

1. F. E. Bachl, F. Lindgren, D. L. Borchers, and J. B. Illian. `inlabru`: an r package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6): 760[U+FFFD]66, 2019. ISSN 2041-210X. doi: 10.1111/2041-210x.13168.
2. F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423[U+FFFD]98, 2011. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2011.00777.x.
   URL `https://dx.doi.org/10.1111/j.1467-9868.2011.00777.x`.
3. H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319[U+FFFD]92, 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2008.00700.x.
   URL `https://dx.doi.org/10.1111/j.1467-9868.2008.00700.x`.

4. D. Simpson, J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue. Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103 (1):49[U+FFFD]0, 2016. ISSN 0006-3444. doi: 10.1093/biomet/asv064.
URL `https://dx.doi.org/10.1093/biomet/asv064`.
5. A. Soriano-Redondo, C. M. Jones-Todd, S. Bearhop, G. M. Hilton, L. Lock, A. Stanbury, S. C. Votier, and J. B. Illian. Understanding species distribution in dynamic populations: A new approach using spatio-temporal point process models. *Ecography*, 42(6):1092–1102, 2019.
6. L. D. Williamson, B. E. Scott, M. R. Laxton, F. E. Bachl, J. B. Illian, K. L. Brookes, and P. M. Thompson. Spatiotemporal variation in harbor porpoise distribution and foraging across a landscape of fear. *Marine Mammal Science*, 2021. ISSN 0824-0469. doi: 10.1111/mms.12839.
7. Y. Yuan, F. E. Bachl, F. Lindgren, D. L. Borchers, J. B. Illian, S. T. Buckland, H. Rue, and T. Gerrodette. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11(4): 2270[U+FFFD]297, 2017. ISSN 1932-6157. doi: 10.1214/17-aoas1078.

Statistics and indicators for the recovery and resilience plan

The prominence of statistical information for the monitoring and effective implementation of the NRRP

La centralità dell'informazione statistica per il monitoraggio e l'efficace attuazione del PNRR

Andrea Petrella

Abstract In this short contribution I discuss three areas of interest where an improvement in the statistical domain is likely to provide the highest returns in terms of effective implementation and social inclusiveness of the NRRP: (i) monitoring; (ii) transparency; (iii) ex-post evaluation.

Abstract *In questo breve contributo discuto tre ambiti in cui un miglioramento nel campo statistico potrebbe fornire i maggiori benefici in termini di efficacia nell'attuazione e inclusione sociale del PNRR: (i) monitoraggio; (ii) trasparenza; (iii) valutazione d'impatto.*

Key words: NRRP, monitoring, policy evaluation

The Italian NRRP encompasses an ambitious set of reforms and investments, with the aim to spur post-pandemic recovery and to increase the resilience of the economy to future shocks. The arrangements made with the European Commission for a timely implementation of the Plan place a great emphasis on performance, measured as the capacity to meet pre-determined qualitative and quantitative goals (so-called 'milestones' and 'targets'). This represents a challenge – and possibly a good practice for the future – for Italian Public administration that is traditionally more inclined to monitor spending rather than attainments. In this context, data infrastructures and high-quality statistical information will be key to pursue the goals of the Plan. Here I will briefly discuss three areas of interest where an improvement in the statistical domain is likely to provide the highest returns in

[1] Andrea Petrella, Bank of Italy; email: andrea.petrella@bancaditalia.it

terms of effective implementation and social inclusiveness of the Plan: (i) monitoring; (ii) transparency; (iii) ex-post evaluation.

Monitoring is the area where most advancements have already been done. As a matter of fact, the orientation to monitoring is intrinsic in the design of the NRRPs that provide for a continuous supervision of the different achievements planned for each project. The variety in the number of interventions, governance layers and implementing bodies make this task extremely complex. This is the reason why the design of a monitoring system for the NRPP has been one of the foremost endeavour in the first year of application of the Plan. The Regis system realized by the State General Accounting Department is the IT infrastructure intended to collect data on the progress of each NRPP project, based on the information flows fed by the responsible administrations. On top of that, it keeps track of firms or other subjects having a contract awarded, links several external data sources to perform cross-validation checks, and also embeds a module to monitor the progress of spending for each intervention. Regis is definitely an ambitious infrastructure, whose contents are being progressively populated as the implementation phase of the NRRP gains momentum. However, the richness of the statistical information it can provide crucially relies on the responsible administrations' capacity to feed data on project advancements in a timely and accurate manner. For this reason, all interested parties should be aware of the importance of a high quality reporting, be provided with harmonized guidelines and have dedicated resources allocated to this task. Regis has especially been designed to monitor NRPP projects, but it may prospectively represent the primary monitoring instrument for all public spending initiatives. This will most probably be the case for the projects financed by the National Complementary Plan, whose implementation timeline has been designed according to the principles inspiring the NRRP. For all other initiatives outside the NRRP boundary, this might require a cultural switch within Italian Public administration toward a more performance-oriented planning.

Given the complexity of the NRRP framework and the richness of the statistical information collected for monitoring purposes, granting a transparent and effortless data access to the wider public is essential to guarantee an inclusive implementation of the Plan, to push forward the accountability of responsible administrations and to promote ex-ante analyses on the expected impact of each investment line. The website ItaliaDomani is intended to be the virtual platform through which information is channelled to the public. It has detailed sections describing the structure of the Plan and the contents of the planned interventions, and has recently launched a new open data section where information is released in a standardized and machine-readable format. At this time, the open data section still needs to be populated with valuable contents, though. In particular, real-time implementation data for each investment is an area where the greatest advances in transparency might be attained. To this aim, a greater integration with the Regis infrastructure might be pursued, designing automated routines that extract information and release them in open data format, of course respecting the relevant statistical confidentiality and personal data protection issues. Another area in which a greater information

standardization would favour transparency is the one of tender notices, which are highly heterogeneous in nature. At the moment, ItaliaDomani acts as a searchable repository for tenders, but the comparability between them is limited by the fact that the information is still unstructured and essentially textual.

Finally, let me stress that monitoring does not typically coincide with evaluation, and a timely implementation of projects does not necessarily imply their effectiveness. Whilst it is essential to deliver projects on time in order to receive the European funds as scheduled, the ultimate interest of both policymakers and the public is that NRPP measures have an impact on the relevant variables on which they are intended to act. As a matter of fact, the milestones and targets agreed upon with the European Commission hardly ever entail indicators of impact (for example, an increase in the graduation rate), while most of the times they concern procedural attainments (for example, the number of scholarships awarded). This is because the latter category of indicators is more directly verifiable than the former, which take considerably more time to materialize. Most importantly, in order to assess impact, it is first necessary to assess causality, which requires suitable data to perform a counterfactual econometric analysis. A rigorous impact evaluation exercise is not only relevant to assess the effectiveness of NRPP projects per se, but also to learn from past experiences, in order to better calibrate future policies. Of course, the evaluation phase of the NRPP measures is distant in time, as we have to wait for the complete rollout of each intervention and the full deployment of their effects. Nonetheless, it is important to figure out now what are the statistical challenges for future evaluation exercises, in order to start planning the necessary actions to take. Two issues should be tackled with greater priority. First, to perform counterfactual analysis it is crucial to have a control group, while the monitoring process is inclined to only record the "treated" units, meaning the firms or entities that receive funds or that are awarded a contract. To build a suitable control group it is instead essential to also record the units that applied but were not selected among the recipients or the contractors. The Regis system has the technical means to coherently process and store this data; it is notwithstanding necessary to activate the related information flows. Second, counterfactual techniques typically leverage on large datasets at the individual level to perform their impact evaluation analyses. Hence, for each variable of interest whose evolution we are interested to measure at the aggregate level, it would be important to observe the contribution of each elementary unit. In principle, the increased availability of administrative microdata makes this task relatively easy. However, on one hand, the necessity to link together the various data sources calls for an increased cooperation between the institutions owning proprietary data; on the other hand, suitable data management infrastructures are needed to handle such high-dimensional information. In recent years, some valuable experiences have been carried out it this direction. The wish is that we can build on them to face the challenges posed by the evaluation of the NRPP.

Big Data Analytics in mobile cellular networks as enabler for innovative statistics to evaluate the effects of Recovery and Resilience Plan actions

L'Analisi dei Big Data delle reti radiomobili cellulari come abilitatore di statistiche innovative per valutare l'effetto delle azioni del Piano di Ripresa e Resilienza

Andrea Zaramella[1], Dario Di Sorte[2], Denis Cappellari[3], Bruno Zamengo[4]

Abstract – The large deployment of packet-switched communication paradigm and the continuous growth of mobile phone services and usage, together with the pervasive deployment of network coverage, have made wide area mobile networks a valuable source of an extreme amount of data. The analytics of this time-space big-data can indeed be an innovative way to explore several insights on subscriber behaviour, presence and flows, while being strictly compliant with the GDPR requirements. In this paper we specifically focus on the tourism framework and put forward new algorithms and KPIs to investigate dynamics which can be overlooked by the official statistics. This should allow to be more effective/efficient to plan/monitor NPRR projects, the goal of which is to improve the touristic offer.

Abstract – *L'ampia diffusione del paradigma di comunicazione a commutazione di pacchetto e la continua crescita dei servizi e dell'utilizzo della telefonia mobile, insieme alla diffusione pervasiva della copertura del segnale, hanno reso le reti mobili geografiche una preziosa e immensa fonte di dati referenziati nel dominio del tempo e dello spazio. L'analisi di questi big-data è uno strumento innovativo per esplorare comportamenti, presenze e flussi degli utenti, in conformità assoluta ai requisiti di privacy del GDPR. In questo articolo ci concentriamo specificamente sul turismo e presentiamo nuovi algoritmi e KPI per indagare dinamiche che possono sfuggire alle statistiche ufficiali, con l'obiettivo di essere più efficaci/efficienti nella pianificazione/monitoraggio dei progetti PNRR in ambito offerta turistica.*

Key words: NRRP, mobile phone big-data, business analytics, tourism case-study

[1] Andrea Zaramella, Vodafone Business Italia, email: andrea.zaramella@vodafone.com

[2] Dario Di Sorte, Vodafone Business Italia, email: dario.disorte@vodafone.com

[3] Denis Cappellari, Motion Analytica, email: denis.cappellari@ motionanalytica.com

[4] Bruno Zamengo, Motion Analytica, email: bruno.zamengo@motionanalytica.com

1 Introduction

The big-data generated by cellular mobile networks may open new perspectives of analysis about users' behaviour, presence and flows in several area analysis related to the tourism and transport/mobility sectors. In Vodafone, the platform to collect and elaborate raw data to quickly calculate insights is called Vodafone Analytics.

Mobile network big-data can help to better understand the behavior of visitors, their movement independently of overnight stays, their preferences (coast lovers, country lovers, explorers), co-visits, the trajectory of their movements. These analyses can be carried out also for specific segments of users (gender, age, nationality, residence). Also, it is also possible to perform cluster analysis to identify experience tourist areas (e.g., wine territory, spa locations) outside the main and classic paths, to proactively know and quantify new trends (e.g., bike tourism), to describe over and under tourism phenomenon, to measure the seasonal adjustment of tourism request. The analytics of big data can be useful to be more effective and efficient to both design and monitor projects as a function of the mobility of people; historical data (up to 18 months backwards), the low-latency and high frequency of data refresh (near-real-time) are precious planning and monitoring instruments. In the framework of the Italian NRRP (National Recovery and Resilience Plan), a field of application of Vodafone Analytics is the area of actions related to tourism, especially those related to the attractivity of *villages and small historical centers* ("borghi"). Similar applications can also be found in the NRRP missions relevant to mobility and transport (sustainable mobility) and ecological transition. Also, some analyses can be carried out to study some aspects crossed to several sectors such as, for instance, the measure of *ecological footprint* of tourists in a geographical area, correlated to presence, mobility and means of transport.

In the following Sections, we first introduce of the main concepts related to big data in cellular mobile networks, then we give an overview of the basic definitions of tourism statistics and map them within the mobile data algorithms. Finally, we describe a case study of the municipality of Padova to show how Vodafone Analytics can integrate the official statistics to give additional interesting insights into the mobility of tourists outside the municipality where they spend the night.

2 Mobile phone Big-Data

The ability of the Vodafone Analytics big data platform to collect (time, space) data from the field and immediately elaborate them depends on:

Space granularity: the density of mobile radio cells is paramount in guaranteeing the supply of reliable data. Vodafone can rely on 200.000 network cells; within densely populated areas the diameter of a cell can be reduced to a few hundred meters, and even to a smaller dimension if a dedicated network coverage to specifically cover certain locations (e.g., malls, stadiums, train stations, airports);

Time granularity: the sampling frequency of phone/SIM position (i.e., the cell is connected to) is of the utmost importance to enable the profiling process together with an accurate analysis. Vodafone can rely on a high frequency sample that guarantees presence notifications thanks to the monitoring of raw data from all the packet-switched interactions between phones and the network (calls, messaging,

notifications, data connections, app interactions etc). A phone can be sampled up to more than one thousand time per day, and this represents a proxy of continuity;

Network coverage extension: the Vodafone mobile network is widely recognized for its quality and strength throughout the whole national territory with a percentage of population (i) covered by 2G close to 100% and (ii) covered by 4G close to 99%;

Customer base: Vodafone counts 23 million of human Italian SIMs which generate a number of raw time-space data in the order of tens of billions;

Privacy by design: Vodafone Analytics services are designed according to the principles of privacy-by-design in compliance with GDPR rules (n. 05/2014, Working Group ex Art. 29) and all data are irreversibly anonymized and aggregated. Vodafone Analytics aims at studying the behaviour of homogeneous groups of people and not the behaviour of the single user. Finally, thanks to a proprietary calibration algorithm, the number of people in a cluster is projected to the entire universe of users and not only those connected to the Vodafone mobile network.

3 Tourism Case Study

Here below the classic definitions of tourism statistics:

- visitor: traveller taking a trip to a main destination outside his/her usual environment, for less than a year, for any main purpose (e.g., business, leisure) other than to be employed by a resident being and is classified as:
  - *tourist* (or overnight visitor), if the trip includes an overnight stay;
  - *same-day visitor* (or excursionist), if the trip does not include a night.

Through Vodafone Analytics it is possible to qualify travellers, visitors, tourists and same-day visitors by observing the most active area during the night:

- prevailing night area: this information approximates the location of the overnight stay of the tourist and corresponds to the coverage of the prevailing network cells where the Italian SIM is mainly registered during the night;
- telco home location (or phone residence): corresponds to the municipality where the Italian SIM has been registered more frequently in the past sliding window, typically 6-12 months long. This way Italian users can be disaggregated by region/province/municipality, based on the telephone residence. Also, as a first approximation, the residence of mobile phone users with a foreign SIM can be associated with the nationality of the SIM.

Thus, we can classify as tourists those users whose prevailing night area differs from their telco home location and as same-day visitors those who visit a municipality for at least 3 hours without having the phone residence there or nearby.

We now represent a trial carried out in the city of Padova, which is commonly recognized as a base of overnight stays for tourists visiting the Veneto region. ISTAT traditionally measures arrivals and nights spent in the territory through a census from recorded flows in official accommodation establishment, carried out on a monthly basis. What we want to study is the daily activity of tourists to Padova; where do they move during the holiday? We have considered all people spending a night in Padova in July 2021 and classified them as follows: (i) inhabitants of Padova and nearby, (ii) inhabitants of Veneto, (iii) Italians, (iv) Foreigners.

A subset of main results is shown in Fig.1. Fig.1a presents the geographical distribution of overnight stays of tourists in the area of Padova municipality. Fig.1b gives a visual representation of tourist mobility the day after their overnight stay in

Padova for both Italians and Foreigners. Fig.1c provides a quantitative description of the time they spend within and outside Padova municipality; a significant result is that a large percentage of daytime is spent outside Padova, with a strong presence, as expected, in Venice (Fig.1d). The behaviours of Italians and Foreigners are quite similar; this is the reason why they have been always represented together.

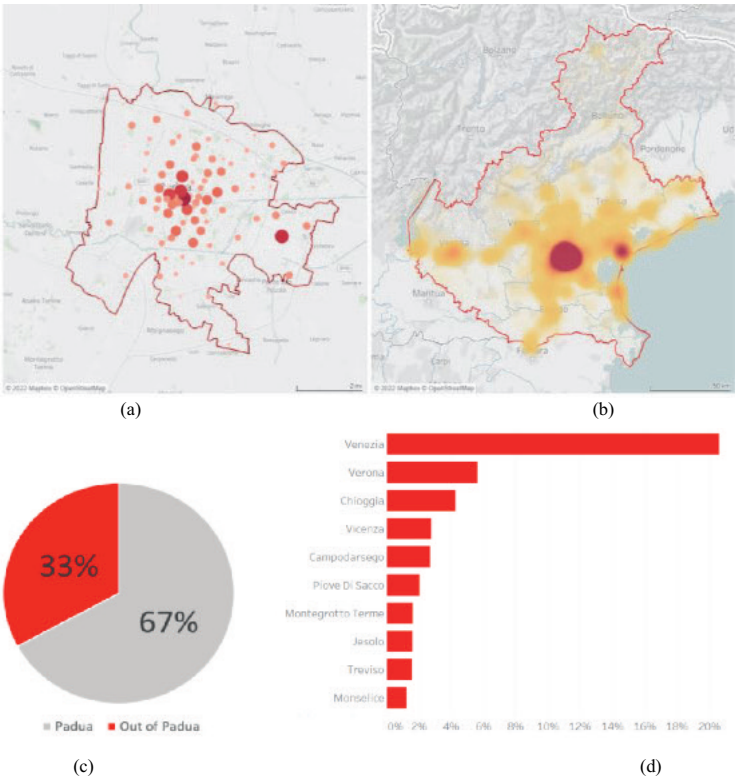


Figure 1: (a) Geographical distribution of overnight stays in Padova; (b) Mobility of tourists of Padova in Veneto region; (c) % daily time (8-22) spent within and outside Padova (d) ranking of municipalities visited by the tourists of Padova (% of total time outside Padova).

4 Conclusions

In this paper we have presented the capacity of mobile phone big data to deeply and promptly describe the behaviour of people on the territory with a high granularity on both time and space domain. NPRR is indeed a great opportunity to deeply investigate and deploy the analytical skills of big data generated near-real-time by mobile network operator for the accurate territorial monitoring. New perspectives of investigations, algorithms and KPIs can be opened in the framework of tourism, sustainable mobility, and ecological transition.

We have presented an application case in the touristic field to show how Vodafone Analytics metrics can support and complement traditional surveys to identify and quantify the presence of visitors in all the small municipalities and/or touristic destinations where there are no official accommodation establishments and the relevant data from survey/census.

Measuring the digital transition within the PA: proposals comparison

Misurare la transizione digitale nella PA: proposte a confronto

Susanna Traversa[1], Enrico Ivaldi[2]

Abstract
Although technological innovation is considered an important driver for economic recovery, interventions that are not accurately assessed against the current national scenario could increase the already existing territorial digital divide. In order to promote a fair and sustainable recovery in the course of this study, three indices of PA digitization are presented and compared, developed following the objectives W1.C1. "Digitization, innovation and security of PA" of the Italian RRF.
Abstract
Sebbene l'innovazione tecnologica sia considerata un importante driver per la ripresa economica, interventi non adeguatamente valutati rispetto all'attuale scenario nazionale potrebbero portare ad un aumento del digital divide territoriale già esistente. Al fine di promuovere una ripresa equa e sostenibile nel corso di questo studio vengono presentati e confrontati tre indici di digitalizzazione della PA, sviluppati in accordo con gli obiettivi W1.C1. "Digitalizzazione, innovazione e sicurezza della PA" del PNRR italiano.

Keywords: “eGovernment”, “Partially Ordered Set”, “Adjusted Mazziotta Pareto Index”.

1 Susanna Traversa, Università degli Studi di Genova, Dipartimento di Economia, email: susanna.traversa@gmail.com

2 Enrico Ivaldi, University of Genoa, Department of Economics and Centro de Investigationes en Econometria, Universidad de Buenos Aires, email: enrico.ivaldi@unige.it

Introduction

The challenges posed by COVID-19 have led to a break with the pre-2020 model of society, initiating a new phase characterised by extremely rapid and pervasive digital transition processes [4,5,19]. One of the areas most affected by digital implementation in recent years is the public sector. As emphasised by the European Institutions, the adoption of eGovernment processes represents an extremely strategic innovation opportunity for the Member States. The correct investment of European funds in the digital transition of the Italian Public Administration would therefore represent an opportunity for both social and economic recovery, in response to the criticalities that emerged during the pandemic. New forms of inequality - both social and industrial - due to the raising of barriers linked to weak digital infrastructure, a poor ability to reconvert business models and, above all, digital illiteracy that is still too present among the population, represent new challenges to be handled today [9,10,11,16,19]. For this reason, three different proposals for synthetic indices for assessing and monitoring the deployment of eGovernment models from a NUTS-2 perspective will be presented in this study. The possibility of having quantitative tools for the study of digitization over time represents a strategic opportunity for policymakers since the success of digital implementation policies represents for the states the possibility of strengthening their competitiveness, providing higher quality services to citizens while ensuring transparency and accountability, and reducing current and future costs [3,5,9,19].

Methodology

The methodological choice fell on the elaboration of two composite indices - i.e., aggregative approach – and a non-aggregative alternative using the Partially Ordered Set technique.
The first composite index hypothesis, referred to as $Digital_{min-max}$ was built by adopting the aggregation technique of Min-Max approach [13]. The second one, on the other hand, exploits a technique of aggregation that is becoming more and more widespread in the study of socio-economic phenomena, such as the Adjusted Mazziotta and Pareto Index ($AMPI^{+/-}$) technique, referred to in this study as $Digital_{AMPI^{+/-}}$ [14,15]. Among the reasons that led to the use of these two methods is the possibility, thanks to the choices made during their elaboration, to conduct the study of a phenomenon also from the point of view of its evolution over time. Although the use of aggregated indices is well consolidated and supported throughout the literature, this approach is not limitless. For this reason, an alternative synthesis method has been suggested employing the Partially Ordered Set – named as $Digital_{POSET}$ [6,7,8]. Among the strengths of $Digital_{POSET}$ is its great versatility of application, since it is suitable for studying qualitative, quantitative, and mixed variables, unlike aggregative methods [6,7,8,18]. In addition, since it does not require any aggregation and weighting of the variables, it avoids some of

the criticalities typical of composite the indicators, such as loss of information and the "flattering" effect concerning the incompatibilities that exist within a system of multi-dimensional variables.
For the construction of the non-aggregative index, one of the main issues in the $Digital_{POSET}$ analysis is the computational aspect. The study was therefore conducted by employing the statistical analysis software R and the CRAN package "parsec" [7]. Thanks to the graphical representation of the $Digital_{POSET}$ by the Hasse diagram, the cover relations between the individual profiles were re-constructed. Subsequently, the indicator was synthesized by exploiting the average rank method, which allows to obtain an order-preserved vector by calculating the average of the ranks assumed by each profile in a set of linear extensions of the original $Digital_{POSET}$ - i.e. $Digital_{POSET}$ subsets composed exclusively of comparable profiles (i.e. chains). Once the rankings were derived for each proposal, the results of the research were presented comparing their performance.

1.1 Selection of basic indicators

The variables used were selected through I.Stat database by a formative approach [1], and with the guidelines included by the Italian Government in the W1.C1 Mission of the RRF, dedicated to the "Digitalisation, innovation and security within the PA" [8]. Therefore, seven variables expressing the development of eGovernment were identified: (1) percentage of public institutions adopting cloud computing services to manage their data (CCS); (2) percentage of people aged 16-74 years who have used the Internet in the last 3 months and have basic digital skills (BDS); (3) percentage of institutions that have used all IT security measures (CS); (4) percentage of households with access to a broadband connection (BBC); (5) percentage of people (>14 years old) who contacted the PA in the last 12 months to obtain information through the internet (DS_I); (6) percentage of people (>14 years old) who interacted with the PA in the last 12months to download pre-filled forms through the internet (DS_PFF); (7) percentage of people (>14 years old) who have contacted the PA in the last 12months to send filled forms via the internet (DS_SM).

Discussion of the results

First of all, thanks to the cluster analysis carried out on the synthetic indices by the k-means method [17], a digital gap in terms of eGov adoption is evident. All three indices show that Sicily and Calabria are the most critical regions compared to the Italian context, with the 19th and 20th rankings, respectively. From the $Digital_{POSET}$ index perspective, however, a discrepancy appears between the two profiles, which are actually incomparable, as observed by studying the original data. Sardinia, Ligurian and Tuscany present distinctive features. While on the one hand the composite indexes place these three regions - two of which are in the center-

north - among those with medium digitalization performances, $Digital_{POSET}$ supplies more information. From the study of the non-aggregative index, in addition to the lack of comparability with any other profile, it's emerged also a great internal variability from the viewpoint of the range between which the ranks are assigned to the regions by the simulations of the final operation on the indicators, a correlation analysis between the three indices was carried out, using Kendall's aggregation coefficient ($\tau - b$) as a technique. After verifying the correct existence of the requirements for the use of Kendall's method, the results for the three indices were obtained. Through the function "corr.test()" implemented in R, an extremely positive correlation has been obtained in all three cases. Between $Digital_{AMPI^{+/-}}$ and $Digital_{min-max}$ for example, $\tau - b$=0.91 - showing a near-superposition between the two rankings. Lower is instead, although positive, $\tau - b$=0.87 calculated between $Digital_{min-max}$ and $Digital_{POSET}$. Finally, between $Digital_{AMPI^{+/-}}$ and $Digital_{POSET}$ there is a correlation $\tau - b$=0.78.
Although the values reported by the correlation index may suggest the choice between $Digital_{min-max}$ and $Digital_{POSET}$, it is necessary to consider the higher robustness obtained by the influence analysis test for the $Digital_{AMPI^{+/-}}$. Therefore, it is suggested that the final comparison between the latter aggregative technique and the $Digital_{POSET}$index should be the non-aggregative alternative.

Conclusions

In the light of what emerged during the analysis of the technical literature and from the considerations during the discussion of the results, we confirm the limitations linked to the use of synthetic indicators to the advantage of non-aggregative measures, which make it possible to understand the real variability that exists for a statistical unit, the existing relations of order between profiles without thinking about incompatibilities incurring a "flattening" risk in the interpretation of the results. Over the coming years, it will be necessary to update the datasets to continue checking the evolution of eGov systems at the NUTS-2 level, to understand the real state the art of government policies in terms of the digital transition.

References

1. Diamantopoulos, A., Riefler, P and Roth K. P. (2008). "Advancing formative measurement models.". *In Journal of business research,* 61(12):1203–1218.
2. Berkhin, P. (2006). "A survey of clustering data mining techniques". In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.
3. European Commission (2020). "2020 Digital Compass. The European way for the digital decade".
4. European Commission, (2020). "Europe's moment: Repair and prepare for the next generation".
5. European Commission (2021). "Digital economy and society index (desi) 2021: thematic chapter". Available online: https://ec.europa.eu/newsroom/dae/redirection/document/80563.
6. Fattore M (2017). "Synthesis of indicators: The non-aggregative approach". In *Complexity in society: From indicators construction to their synthesis*, pages 193–212. Springer.

7. Fattore M and Arcagni A., (2014). “Parsec: An R package for poset-based evaluation of multidimensional poverty.”. In *Multi-indicator systems and modelling in a partial order*, pages 317–330. Springer
8. Fattore M., (2016). “Partially ordered sets and the measurement of multidimensional ordinal deprivation.”. In *Social Indicators Research*, 128(2):835–858.
9. Governo italiano, (2021). “Piano Nazionale di ripresa e resilienza”. Available online: https://www.governo.it/sites/governo.it/files/PNRR_0.pdf.
10. K. K. Larsson., (2021). “Digitization or equality: When government automation covers some, but not all citizens.”. In *Government Information Quarterly*, 38(1).
11. Kuc-Czarnecka, M., (2020). “COVID-19 and digital deprivation in Poland.”. In *Oeconomia Copernicana*, *11*(3), 415-431.
12. Lallmahomed, M.Z, Lallmahomed, N., and Lallmahomed G. M., (2017). “Factors influencing the adoption of e-government services in Mauritius.”. In *Telematics and Informatics*, 34(4):57–72.
13. Maggino, F, (2017). “Complexity in Society: From Indicators Construction to their Synthesis, volume 70 of Social Indicators Research Series.”. In *Springer International Publishing*, Cham, 2017. ISBN 978-3-319-60593-7 978-3-319-60595-1.
14. Mazziotta M and Pareto A., (2012). “A non-compensatory approach for the measurement of the quality of life.”. In *Quality of life in Italy*, pages 27–40. Springer.
15. Mazziotta M and Pareto A., (2016). “On a generalized non-compensatory composite index for measuring socio-economic phenomena.”. In *Social indicators research*, 127(3): 983–1003. ISBN: 1573-0921 Publisher: Springer.
16. Meliciani V. and Pini M, (2021). “Digitalizzazione e produttività in Italia: Opportunità e rischi del PNRR.”.
17. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman H, and Giovannini E, (2005). “Handbook on constructing composite indicators: methodology and user guide. Organisation for economic cooperation and development (OECD).”. In *Statistics Working Paper JT00188147*, OECD, France.
18. Penco, L., Ivaldi, E., & Ciacci, A., (2021). “Entrepreneurial ecosystem and well-being in European smart cities: a comparative perspective.”. In *The TQM Journal*.
19. Plekhanov, D., (2020). “Digital Economy Outlook 2020”. *OECD*.

Guest Session - European Network for Business and Industrial Statistics (ENBIS)

Interpretability in functional clustering with an application to resistance spot welding process in the automotive industry

Interpretabilità nelle tecniche di clustering di dati funzionali mediante un'applicazione al processo di saldatura a resistenza per punti nell'industria automobilistica

Christian Capezza, Fabio Centofanti, Antonio Lepore, Biagio Palumbo

Abstract The functional clustering problem is recurrent in the Industry 4.0 framework with the ultimate goal to turn homogeneous cluster identification of profile data into a valuable interpretation of the phenomenon at hand. The concept of interpretability was recently addressed in functional cluster analysis by developing sparse methods able also to detect the portion of profile domain determining the cluster mean differences. This contribution aims to practically motivate the need for spreading sparse functional clustering methods in the industry through an application of the SaS-Funclust method proposed in [3] on the `ICOSAF` project functional data set which pertains to the automotive industry and contains observations of dynamic resistance curves, commonly recognized as the complete technological signature of the resistance spot welding process.

Abstract *Le tecniche di clustering di dati funzionali sono molto diffuse nel paradigma Industria 4.0 con l'obiettivo di interpretare il fenomeno in esame attraverso l'identificazione di sottogruppi omogenei dei segnali osservati. Il concetto di interpretabilità viene tradotto nell'ambito del clustering mediante lo sviluppo di metodi sparsi, in grado di identificare le porzioni di segnale che impattano maggiormente sulle differenze in media presenti tra i diversi gruppi identificati. Questo lavoro intende motivare la diffusione nell'industria di tecniche di clustering interpretabili, promuovendo l'uso del metodo SaS-Funclust proposto in [3] attraverso l'applicazione al dataset `ICOSAF` nell'ambito dell'industria automobilistica, che contiene osservazioni di curve di resistenza dinamica, comunemente riconosciute come firma tecnologica del processo di saldatura a resistenza per punti.*

Key words: Functional data analysis, Functional clustering, Sparse clustering, Interpretability

---

Christian Capezza, Fabio Centofanti, Antonio Lepore*, Biagio Palumbo
Department of Industrial Engineering
University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy
*e-mail: `antonio.lepore@unina.it`

1 Introduction

The dramatic advances in computational power and technology have allowed scientists and practitioners in business and industry to acquire and store massive and complex data apt to be modelled as continuous random functions defined on a compact domain, which are usually, and hereinafter, referred to as *functional data*, *profile data* or simply *profiles*. However, the most common practice in industrial data analysis is to use any domain knowledge to extract univariate or multivariate attributes from observed profiles, even though this is markedly criticized as problem-specific, arbitrary, and risky of hiding useful information contained by the original profile. To circumvent this issue, the most natural idea is to cross-fertilize industrial best practices with functional data analysis (FDA) techniques [13, 7, 4, 8] and use or develop FDA methods to be directly applied on functional data as founding elements. This applies also to the so-called functional *clustering problem* in the unsupervised statistical learning setting, that is the identification of homogeneous subgroups (clusters) in a functional data set, without having specific knowledge about the true underlying clustering structure. The term *homogeneus* means that data falling in each group are more similar than those falling in different groups, with respect to a given similarity measure. As in any FDA problem, the intrinsic infinite dimensionality of functional data does not make the functional cluster analysis a mere extension of multivariate clustering. Classical overview of functional clustering methods can be found in [13, 4]. The functional clustering problem is recurrent in the Industry 4.0 framework where the quality characteristic of interest is often in the form of profile and has the ultimate goal to turn homogeneous cluster identification of profiles into a valuable interpretation of the process, or more in general, the phenomenon at hand.

The concept of *interpretability* is a broader issue to be faced in the development of insightful statistical approaches not only in business and industry but also in a large variety of applications such as medical sciences, law and justice. This concept was recently discussed during the ENBIS (European Network for Business and Industrial Statistics) Workshop "Interpretability for Industry 4.0" that was held at the University of Naples Federico II (Italy) on July 12-13, 2021 [9] and offered real-world industrial motivations and deep methodological insights on this topic [10]. Even though there is a lack of consensus about the rigorous definition, interpretability essentially refers to a profound cognitive process as the ability of a model or technique (or any element related to them, e.g., inputs, outputs, predictions) to support human decisions based on them [10]. This ability may have positive consequences on the acceptability of any proposed tool and its relative industrial deployment. Interpretability in functional clustering was recently addressed by developing *sparse* methods which are able to jointly cluster profiles and detect the portion of the profile domain that mostly determines the clustering, hereinafter referred to as *informative portion*. As in the multivariate setting, [6, 12, 11, 15], where some attributes could be completely noninformative to uncover the clustering structure of interest, sparse functional clustering methods [5, 14, 3] improve the interpretability of the solution, by imputing the presence of the clustering structure to the informative portion, as

well as its accuracy, because it avoids noninformative portions to possibly hide the actual clustering structure.

The paper aims to practically motivate the need for spreading sparse functional clustering methods in industry, by promoting the use of the SaS-Funclust method that was proposed in [3] and shown to outperform other methods already appeared in the literature before. SaS-Funclust is based on a functional Gaussian mixture model whose parameters are estimated by maximizing an objective function obtained by penalizing a log-likelihood function with roughness and functional adaptive pairwise penalties. The roughness penalty is introduced to impose some smoothness to the estimated cluster means, while the functional adaptive pairwise penalty identifies the informative portion by shrinking the means of separated clusters to some common values.

The remainder of the paper is as follows. Section 2 motivates the need for sparsity through a simulated numerical toy example. Section 3 applies the SaS-Funclust method to the `ICOSAF` project data set, which is a functional data set acquired during lab tests at Centro Ricerche Fiat (CRF) to characterize a resistance spot welding process in the automotive industryand openly available online [1]. A brief conclusion is presented in Section 4.

2 A simulated toy example

Figure 1 shows the cluster means estimated for a simulated data set, in which the real number of clusters is $G = 3$, by the SaS-Funclust method through the R package `sasfunclust` available on CRAN []. The informative portion of the domain for each pair of clusters is correctly recovered. The estimated cluster means are indeed pairwise fused over approximately the same portion of the domain as the true cluster means pairs. Note that, for the clusters whose true means are equal over $t \in (0.2, 1.0]$, the SaS-Funclust method identifies the informative portion of domain roughly in $[0.0, 0.2]$.

3 An application to resistance spot welding process in the automotive industry real-case study

Starting from the idea given by the simulated toy example of Section 2, in this real application we want to demonstrate the practical advantages of a sparse functional cluster analysis, in terms of interpretability. The `ICOSAF` project data set mentioned in the introduction contains 538 dynamic resistance curves (DRCs) acquired over a regular grid of 238 points equally spaced by 1 ms. DRC is recognized as the full technological signature in the resistance spot welding processes. Further details on this data set can be found in [2]. In this application, we focus on the DRCs estimated by means of the central differences method applied to the DRC values sampled each

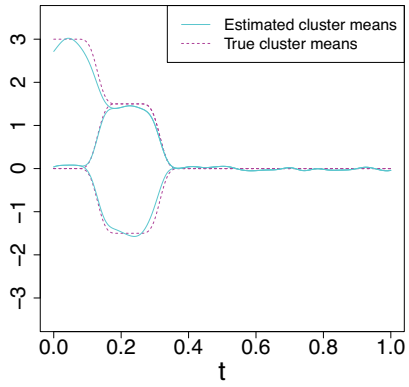
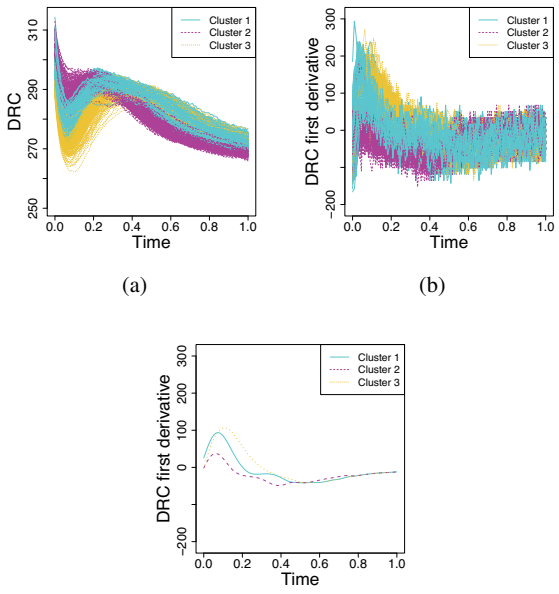


Fig. 1: True and estimated cluster means obtained through the SaS-Funclust method for a simulated data set in which the real number of clusters is $G = 3$.

2 ms. Figure 2 (a) shows the 538 functional observation of DRCs warped, without loss of generality, on the same compact domain $[0, 1]$, whereas Figure 2 (b) shows the corresponding derivative functions. In this setting, the aim of the analysis is to cluster DRCs and identify homogenous groups of spot welds that share common mechanical and metallurgical properties. Based on the considerations provided by [2] as well as on cluster number selection methods that are described for the SaS-Funclust and competing methods in [3], the number of clusters is set equal to $G = 3$. The estimated cluster mean of DRC derivative functions are displayed in Figure 2 (c), colored by cluster identified by the SaS-Funclust method. The same colours are used accordingly in Figure 2 (a) and (b), for graphical convenience. Even though at the first glance of Figure 2 the SaS-Funclust method seems to provide partitions that are similar to those obtained in [2] through the FPCA-based methods, the former clearly enables a more insightful interpretation of the results. The SaS-Funclust method is in fact able to effectively fuse cluster mean functions over noninformative portions of the DRC domain.

The mean function of DRC derivatives in clusters 1 and 3 are fused approximately from 0.5 to 1, due to comparable decreasing rate of the DRCs over these clusters. Instead, the mean of cluster 2 is fused with other cluster means between 0.8 and 1, only. Differences between mean functions of the three clusters are plainly visible in the first part of the domain. In particular, note that DRCs of cluster 2 show a smaller mean of the derivative function and reach their peaks (i.e., zeros of the first derivative) earlier than those of clusters 3 and 1. Roughly speaking, the plot in of Figure 2 (c) is a powerful support to display how the mean function behaviours differ over informative portions, as it allows practitioners to effectively filter out the focus of the analysis from the portions of domains where the estimated cluster mean functions are fused.



(a) (b)

(c)

Fig. 2: (a) 538 DRCs and (b) the corresponding derivative functions from the `ICOSAF` project data set colored by cluster identified by the SaS-Funclust method; (c) estimated cluster mean functions.

4 Conclusion

The advantages of the SaS-Funclust method, as a nice example of a sparse functional clustering method, are used to encourage the use of functional clustering methods in the industry, in place of traditional univariate or multivariate techniques. These in fact force practitioners to extract scalar attributes from quality characteristic naturally observed as functional data. The sparsity of the final solution induced by the functional adaptive pairwise fusion penalty of the SaS-Funclust method [3] allows pairs of cluster mean functions to be separated only over informative portions of profiles, or, in other terms, to be exactly equal over noninformative portions of the domain. In many applications, as in the `ICOSAF` project data set analyzed, such informative portions are very limited. In this case, a sparse clustering method may automatically improve interpretability. The specific application to the `ICOSAF` project data set identified homogenous groups of DRCs with different rate of change in the first part of the process alone. The identification of this behaviour, i.e., informative portion of DRC domain has been confirmed by CRF experts as a novel insight into the resistance spot welding process characterization which can naturally guide prac-

titioners to define, in a later stage of process learning, the most effective proxy of the final quality of spot welds produced.

Acknowledgements The authors acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support. The present work was developed within the activities of the project ARS01_00861 "Integrated collaborative systems for smart factory - ICOSAF" coordinated by CRF (Centro Ricerche Fiat Scpa - `www.crf.it`) and financially supported by MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca). The authors are extremely grateful to CRF WCM R&I (World Class Manufacturing Research and Innovation) sites of Orbassano and Campus Melfi, for the access to experimental data, and the general support over the course of these activities. In particular, the authors wish to thank Alessandro Zanella, Gianmarco Genchi, and Mariano Quagliano for their technological insights in the interpretation of the results.

References

1. Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: funclustrsw: functional clustering for resistance spot-welding data. https://github.com/unina-sfere/funclustRSW (2020)
2. Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: Functional clustering methods for resistance spot welding process data in the automotive industry. Applied Stochastic Models in Business and Industry **37**(5), 908–925 (2021)
3. Centofanti, F., Lepore, A., Palumbo, B.: Sparse and smooth functional data clustering (2021). DOI 10.48550/ARXIV.2103.15224. URL https://arxiv.org/abs/2103.15224
4. Ferraty, F., Vieu, P.: Nonparametric functional data analysis theory and practice. Springer Science Business Media (2006)
5. Floriello, D., Vitelli, V.: Sparse clustering of functional data. Journal of Multivariate Analysis **154**, 1–18 (2017)
6. Friedman, J.H., Meulman, J.J.: Clustering objects on subsets of attributes (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) **66**(4), 815–849 (2004)
7. Horváth, L., Kokoszka, P.: Inference for functional data with applications. Springer Science & Business Media (2012)
8. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
9. Lepore, A., Palumbo, B., Poggi, J.M.: Interpretability for Industry 4.0, ENBIS workshop, University of Naples Federico II, Italy, July 12-13 2021
10. Lepore, A., Palumbo, B., Poggi, J.M. (eds.): Interpretability for Industry 4.0: Statistical and Machine Learning Approaches, to appear, Springer
11. Maugis, C., Celeux, G., Martin-Magniette, M.L.: Variable selection for clustering with gaussian mixture models. Biometrics **65**(3), 701–709 (2009)
12. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. Journal of the American Statistical Association **101**(473), 168–178 (2006)
13. Ramsay, J.O., Silverman, B.W.: Functional data analysis. Wiley Online Library (2005)
14. Vitelli, V.: A novel framework for joint sparse clustering and alignment of functional data. arXiv preprint arXiv:1912.00687 (2019)
15. Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. Journal of the American Statistical Association **105**(490), 713–726 (2010)

Statistical process monitoring of thermal images in additive manufacturing: a nonparametric solution for in-situ monitoring

Monitoraggio statistico di processo di immagini termiche in manifattura additiva: una soluzione non parametrica per il monitoraggio in situ

Panagiotis Tsiamyrtzis[1], Marco Luigi Giuseppe Grasso[2] and Bianca Maria Colosimo[3]

Abstract Statistical Process Control and Monitoring (SPC/M) is a well-defined area of statistics, which aims to provide tools (typically in the form of control charts) that can be used in monitoring the quality in an ongoing process. Specifically, the goal is to determine whether a process works under statistical stability (i.e. with endogenous to the process variation) also called "In Control" (IC) state, or if any assignable (i.e. exogenous to the process) variation exists, named "Out of Control" (OOC) state. A powerful control chart will be able to detect a transition form the IC to the OOC state "soon" after its occurrence, maintaining a "low" false alarm rate. Once the IC state distribution is estimated, then a control chart can be calibrated and used for the process monitoring. The control chart choice will depend on the data dimension (univariate/multivariate), on the type of parameter shift that we aim to detect (transient/persistent) and are built under the Frequentist, Non-parametric or the Bayesian approach.

In this work our focus is in the use of SPC/M to a novel field of engineering called Additive Manufacturing (AM), where a three dimensional object is build, layer by layer, from a Computer Aided Design (CAD), permitting custom made products that would be impossible to construct otherwise. In AM we have a new quality monitoring framework as we move from mass production to a single custom made product. This is the first challenge that we face since we are not able to talk about the quality stability over multiple products, as we do in SPC/M, but we need to guarantee a stable (IC) state over a single product/process (Colosimo et al, 2018).

[1] Panagiotis Tsiamyrtzis, Dept. of Mechanical Engineering, Politecnico di Milano, panagiotis.tsiamyrtzis@polimi.it
[2] Marco Luigi Giuseppe Grasso, Dept. of Mechanical Engineering, Politecnico di Milano, marcoluigi.grasso@polimi.it
[3] Bianca Maria Colosimo, Dept. of Mechanical Engineering, Politecnico di Milano, biancamaria.colosimo@polimi.it

As AM processes move from one layer to the next, opportunities arise as quality characteristic measurements are of non-invasive nature and in-situ in-line monitoring represent a significant opportunity for first-time right production (Grasso et al, 2021). The most widely used sensors in such cases are cameras (infrared and/or visual), which can record in near real time the ongoing process, providing time-stamped images that can be used in monitoring the quality, i.e. we have unstructured data. Nowadays, video based data become more and more informative of the process, as the frame rate and the resolution increases, producing orders of magnitude larger volumes of data compared to what we used to have in the past. In the industry 4.0 era, where big data streams are available, the SPC/M (and statistics in general) faces the challenge of handling efficiently all this flow of information. Opportunities and challenges of moving SPC/M to image data has been discussed in the literature (Megahed et al., 2011) and several approaches have been proposed in the framework of image-based SPC/M for AM (Grasso et al., 2018; Colosimo and Grasso, 2018 and Yan et al, 2021).

In a recent work, Tsiamyrtzis et al. (2021) presented two novel non-parametric process monitoring methodologies that utilized the concept of partial first order stochastic dominance (PFOSD) in handling all the above mentioned challenges. In this work they used infrared thermography (i.e. univariate) pixel based data for which spatial information was neglected permitting complex underlying dynamics and tested the performance on simulated and real data from the production of zinc samples, demonstrating efficient performance in the presence of different OOC scenarios of various severity levels.

In this contribution, the PFOSD approach and different extensions are presented and discussed in the framework of in-situ SPC/M for AM. In particular, a multivariate extension of the approach is discussed to improve performance of the proposed method by enhancing the informative content to be monitored layerwise. A self-starting solution is also presented as very appealing to start monitoring layerwise with the beginning of the process, breaking free of the usual off-line calibration phase.

References

1. Colosimo, B. M., Huang, Q., Dasgupta, T., & Tsung, F. (2018). Opportunities and challenges of quality engineering for additive manufacturing. Journal of Quality Technology, 50(3), 233-252.
2. Colosimo, B. M., & Grasso, M. (2018). Spatially weighted PCA for monitoring video image data with application to additive manufacturing. Journal of Quality Technology, 50(4), 391-417.
3. Grasso, M., Demir, A. G., Previtali, B., & Colosimo, B. M. (2018). In situ monitoring of selective laser melting of zinc powder via infrared imaging of the process plume. Robotics and Computer-Integrated Manufacturing, 49, 229-239.
4. Grasso, M. L. G., Remani, A., Dickins, A., Colosimo, B. M., & Leach, R. K. (2021). In-situ measurement and monitoring methods for metal powder bed fusion–An updated review. Measurement Science and Technology.
5. Megahed, F. M., Woodall W. H., and Camelio J. A. (2011). A review and perspective on control charting with image data. Journal of Quality Technology 43 (2), pp. 83–98. doi:10.1080/00224065.2011.11917848.
6. Tsiamyrtzis P., Grasso M. & Colosimo B.M. (2021). Image based statistical process monitoring via partial first order stochastic dominance. Quality Engineering (online), DOI: 10.1080/08982112.2021.2008974.
7. Yan, H., Grasso, M., Paynabar, K., & Colosimo, B. M. (2021). Real-time detection of clustered events in video-imaging data with applications to additive manufacturing. IISE Transactions, 54(5), 464-480.

Guest Session - International Biometric Society (IBS) - Italian region

Multiple Arrows in the Bayesian Quiver: Bayesian Learning of Partially Directed Structures from Heterogeneous Data

Frecce Multiple all'Arco Bayesiano: Apprendimento Bayesiano da Dati Eterogenei di Strutture Parzialmente Orientate

L. La Rocca, F. Castelletti, S. Peluso, F.C. Stingo and G. Consonni

Abstract Motivated by the identification of complex dependencies in biological networks, we present a Bayesian method for structural learning of graphical models that exhibits two distinctive features: i) it does not assume a priori an ordering of the variables, but it learns arrows when possible and lines otherwise; ii) it assumes that the observations form subgroups having different but similar structures.
Abstract *Motivati dall'identificazione di dipendenze complesse in reti biologiche, presentiamo un metodo bayesiano per l'apprendimento strutturale di modelli grafici che esibisce due caratteristiche distintive: i) non assume a priori un ordinamento delle variabili, ma apprende frecce quando possibile e linee altrimenti; ii) assume che le osservazioni formino sottogruppi aventi strutture diverse ma simili.*

Key words: Markov equivalence, Markov random field, objective Bayes

---

Luca La Rocca
Department of Physics, Informatics and Mathematics, Università degli Studi di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy, e-mail: luca.larocca@unimore.it

Federico Castelletti
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Edificio Lanzone 18, 20123 Milano, Italy, e-mail: federico.castelletti@unicatt.it

Stefano Peluso
Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy. e-mail: stefano.peluso@unimib.it

Francesco Claudio Stingo
Department of Statistics, Computer Science, Applications "G. Parenti", Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy, e-mail: francescoclaudio.stingo@unifi.it

Guido Consonni
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli 1, Edificio Lanzone 18, 20123 Milano, Italy, e-mail: guido.consonni@unicatt.it

1 Introduction

Biological networks, where biomolecules are represented by nodes and molecular influences by edges, are crucial to modern biology [14]. The representation can be made precise as a graphical model for a vector of molecular variables to be measured in biological samples, but it is important to understand that there is no unique way to take this step: Section 2 illustrates this point. A biological network interpreted as a graphical model can be learned from data [8], and this can be useful to identify the complex dependencies represented by its structure.

It is not uncommon for biological samples to form subgroups in such a way that greater similarity is expected within groups than across groups: Kornblau et al. [11], for instance, measured protein levels on leukemia patients classified in 4 subtypes (17 subjects of type M0, 34 subjects of type M1, 68 subjects of type M2, and 59 subjects of type M4) disregarding subtypes with fewer observations. In such a case, if a single structure is learned from all samples, the estimate will be based on 178 observations, but the differences between subtypes will be lost. On the other hand, if individual structures are learned from each group of samples, the estimates will be based on 68 observations at best (and 17 observations at worst) despite the fact that some common structure across subtypes is only to be expected. This tension, which is at the heart of multiple structural learning, was overcome by Peterson et al. [17] using a Bayesian method to borrow strength across subgroups.

It turns out that Bayesian methods are well-suited to graphical models: besides their ability to express uncertainty in natural terms, using simple concepts like the probability of inclusion for a given edge, they can incorporate prior information and encourage sparsity; see Ni et al. [15] for a recent overview geared towards modern biological applications.

Peterson et al. [17] dealt with undirected graphical models. We [6, 12] deal with (acyclic) directed graphical models, which introduce the problem of directing the edges of the graph (without forming cycles). If an ordering of the variables is known a priori, the graphical model reduces to a product of regression models and can be learned very effectively; see Altomare et al. [1] for a treatment of the simpler case without subgroups. If no ordering of the variables is know a priori, one faces Markov equivalence: different graphs may give rise to the same statistical model, so that observational data are unable to distinguish one from another, and not all arrows can be directed; see Section 2 for more information. We work in this second, harder, but more realistic, scenario, as described in Section 3.

2 Graphical Models

Let $Y_1, \ldots, Y_q$ be q molecular variables of interest (random variables representing a population of interest) collected together in a vector $\mathbf{Y} = (Y_j)_{j \in V}$ indexed by the set of biomolecules $V = \{1, \ldots, q\}$. Each variable will be associated to a node in a graph $\mathbb{G}$, like those depicted in Fig. 1, with edges representing molecular influences.

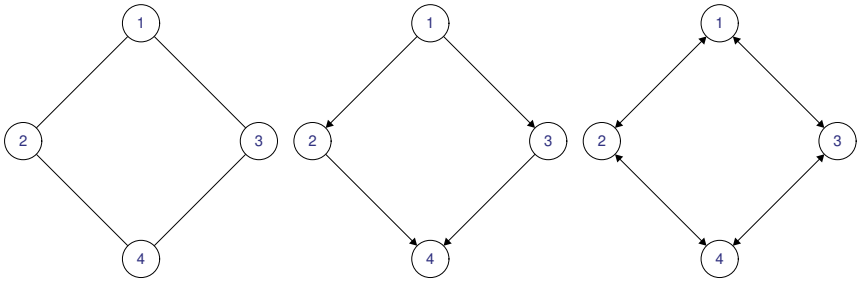


Fig. 1 Three example graphs for four variables: an undirected graph (left), an acyclic directed graph (center), and a bidirected graph (right).

The lacking edges of $\mathbb{G}$ can be read as (conditional) independence statements about subvectors of $\mathbf{Y}$, thus defining a statistical model for $\mathbf{Y}$, by choosing a specific Markov property [9, 13, 19], which is usually suggested by the type of edges that form the graph. For instance, the left graph in Fig. 1 consists of undirected edges (lines) and is usually interpreted as stating

$$Y_1 \perp\!\!\!\perp Y_4 \mid Y_2, Y_3 \quad \text{and} \quad Y_2 \perp\!\!\!\perp Y_3 \mid Y_1, Y_4 \tag{1}$$

where $\perp\!\!\!\perp$ denotes independence and $\mid$ conditioning. On the other hand, the central graph in Fig. 1 consists of directed edges (arrows) and is typically interpreted as stating

$$Y_1 \perp\!\!\!\perp Y_4 \mid Y_2, Y_3 \quad \text{and} \quad Y_2 \perp\!\!\!\perp Y_3 \mid Y_1 \tag{2}$$

where the second independence statement is no longer conditional on Y_4. Finally, the right graph in Fig. 1 consists of bidirected edges (double arrows) and is typically interpreted as stating

$$Y_1 \perp\!\!\!\perp Y_4 \quad \text{and} \quad Y_2 \perp\!\!\!\perp Y_3 \tag{3}$$

where there is no conditioning at all. Clearly, equations (1), (2) and (3) have different implications on the data that can be observed from $\mathbf{Y}$.

We remark that the left and right graphs in Fig. 2 are essentially the same graph, because they only differ in the choice of depicting symmetric influences as lines or double arrows, but the statistical models given by (1) and (3) are radically different, because the pictorial choice leads to using different Markov properties. One should be wary of a potential mismatch between the substantial meaning of a biological network and its implications on data when it is interpreted as a graphical model (statistical model specified by a graph through a Markov property).

In the following, we restrict our attention to acyclic directed graphs (like the central one in Fig. 1). These graphs are especially interesting, because they explicitly provide a data generating mechanism and can be used for causal reasoning [16]. It should be noted, however, that bidirected graphs (like the right one in Fig. 2) can emerge from acyclic directed graphs under marginalization [9]; this implies that we are assuming all relevant variables are observed as components of $\mathbf{Y}$.

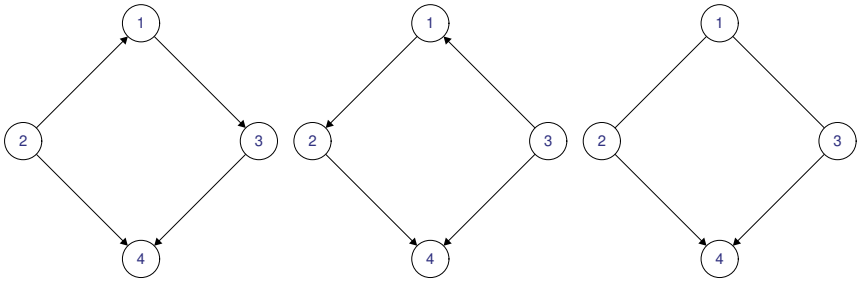


Fig. 2 Two equivalent acyclic directed graphs (left, center) and their essential graph (right); the latter has $\{1,2,3\}$ and $\{4\}$ as chain components.

Two graphs are Markov equivalent (under a given Markov property) if they define the same statistical model: this happens, for acyclic directed graphs, if and only if [20] they have the same skeleton (undirected graph obtained turning all arrows into lines) and the same immoralities (sugbraphs $i \to k \leftarrow j$ with i and j not joined by an edge). For instance, the left and central graphs in Fig. 2 are Markov equivalent: they share the left graph in Fig. 1 as skeleton and have $2 \to 4 \leftarrow 3$ as their unique immorality. They are also equivalent to the central graph in Fig. 1 and it is clear that data observed from $\mathbf{Y}$ will not be able to discriminate between these three graphs. Hence, in effect, we will be learning a Markov equivalence class.

Each Markov equivalence class can be represented by an essential graph [2] like the right graph in Fig. 2. The nodes of an essential graph can be partitioned in blocks, called chain components, such that: i) nodes in the same block are connected by lines; ii) nodes in different blocks can be joined by arrows; iii) there are no cycles containing arrows. This partitioning can be used to factorize the probability density function of $\mathbf{Y}$ (conditional on the essential graph $\mathbb{G}$) as

$$f_{\mathbb{G}}(\mathbf{y}) = \prod_{\tau \in \mathcal{T}} f_{\mathbb{G}}(\mathbf{y}_\tau \mid \mathbf{y}_{\mathrm{pa}(\tau)}), \quad \mathbf{y} \in \mathbb{R}^q \tag{4}$$

where $\mathcal{T}$ is the family of all chain components of $\mathbb{G}$, τ denotes the generic chain component, $\mathrm{pa}(\tau) = \{i \in V \mid \mathbb{G} \text{ contains } i \to j \text{ for some } j \in \tau\}$ is the parent set of τ, and $\mathbf{y}_\tau = (y_j)_{j \in \tau}$ denotes the subvector of $\mathbf{y}$ indexed by τ. Equation (4) defines the same statistical model as the acyclic directed graphs in the class represented by $\mathbb{G}$; see Andersson et al. [3] for a discussion of the underlying Markov property.

Since a characterization of essential graphs is available [2, 18], equation (4) can be used to carry out structural learning in the space of essential graphs [5]. This leads to learning a graph where edges are arrows if data permit and lines if data do not permit, which is a fair graphical summary of the information contained in the data. In case interventional data are also available, that is, data collected after intervening on same variables, it may be possible to learn more arrows [4, 10], but an essential graph is the best possible output, in terms of learning arrows, when the analysis is based on observational data (as opposed to interventional data).

3 Learning Method

Let $\mathbf{Y}_{[k]}$ be the $n_k \times q$ data matrix obtained observing $\mathbf{Y}$ in group k, for $k = 1, \ldots, K$, and obtain the full data matrix $\mathbf{Y}_{[1:K]}$ by stacking $\mathbf{Y}_{[1]}, \ldots, \mathbf{Y}_{[K]}$ on top of one another. Then, conditional on a multiple essential graph $\mathbb{G}_{[1:K]} = (\mathbb{G}_{[1]}, \ldots, \mathbb{G}_{[K]})$, assuming independence across groups (as well as within groups) and Gaussian observations, the likelihood, constrained by (4), can be written as

$$f_{\mathbb{G}_{[1:K]}}(\mathbf{Y}_{[1:K]} \mid \boldsymbol{B}_{[1:K]}, \boldsymbol{\Omega}_{[1:K]}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} \prod_{\tau \in \mathcal{T}_k} f_{\mathbb{G}_{[k]}}(\mathbf{y}_{[k]i\tau} \mid \mathbf{y}_{[k]i\mathrm{pa}(\tau)}, \boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau}) \quad (5)$$

where $\mathcal{T}_k$ is the family of all chain components of $\mathbb{G}_{[k]}$, $\mathbf{y}_{[k]i}$ is the i-th row of $\mathbf{Y}_{[k]}$, $\mathbf{B}_{[k]\tau}$ is a $\{|\mathrm{pa}(\tau)| + 1\} \times |\tau|$ matrix of regression coefficients (including intercepts), $\boldsymbol{\Omega}_{[k]\tau}$ is a $|\tau| \times |\tau|$ precision matrix constrained by the lacking lines between nodes in the chain component τ of $\mathbb{G}_{[k]}$, and $f_{\mathbb{G}_{[k]}}(\mathbf{y}_{[k]i\tau} \mid \mathbf{y}_{[k]i\mathrm{pa}(\tau)}, \boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau})$ denotes the $|\tau|$-dimensional Gaussian density that regresses $\mathbf{y}_{[k]i\tau}$ on $\mathbf{y}_{[k]i\mathrm{pa}(\tau)}$.

We eliminate the nuisance parameters $\boldsymbol{B}_{[k]\tau}$ and $\boldsymbol{\Omega}_{[k]\tau}$ from (5) by assigning their parameter priors $\pi_{\mathbb{G}_{[k]}}(\boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau})$ independently over k and τ, so that the resulting marginal likelihood also factorizes over k and τ, and the posterior probability for the multiple essential graph (parameter of interest) can be written as

$$\Pr(\mathbb{G}_{[1:K]} \mid \mathbf{Y}_{[1:K]}) \propto \Pr(\mathbb{G}_{[1:K]}) \prod_{k=1}^{K} \prod_{\tau \in \mathcal{T}_k} m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau} \mid \mathbf{Y}_{[k]\mathrm{pa}(\tau)}) \quad (6)$$

where $\Pr(\mathbb{G}_{[1:K]})$ is the corresponding prior probability (specified below), $\mathbf{Y}_{[k]\tau}$ is the submatrix of $\mathbf{Y}_{[k]}$ formed by the columns of $\mathbf{Y}_{[k]}$ indexed by τ, and the quantity $m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau} \mid \mathbf{Y}_{[k]\mathrm{pa}(\tau)}) = \int \pi_{\mathbb{G}_{[k]}}(\boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau}) \prod_{i=1}^{n_k} f_{\mathbb{G}_{[k]}}(\mathbf{y}_{[k]i\tau} \mid \mathbf{y}_{[k]i\mathrm{pa}(\tau)}, \boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau})$ is available in closed form [6] for the objective parameter priors of Consonni et al. [7].

We specify the prior probability of $\mathbb{G}_{[1:K]}$ in two steps: we first use the prior of Peterson et al. [17] for the skeletons of $\mathbb{G}_1, \ldots, \mathbb{G}_K$, so that we encourage similarity among them and control their sparsity; we then assume, for the sake of simplicity, that all multiple essential graphs with given skeletons are equally probable. The prior of Peterson et al. [17] is a Markov random field depending on i) a vector $\boldsymbol{\nu}$ of sparsity parameters (one per each pair of nodes) and ii) a symmetric matrix $\boldsymbol{\Theta}$ of pairwise association parameters (one per each pair of groups). The prior for $\boldsymbol{\nu}$ controls sparsity, while the prior for $\boldsymbol{\Theta}$ encourages similarity, but the data are free to suggest which nodes are joined by an edge and which groups are not similar.

We target the joint posterior distribution of $\mathbb{G}_{[1:K]}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\nu}$ with a Markov chain of Metropolis-Hastings type, which we marginalize to approximate the posterior distribution (6) on the set of all multiple essential graphs. As a point estimate from this distribution, we resort to the *projected median probability graph model* [5]. We refer the reader to Castelletti et al. [6] for details, as well as for simulations validating the method and some promising results on real data (including those of Kornblau et al. [11] presented in the Introduction).

Acknowledgements Luca La Rocca received financial support from the FAR IMPULSO 2020 grant of the University of Modena and Reggio Emilia.

References

1. Altomare, D., Consonni, G., La Rocca, L.: Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. Biometrics **69**, 478–487 (2013)
2. Andersson, S.A., Madigan, D., Perlman, M.D.: A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics **25**, 505–541 (1997)
3. Andersson, S.A., Madigan, D., Perlman, M.D.: Alternative Markov properties for chain graphs. Scandinavian Journal of Statistics **28**, 33–85 (2001)
4. Castelletti, F., Consonni, G.: Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. The Annals of Applied Statistics **13**, 2289–2311 (2019)
5. Castelletti, F., Consonni, G., Della Vedova, M., Peluso, S.: Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. Bayesian Analysis **13**, 1235–1260 (2018)
6. Castelletti, F., La Rocca, L., Peluso, S., Stingo, F.C., Consonni, G.: Bayesian learning of multiple directed networks from observational data. Statistics in Medicine **39**, 4745–4766 (2020)
7. Consonni, G., La Rocca, L., Peluso, S.: Objective Bayes covariate-adjusted sparse graphical model selection. Scandinavian Journal of Statistics **44**, 741–764 (2017)
8. Drton, M., Maathuis, M.H.: Structure learning in graphical modeling. Annual Review of Statistics and Its Application **4**, 365–393 (2017)
9. Evans, R.: Markov properties for mixed graphical models. In: Handbook of Graphical Models, pp. 39–60. CRC Press (2019)
10. Hauser, A., Bühlmann, P.: Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **77**, 291–318 (2015)
11. Kornblau, S.M., Tibes, R., Qiu, Y.H., Chen, W., Kantarjian, H.M., Andreeff, M., Coombes, K.R., Mills, G.B.: Functional proteomic profiling of AML predicts response and survival. Blood **113**, 154–164 (2009)
12. La Rocca, L., Castelletti, F., Peluso, S., Stingo, F.C., Consonni, G.: Bayesian learning of multiple essential graphs. In: Book of Short Papers SIS 2020, pp. 447–452. Pearson (2020)
13. La Rocca, L., Roverato, A.: Discrete graphical models and their parameterization. In: Handbook of Graphical Models, pp. 191–216. CRC Press (2019)
14. Mukherjee, S., Oates, C.: Graphical models in molecular systems biology. In: Handbook of Graphical Models, pp. 497–512. CRC Press (2019)
15. Ni, Y., Baladandayuthapani, V., Vannucci, M., Stingo, F.C.: Bayesian graphical models for modern biological applications. Statistical Methods & Applications **Online first**, 1–29 (2021). With discussion
16. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
17. Peterson, C., Stingo, F.C., Vannucci, M.: Bayesian inference of multiple Gaussian graphical models. Journal of the American Statistical Association **110**, 159–174 (2015)
18. Roverato, A.: A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs. Scandinavian Journal of Statistics **32**, 295–312 (2005)
19. Studený, M.: Conditional independence and basic Markov properties. In: Handbook of Graphical Models, pp. 3–38. CRC Press (2019)
20. Verma, T., Pearl, J.: Equivalence and synthesis of causal models. In: Uncertainty in Artificial Intelligence 6, pp. 255–270. Elsevier Science Publishers (1991)

4 Contributed Sessions

Applications in Machine Learning

A neural network approach to survival analysis with time-dependent covariates for modelling time to cardiovascular diseases in HIV patients

Un approccio basato su reti neurali per l'analisi di sopravvivenza con variabili tempo-dipendenti per predire il tempo all'evento cardiovascolare in pazienti affetti da HIV

Federica Corso, Agostino Lurani Cernuschi, Laura Galli, Chiara Masci, Camilla Muccini, Anna Maria Paganoni and Francesca Ieva

Abstract In this study, we investigate the impact of AntiRetroviral Therapies (ART) and clinical features of HIV patients on the risk of a cardiovascular event, with a survival analysis approach. Clinical and demographic features of 4512 HIV patients are collected during a follow-up time of 15 years. Two different methods for survival analysis are adopted: the classical Cox Proportional Hazard model and a deep-learning approach, DeepHit, based on neural-network. Models are implemented including both time-invariant and time-varying covariates. Results of the two methods are compared in terms of predictive performance and interpretability, with a special focus on the different potentialities of the two approaches.

Abstract *In questo studio, analizziamo l'impatto delle Terapie AntiRetrovirali (ART) e delle caretteristiche cliniche di pazienti affetti da HIV sul rischio di evento cardiovascolare, applicando metodi di analisi di sopravvivenza. Il dataset è composto da 4512 pazienti affetti da HIV di cui sono note le informazioni cliniche e demografiche. Per l'analisi di rischio cardiovascolare adoperiamo due approacci: il classico modello di Cox a rischi proporzionali e il DeepHit basato su reti neurali. I modelli includono sia covariate tempo-invarianti che tempo-dipendenti, monitorate lungo il follow-up. I risultati dei modelli sono confrontati sia in termini di capacità predittiva che di interpretabilità, con una particolare attenzione alle diverse potenzialità dei metodi confrontati.*

Key words: Survival analysis, Neural Network, DeepHit, Time-Dependent variables, HIV, Cardiovascular Disease

---

Corso F., Politecnico di Milano, e-mail: federica.corso@polimi.it
Lurani Cernuschi A., Politecnico di Milano, e-mail: agostino.lurani@mail.polimi.it
Galli L., IRCSS San Raffaele Scientific Institute, e-mail: galli.laura@hsr.it
Masci C., Politecnico di Milano, e-mail: chiara.masci@polimi.it
Muccini C., IRCSS San Raffaele Scientific Institute, e-mail: muccini.camilla@hsr.it
Paganoni A.M., Politecnico di Milano, e-mail: anna.paganoni@polimi.it
Ieva F., Politecnico di Milano, e-mail: francesca.ieva@polimi.it

1 Introduction

The Antiretroviral Therapy (ART) has allowed people affected by Human Immunodeficiency Virus (HIV) to live longer, especially with the introduction of drugs such as Protease Inhibitors (PIs) combined with Nucleoside Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), with a decrease of 60% in the expected rate of death [8]. Recently, the link between ART and CardioVascular Diseases (CVDs) has been explored [3] and some results have highlighted an association between the exposure to these drugs and the risk of CVD events [1]. In this work, we analyse the occurrence of a CVD event in 4512 patients affected by HIV, with a follow-up of 15 years from the beginning of the ART. Patients data are collected at each medical visit, producing a longitudinal dataset that tracks patients information through time. Given the complexity of the dataset, we adopt two different survival analysis approaches, the Cox Proportional Hazard (PH) model [2, 4] and DeepHit [6, 7], including both variables measured at the baseline, i.e., at the ART start, and during the follow-up, i.e., longitudinal data. The goal of this work is twofold: to investigate the relationship between ART and other patient's characteristics with the risk of a CVD event over time and to compare a classical survival method with an advanced machine learning one in which the assumptions of linear relationships and of proportional hazards are relaxed. To the best of our knowledge, the applications of time-dependent neural networks for survival analysis are still very rare in the literature and part of our work consists in the exploration of the potentialities of these methods in modelling and predicting time-to-event data.

2 HIV patients' data

A cohort of 4512 people living with HIV infection, attending the IRCCS Ospedale San Raffaele for routine care, who started ART since 1998 were included in the present study. For each patient, medical information was registered including demographic variables (e.g., sex, race, age, etc.), laboratory parameters (e.g., viremia, cholesterol, etc.) and time-exposure to ART drugs. 24 variables are selected according to clinician's experience and proportion of missing data. Specifically, our data include five variables describing the ART: four of them regard the cumulative time of exposure to drugs, each classified by the stage of the cycle they inhibit, and the last one regards the year of ART start, treated as categorical variable with 2007 as cut-off. Among the 4512 patients, only 90 (2%) experienced a CVD event within 15 years of follow-up.

3 Methods

Survival analysis is a common method used to predict the time-to-event, e.g., the time until a CVD event happens. The target variable is defined as the couple of the survival time $T_i = min(T_i^*, C_i)$ and the censoring indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$, for each observation $i = 1, \ldots, N$, where C_i is the censoring time and T_i^* is the CVD event time, if any.

3.1 The Cox Proportional Hazard model

Cox Proportional Hazard model is a semi-parametric method that models the association between covariates and the event risk, through time. The Cox PH model assumes the hazard function for an individual i to be described as follows:

$$h_i(t|\mathbf{x}_i) = h_0(t)e^{\mathbf{x}_i^T \boldsymbol{\beta}} \tag{1}$$

where $h_0(t)$ is the baseline hazard function, $\mathbf{x}_i^T$ is the covariates vector for the i^{th} individual and $\boldsymbol{\beta}$ is the vector of the regression coefficients. The parameter estimation is obtained by maximizing the *partial likelihood*, that considers only the probabilities for the patients who experience the event. The Cox PH model has two strong assumptions: the hazard ratios are assumed to be constant over time (proportional hazards) and the relationship between the *log* hazard and the covariates is assumed to be linear. Cox PH model can also handle time-dependent covariates. In this case, the hazard function at time t depends on the time-invariant covariates and on the values of the time-dependent covariates only at time t:

$$h_i(t|\mathbf{x}_i(t)) = h_0(t)e^{\left[\sum_{k=1}^{P_{fix}} x_{ik}\beta_k + \sum_{z=1}^{P_{td}} x_{iz}(t)\delta_z\right]} \tag{2}$$

where $\mathbf{x}_i(t)$ is the vector of covariates of the i^{th} patient, including P_{fix} fixed variables with coefficient vector $\boldsymbol{\beta}$ and P_{td} time-dependent variables, assuming different values at each time t, with coefficient vector $\boldsymbol{\delta}$. It is worth to underline that the coefficients are assumed to be time-invariant.

3.2 DeepHit: Neural-network based approach to survival analysis

DeepHit [7] is a Neural-Network (NN) based model that is able to capture complex patterns such as interactions among variables and non-linear relationships between the covariates and the outcome. DeepHit can handle k competing risks and is structured by feed-forward blocks such as a shared sub-network and k cause-specific networks. The shared sub-network captures the patterns of the relationships

between covariates and the time-to-event that are common to all different k events. Each cause-specific network captures the features of a specific event. The cause-specific cumulative incidence function expresses the probability that a particulat event $k^* \in K$ occurs on or before time t^* conditional on the covariates $\mathbb{X}^*$. The parameter estimation is based on the minimization of the weighted sum of two loss functions, *L1* and *L2*. *L1* is the log-likelihood of the joint distribution modified for right censored data, while *L2* is the ranking loss that incorporates the estimated cumulative incidence functions to finetune the network to each cause-specific event. The extension of the DeepHit structure for longitudinal data [6] is a recurrent NN where the shared block is replaced with a shared recurrent NN and the k specific networks. In this case, a loss function, that handles time-dependent variables and right censored data, is introduced in the shared NN. In particular, it is defined as the sum of three sub-loss functions, including *L3* loss function that makes predictions on the one step-ahead covariate x.

4 Results and model evaluation

We apply Cox PH method and Deephit method, both at baseline, i.e. at the ART start, and at the end of the follow-up, by including time-dependent covariates, to model the time to a CVD event in HIV patients, within 15 years from the baseline. For each method, full (24 covariates) and reduced models are implemented. DeepHit results are interpreted by means of the Permutation Feature Importance (PFI) and the Shapley value Additive Explanation [5]. Results are compared in terms of interpretability and predictive power (on training and test sets).

4.1 Cox and DeepHit models at baseline

At baseline, the full Cox PH model enables to understand which are the significant variables and how they contribute in the prediction of the time-to-CVD. The Hazard Ratios of the Cox model allow to deduce that an hypertensive patient has a risk of a CVD event 2.74 times higher than a non-hypertensive patient. Regarding the protective factors, a patient who started the ART before 2007 is less likely to experience a CVD event. Results of DeepHit are less quantitatively interpretable but, by using Shapley values, we can understand how the covariates interact on the risk of having a CVD event. By looking at the Shapely values, we observe that hypertension is again a risk factor and the year of ART start is still a protective factor. The Cox model predicts with a C-index of 0.66 on the test set while the DeepHit reaches 0.74 on the test set. Given the low percentage of observations with CVD event and the high number of covariates, we proceed with models reduction to obtain more robust performances. While the Cox models is reduced with a one-step procedure based on the Hazard Ratios, the feature selection for the DeepHit model is computed by

adopting Shapley values of the variables in Fig.1. Some interesting behaviors are highlighted by the DeepHit model: a short exposure to NRTIs increases the risk of CVD events while a medium-long exposure decreases it. Moreover, Shapley values allow to visualize interactions between couples of covariates: if we look at the interaction between exposure to NRTIs and patient's age (Fig. 2), we observe that the effect of therapies is stronger in old patients both in increasing the risk (for short exposure) and in decreasing it (for long exposure). The reduced Cox model achieves slightly better performance on the test set (C-index 0.69) while the reduced DeepHit model is similar to the complete model.

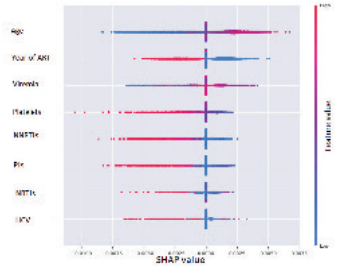


Fig. 1 The Shapley value importance of the reduced DeepHit model at the baseline.

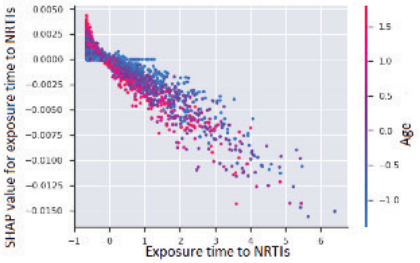


Fig. 2 The Shapley value dependency on the time of exposure to NRTIs as the age of the patients varies for the reduced DeepHit model at the baseline.

4.2 Time-dependent Cox and dynamic DeepHit models

The results of the time-dependent Cox model are interpreted through the Hazard Ratios and, again, the most significant variables are the year of ART start and hypertension. Since the computational cost of Shapaley values for logitudinal data is prohibitive, we measure the contribution of each covariate in the dynamic DeepHit with the PFI using the C-index on the training set. The NNRTIs and PIs drugs are the most important variables differently from the time-dependent Cox model. This is possibly due to DeepHit ability of capturing non-linear relationships. The dynamic DeepHit reaches a C-index of 0.77 while the Cox model achieves a lower C-index on the test set. We identify from the reduced time-dependent Cox model that the time to exposure to NNRTIs is a protective factor and the reduced dynamic DeepHit confirms the importance of the covariates NNRTIs and PIs. These results underline that time to exposure to drugs, which is an information that varies over time, remains the most important features for the prediction in both the full and the reduced models. The time-dependent Cox model with reduced covariates obtains a C-index of 0.70 in the test set while the Dynamic DeepHit reduced model does not perform as well as the complete model.

5 Conclusions

So far, the relationship between ART and CVD events was approached with classical regression and classification methods and only few studies adopted survival analysis tools or machine learning algorithms. In this study, the classic Cox PH model is compared with DeepHit on a real dataset with fixed and time-dependent variables. Specifically, the DeepHit models show that the time of exposure to ART drugs is associated with lower risk of CVD event and only under short exposure the risk might increase. Both the models find that the NNRTI drugs are protective factors even if the correlation among ART drugs might mask the effect of other variables. Moreover, patients who started the ART before 2007 are found to be less exposed to CVD risk. This work point out that in long term analysis the influence of ART drugs decreases the risk of CVD events. Better performances are achieved by including longitudinal data and in particular by the dynamic DeepHit that allows to analyse the relationship between covariates and time to CVD event in a more flexible way.

References

[1] Bavinger, C., Bendavid, E., Niehaus, K., Olshen, R. A., Olkin, I., Sundaram, V., Wein, N., Holodniy, M., Hou, N., Owens, D. K., et al. (2013). Risk of cardiovascular disease from antiretroviral therapy for hiv: a systematic review. *PloS one*, 8(3):e59551.

[2] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[3] Currier, J. S. (2020). Management of cardiovascular risk (including dyslipidemia) in patients with hiv. *UpToDate, Waltham, MA (Accessed on January 15, 2020.)*.

[4] Kleinbaum, D. G. and Klein, M. (2015). *Survival Analysis*. Springer, 3 edition.

[5] Knuth and E., D. Explain any models with the shap values — use the kernelexplainer.

[6] Lee, C., Yoon, J., and van der Schaar, M. (2020). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering (TBME)*.

[7] Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. *Conference on Artificial Intelligence (AAAI)*.

[8] Pillay-van Wyk, V., Msemburi, W., Dorrington, R., Laubscher, R., Groenewald, P., and Bradshaw, D. (2019). Hiv/aids mortality trends pre and post art for 1997-2012 in south africa–have we turned the tide? *South African Medical Journal*, 109(11b):41–44.

Analyzing the Correlation Structure of Financial Markets Using a Quantile Graphical Model

Analisi della struttura di correlazione dei mercati finanziari usando un modello grafico quantile

Beatrice Foroni, Luca Merlo and Lea Petrella

Abstract In this paper we develop a quantile graphical model to identify the tail conditional correlation structure in multivariate data across different quantiles of the marginal distributions of the variables of interest. To implement the procedure, we consider the Multivariate Asymmetric Laplace distribution and exploit its location-scale mixture representation to build a penalized EM algorithm for estimating the sparse precision matrix of the distribution by means of an L_1 penalty. The empirical application is performed on a set of market indexes, cryptocurrencies and commodities.

Abstract *In questo articolo sviluppiamo un modello grafico quantile per identificare la struttura di correlazione condizionata di coda attraverso lo studio dei quantili delle distribuzioni marginali delle variabili di interesse. Per implementare la procedura, consideriamo la distribuzione di Laplace asimmetrica multivariata e sfruttiamo la sua rappresentazione a mistura per costruire un algoritmo EM penalizzato per la stima della matrice di precisione sparsa della distribuzione mediante una penalità L_1. La metodologia presentata viene applicata sui rendimenti finanziari dei principali indici di mercato, criptovalute e materie prime.*

Key words: EM Algorithm, Cryptocurrencies, Graphical Models, Multivariate Asymmetric Laplace Distribution

Beatrice Foroni
MEMOTEF Department, Sapienza University, e-mail: beatrice.foroni@uniroma1.it

Luca Merlo
Department of Statistical Sciences, Sapienza University, e-mail: luca.merlo@uniroma1.it

Lea Petrella
MEMOTEF Department, Sapienza University, e-mail: lea.petrella@uniroma1.it

1 Introduction

In recent years, the urge to identify how the impact of financial stress events can spread across the whole financial global system has made network science a useful tool for describing the propagation of systemic risk. Within this literature, Gaussian Graphical Models (GGM) have received an enormous attention because they provide a simple method to model the pair-wise conditional dependence structure of a collection of stochastic variables. In a high-dimensional framework, when a large set of random variables is considered, we are interested in identifying only a smaller subset of variables that exhibits the most relevant and strongest dependencies. Among the several methods proposed in literature, there is the *Graphical LASSO* (*glasso*) algorithm of [4], which maximizes the likelihood of the model penalized by the L_1-norm of the elements of the precision matrix. However, several empirical studies show that financial returns exhibit most of the well known stylized facts like fat tails, leptokurtosis and skewness, and deviation from normality makes it harder to characterize conditional dependence structures. The literature regarding non-Gaussian graphical models is fairly limited. In this context, the work of [3] provides a tool for robust model selection using multivariate *t*-distributions to model the data. Moreover, these proposals are not able to recover the dependencies in the tails of the distribution. Be able to understand and focus on specific part of a distribution such as the tails can really improve the knowledge in areas like financial contagion and systemic risk, where the dynamic of extreme events is of utmost importance. In this paper we develop a quantile graphical model to estimate the conditional tail correlation structure in multivariate data at different quantile levels of interest, without relying on the restrictive assumption of normally distributed data. In order to model the conditional correlation structure of multiple random variables at quantile-specific indices, we generalize the work of [7], which consider a reparametrization of the Multivariate Asymmetric Laplace (MAL) distribution of [6] to jointly model conditional quantiles of multiple random variables in a likelihood framework. Following [3], we demonstrate that the precision matrix of the MAL distribution completely characterizes the conditional dependence structure among the random variables at each quantile level, and allows us to construct a graph whose edges correspond to relations of conditional dependency. To induce sparsity in the precision matrix, we exploit the Gaussian location-scale mixture of the MAL and apply the *glasso* algorithm. In particular, following [5], we build a suitable Penalized EM (PEM) algorithm based on the maximization of the likelihood of the model penalized by the L_1-norm of the off-diagonal elements of the

precision matrix. The estimated networks can be analyzed with respect to centrality measures as functions of the quantile level. The relevance of our approach is shown empirically among the cryptocurrency, commodity and stock market sectors from 2017 to 2021, and the modeling approach we propose is able to identify the connectedness as more serious levels of distress are considered, and can describe the topological structure of the underlying graph at different quantile levels of interest.

2 Model Specification

Given p quantile indexes $\tau = [\tau_1, \ldots, \tau_p]'$, with $\tau_j \in (0,1)$, for $j = 1, \ldots, p$, let $\mathbf{Y}_t = [Y_t^{(1)}, \ldots, Y_t^{(p)}]$ denote a continuous p-dimensional random vector for $t = 1, \ldots, T$. Generalizing the approach of [7], our objective is to develop a quantile graphical model for learning the conditional tail dependence structure among the components of $\mathbf{Y}_t$ at different quantile levels of interest τ. Specifically, we employ the MAL distribution, $\mathscr{MAL} \sim (\mu, \mathbf{D}\tilde{\xi}, \mathbf{D}\Sigma\mathbf{D})$, (see [6]) as:

$$f_{\mathbf{Y}}(\mathbf{y}_t) = \frac{2\exp\left\{(\mathbf{y}_t - \mu)'\mathbf{D}^{-1}\Sigma^{-1}\tilde{\xi}\right\}}{(2\pi)^{p/2}|\mathbf{D}\Sigma\mathbf{D}|^{1/2}} \left(\frac{\tilde{m}_t}{2+\tilde{d}}\right)^{\nu/2} K_\nu\left(\sqrt{(2+\tilde{d})\tilde{m}_t}\right), \quad (1)$$

where μ is the location parameter, $\mathbf{D}\xi \in \mathscr{R}^p$ is the scale (or skew) parameter, with $\mathbf{D} = \text{diag}[\delta_1, \delta_2, \ldots, \delta_p]$, $\delta_j > 0$ and $\xi = [\xi_1, \xi_2, \ldots, \xi_p]'$, having generic element $\xi_j = \frac{1-2\tau_j}{\tau_j(1-\tau_j)}$. Σ is a $p \times p$ positive definite matrix such that $\Sigma = \Lambda\Psi\Lambda$, with Ψ being a correlation matrix and $\Lambda = \text{diag}[\sigma_1, \sigma_1, \ldots, \sigma_p]$, with $\sigma_j^2 = \frac{2}{\tau_j(1-\tau_j)}$, $j = 1, \ldots, p$. Finally, $\tilde{m}_t = (\mathbf{y}_t - \mu)'(\mathbf{D}\Sigma\mathbf{D})^{-1}(\mathbf{y}_t - \mu)$, $\tilde{d} = \tilde{\xi}'\Sigma^{-1}\tilde{\xi}$, and $K_\nu(\cdot)$ denotes the modified Bessel function of the third kind with index parameter $\nu = (2-p)/2$. One of the key benefits of the MAL distribution is that, using (1) and following [6], the $\mathscr{MAL} \sim (\mu, \mathbf{D}\tilde{\xi}, \mathbf{D}\Sigma\mathbf{D})$ admits the following location-scale mixture representation:

$$\mathbf{Y} = \mu + \mathbf{D}\tilde{\xi}W + \sqrt{W}\mathbf{D}\Sigma^{1/2}\mathbf{Z} \quad (2)$$

where $\mathbf{Z} \sim \mathscr{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ denotes a p-variate Normal distribution and $W \sim \text{Exp}(1)$ has a standard Exponential distribution, with $\mathbf{Z}$ being independent of W. Hence, the mixture representation in (2) represents the generating process of a MAL random vector $\mathbf{Y}$ from a latent Gaussian random vector $\mathbf{Z}$ with correlation matrix Ψ and a single latent Exponential variable with mean 1. In particular, the constraints imposed on $\tilde{\xi}$ and Λ represent necessary conditions for model identifiability for any fixed quantile level $\tau_1, \ldots, \tau_p$ and guarantee that $\mu^{(j)}$ is the τ_j-th quantile of $Y_t^{(j)}$, for

$j = 1, \ldots, p$.

To build the graphical model, let $G = (V, E)$ be an undirected graph where $V = \{1, \ldots, p\}$ is the set of nodes, such that each component of the random variable $\mathbf{Y}_t$ corresponds to a node in V, and $E \subseteq V \times V$ represents the set of undirected edges. In order to study the conditional dependence structure of $\mathbf{Y}_t$ through the graph G, we exploit the MAL representation in (2). For notational convenience and to illustrate the similarities with the GGM, we define the precision matrix $\mathbf{K} = \Psi^{-1}$. Following the t-distribution graphical model approach in [3], we establish the following proposition.

Proposition 1. *For a fixed p-dimensional vector of quantile levels* $\tau = [\tau_1, \tau_2, \ldots, \tau_p]'$ *such that* $\tau_j \in (0,1)$*, for* $j = 1, \ldots, p$*, let* $\mathbf{Y} \sim \mathscr{MAL}(\mu, \mathbf{D}\tilde{\xi}, \mathbf{D}\Lambda\mathbf{K}^{-1}\Lambda\mathbf{D})$*. If two nodes* j *and* k*, with* $j,k \in V$ *and* $j \neq k$*, of the graph are separated by a set of nodes* $C \in V$*, then* $\mathbf{Y}^{(j)}$ *and* $\mathbf{Y}^{(k)}$ *are conditionally uncorrelated given* $\mathbf{Y}^{(C)}$*.*

The proof of Proposition 1 follows directly from the mixture representation of the MAL in (2) and the closure property of the Normal distribution under conditioning of its components. Most importantly, from Proposition 1 follows that the zero entries in the precision matrix $\mathbf{K}$ imply the conditional uncorrelation between the components of $\mathbf{Y}_t$ at each given quantile level τ. To estimate and make inference on the model parameters we develop a suitable Expectation-Maximization (EM) algorithm, which exploits the mixture representation of the MAL distribution, treating W as missing data. In order to identify only a smaller subset of variables that exhibit the most relevant and strongest dependencies, we construct a PEM algorithm by adding an L_1-norm penalty of the off-diagonal elements of $\mathbf{K}$ to the likelihood of the model. Specifically, for a given vector $\tau = [\tau_1, \tau_2, \ldots, \tau_p]'$, the penalized complete log-likelihood function is proportional to:

$$\ell_c(\Phi_\tau) \propto \frac{T}{2}\log|\mathbf{D}^{-1}\mathbf{K}\mathbf{D}^{-1}| - \frac{T}{2}\mathrm{tr}\{\mathbf{K}\mathbf{S}\} - \rho||\mathbf{K}||_1 \tag{3}$$

with

$$\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{W_t}\Lambda^{-1}\mathbf{D}^{-1}(\mathbf{Y}_t - \mu - \mathbf{D}\xi W_t)(\mathbf{Y}_t - \mu - \mathbf{D}\xi W_t)'\mathbf{D}^{-1}\Lambda^{-1} \tag{4}$$

and where W_t is an Exponential random variable with mean 1.

As it can be noticed, the likelihood function in (3) is convex in $\mathbf{K}$. Therefore, at each iteration of the PEM, this feature allows us to adopt the *glasso* algorithm for efficient estimation of the sparse precision matrix $\mathbf{K}$.

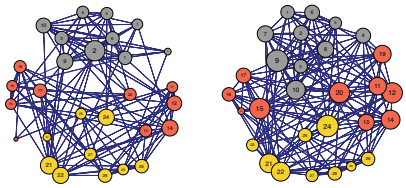


Fig. 1 Graphs for $\tau = 0.50$ (left) and $\tau = 0.95$ (right). Yellow, grey and red nodes represent respectively indexes, cryptocurrencies and commodities while the vertex labels are illustrated in Table 1.

3 Main Results and Conclusions

The empirical analysis is performed using the R software on the log-returns of 29 financial assets comprising stock market indices, commodity futures and digital currencies from September 14, 2017 and to June 21, 2021. We set $\tau = \tau_j$, $j = 1,\dots,29$, and fit the proposed model for a sequence of 99 quantile levels $\tau = [0.01, 0.02, \dots, 0.98, 0.99]'$. Then, for each τ, we construct the corresponding graph G_τ.

Sectors					
Cryptos		Commodities		Stock Index	
1: Bitcoin	**6: Litecoin**	**11: Gold**	**16: Brent**	**21: S&P 500**	**26: Dax**
2: Ethereum	**7: Binance Coin**	**12: Silver**	**17: Gasoline**	**22: Dow Jones**	**27: Ftse Mib**
3: Ripple	**8: Eos**	**13: Palladium**	**18: Heating oil**	**23: Nasdaq**	**28: Cac**
4: Tether	**9: Stellar**	**14: Platinum**	**19: Natural Gas**	**24: Shangai composite index**	**29: Euro Stoxx50**
5: Bitcoin Cash	**10: Tron**	**15: Crude Oil**	**20: Ethanol**	**25: Nikkei 225**	

Table 1 Financial sectors considered in the analysis. Numbers identify vertex labels in Figure 1.

In Figure 1 we represent the estimated graph at $\tau = 0.50$ and $\tau = 0.95$ to show how the density of the network changes between tranquil ($\widehat{G}_{0.50}$) and bullish periods ($\widehat{G}_{0.95}$), respectively. The size of the node is proportional to the degree centrality and the width of the edges is determined by the magnitude of the estimated correlations in $\widehat{K}$. A deeper analysis to describe how the interconnectedness and contagion risk change as a function of the quantile index τ, is conducted by showing in Figure 2 the edge density (left) and the modularity measure (right) as functions of τ. The edge density shows a highly interconnected system, even for the smallest ratio of value at the center of the distribution. It is evident a strongest dependency during bearish and bullish periods, as the edge density is the largest in the tails.The presence of community structures is summarized in the modularity measure plot of Figure 2.
Consistently with the estimated graph $\widehat{G}_{0.50}$ in Figure 1, during tranquil periods the reduced number of connections brings out heterogeneity in the distribution of

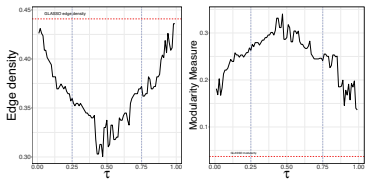


Fig. 2 From left to right, Edge density and Modularity measure as a function of τ. Blue dotted lines identify the 25th and 75th percentiles to mark respectively crisis and bullish markets periods. Red dotted line identify the centrality measure associated with the *glasso*.

edges, i.e., high concentration of edges within groups of nodes and low concentrations between groups. This behavior can be explained by the so-called co-explosion of cryptoassets already discussed by [1],which evidences that price explosivity in one cryptocurrency can lead to explosivity in other cryptocurrencies. In conclusion, with our approach we are able to recover valuable information at each quantile level even without the assumption of normality. The whole analysis conveys a highly connected network which becomes even more dense during bearish and bullish markets periods, and the results are in line with existing studies ([2, 1]). With this model we strengthen the existing literature in this field, implementing a technique to adjust the *glasso* algorithm to a quantile structure of dependence.

References

1. Bouri, E., Shahzad, S.J.H., Roubaud, D.: Co-explosivity in the cryptocurrency market. Finance Research Letters **29**, 178–183 (2019)
2. Demirer, M., Diebold, F.X., Liu, L., Yilmaz, K.: Estimating global bank network connectedness. Journal of Applied Econometrics **33**(1), 1–15 (2018)
3. Finegold, M., Drton, M.: Robust graphical modeling of gene networks using classical and alternative t-distributions. The Annals of Applied Statistics pp. 1057–1080 (2011)
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. Biostatistics **9**(3), 432–441 (2008)
5. Green, P.J.: On use of the EM algorithm for penalized likelihood estimation. Journal of the Royal Statistical Society: Series B (Methodological) **52**(3), 443–452 (1990)
6. Kotz, S., Kozubowski, T., Podgorski, K.: The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Springer Science & Business Media (2012)
7. Petrella, L., Raponi, V.: Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. Journal of Multivariate Analysis **173**, 70–84 (2019)

Neural Network for statistical process control of a multiple stream binomial process with an application to HVAC systems in passenger rail vehicles

Rete neurale per il controllo statistico di un processo binomiale a flussi multipli con un'applicazione per sistemi HVAC nei veicoli ferroviari a trasporto passeggeri

Gianluca Sposito, Antonio Lepore, Biagio Palumbo, Giuseppe Giannini

Abstract A multiple stream binomial process (MSBP) is a process at a point in time that generates several output streams that can be modelled as binomial processes. A multilayer perceptron is embedded in a control charting scheme and is trained to classify whether each sample is drawn from the in-control process distribution or not. The performance of the proposed scheme is evaluated through a wide Monte Carlo simulation and is compared with the Wludyka and Jacobs's MSBP control charts in terms of out-of-control average run length. Finally, the practical applicability of the proposed approach is illustrated through real operational data from heating, ventilation and air conditioning (HVAC) systems installed on-board modern passenger trains.

Abstract *Un processo binomiale a flussi multipli (MSBP) è un processo descritto nel tempo da flussi concorrenti che possono essere modellati come processi binomiali. Viene presentato uno schema di controllo statistico di un MSBP che integra un percettrone multistrato addestrato a classificare se i dati raccolti siano generati da un processo in controllo o fuori controllo statistico. Le prestazioni dell'approccio proposto sono valutate mediante un'estesa simulazione Monte Carlo e confrontate con le carte di controllo per MSBP di Wludyka e Jacobs, in termini di lunghezza media delle sequenze fuori controllo. Infine, la pratica applicabilità dell'approccio proposto viene illustrata attraverso un caso studio reale su dati operativi provenienti da sistemi di riscaldamento, ventilazione e condizionamento dell'aria, installati a bordo di moderni veicoli ferroviari.*

Gianluca Sposito, Antonio Lepore, Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy, e-mail: gianluca.sposito@unina.it; antonio.lepore@unina.it; biagio.palumbo@unina.it

Giuseppe Giannini
Head of Operation Service and Maintenance Product Evolution, Hitachi Rail Group, Naples, Italy, e-mail: giuseppe.giannini@hitachirail.com

Key words: Artificial neural network, Statistical process control, Multiple stream binomial process, Railway HVAC system

1 Introduction

In the railway transport sector, it is of paramount importance to ensure high comfort levels of passengers on board through an efficient regulation of the air conditioning. Recent European regulations, such as UNI EN 14750 [1], prescribe operational requirements for passenger rail coaches, in terms of air quality and comfort level, and the continuous monitoring of the heating, ventilation and air conditioning (HVAC) system installed on modern passenger railway vehicles. For this purpose, the high sensorization and digitalization of modern trains make available large amounts of data and create an urgent call for statistical models and methods able to transform this information into value. Typically, each passenger coach is equipped with a dedicated HVAC system, and the simultaneous statistical process control/monitoring (SPC/M) of the sensor signals coming from each HVAC data acquisition system installed on-board each coach can be regarded as a multiple stream process (MSP) [8]. In particular, referring to a MSP [6] control by attributes, a multiple stream binomial process (MSBP) is a process at a point in time that generates several output streams that can be modelled as binomial processes. From the SPC/M point of view, the quality of interest and its specifications are identical in all streams. Motivated by the industrial context and the latest attempts at the integration of NNs into classical control charting procedures [12, 7, 4], this paper aims to develop a new approach based on artificial neural networks (NNs) in the field of SPC/M of MSBPs with the twofold objective of detecting changes both in the individual stream means and the overall process mean.

In what follows, we present (i) the proposed control charting procedure by briefly framing MSBPs and NNs; (ii) the proposed NN training through a wide Monte Carlo simulation and its comparison with traditional Wludyka and Jacobs' Chi-squared [10] and group p [11] MSBP control charts; (iii) an application to the operational data acquired for the monitoring of HVAC systems installed on board of passenger railway vehicles, courtesy the rail transport company Hitachi Rail STS (https://www.hitachirail.com/).

2 Materials and Methods

Let us suppose that over a subgroup of N consecutive points at time t, the quality of interest is a variable counting the number of times a particular event occurs. Let Y_{tj} denote this counting variable for each stream $j = 1,...,s$ and time index $t = 1,...,T$. A MSBP assumes Y_{tj} distributed as independent binomial random variables $Y_{tj} \sim \text{Bin}(N, p_j)$ over time and streams, where p_j denotes the probability that

the event of interest occurs for the jth stream at each subgroup point. For an in-control (IC) homogeneous MSBP p_j is assumed, for each j, equal to the same constant $p \in (0,1)$, which therefore represents the overall process mean. The process is said to be out of control (OC) if an assignable cause affects (i) all of the streams uniformly (the overall p changes) or (ii) one or more streams differently (p_j changes for one or more streams). Inspired by the real-case study, for each stream, we generate pseudorandom observations of Y_{tj} from a binomial random variable with $s = 6$, $N = 30$, and $p = 0.52$. To the best of our knowledge, Wludyka and Jacobs are the only authors to have investigated the SPC/M of a MSBP.

In this paper, a multilayer perceptron (MLP) [3], which is one of the most widely used NN, is designed to be used for the SPC/M of a MSBP, that is, to classify a sample as drawn from an IC or OC process. Specifically, the MLP is trained such that, when a sample (called *negative sample*) is drawn from an IC process, the target value of the output neuron is coded as zero, otherwise, when a sample (called *positive sample*) is drawn from an OC process, it is coded as one. The MLP has s neurons in the input layer to represent the s monitored statistics Y_{tj} at time t, for each stream $j = 1,2,...,s$. In order to train and properly design the MLP, 300 positive samples are generated through a Monte Carlo simulation from all possible $\sum_{l=1}^{s} \binom{s}{l}$ OC scenarios in which $l = 1,\ldots,s$ streams shift off the target p at the same time by as much as $\Delta p = -1.0\sigma, -2.0\sigma, -3.0\sigma, 1.0\sigma, 2.0\sigma, 3.0\sigma$, where $\sigma = \sqrt{p(1-p)/N}$. To get a balanced data set, the total number of negative and positive samples must be equal. The MLP is trained through the back-propagation algorithm [3]. To compare different NN architectures, the area under the receiver operating characteristic (ROC) curve (AUC) [2] is used as a performance measure and is calculated on the validation set, obtained by retaining the 30% of the generated samples. Then, a MLP with one hidden layer with ten neurons, a rectified linear unit (ReLU) and a sigmoid activation function [3] for the hidden and output layer, respectively, is shown to have the best performance in terms of AUC. Furthermore, to allow a fair comparison with the traditional MSBP control charts, the NN must be tuned to a corresponding level of discriminating power [4], that is, the cut-off value (CV) of the output neuron must be properly fixed to set the Type-I and Type-II error rates of the resulting control charting scheme denoted by α and β, respectively. Or, equivalently, one may set the IC or OC average run lengths [6], respectively, defined as $ARL_0 = \frac{1}{\alpha}$ and $ARL_1 = \frac{1}{1-\beta}$. Further 100000 Monte Carlo simulations from an IC process are generated to set the CV to achieve an ARL_0 as close as possible to the desired one. By simulation, α is the proportion of the wrongly classified samples as drawn from an OC process. At a fixed ARL_0, we can evaluate the performance of the proposed approach in terms of ARL_1 by simulating 100000 data sets from each of the OC scenarios in which at least one stream of the MSBP shifts off-target at different mean shift sizes. By simulation, β is the proportion of negative samples that the MLP incorrectly classifies as drawn from an IC. The proposed NN based control charting procedure is compared with the competitor Wludyka and Jacobs' Chi-squared [10] and group p [11] MSBP control charts by means of the ARL_1, at fixed $ARL_0 = 20, 100, 358$, that is $\alpha = 0.05, 0.01, 0.0028$, respectively.

Table 1 Comparison of ARL_1 of the Chi-squared control chart [10] and of the proposed NN approach based on 100000 simulations of a MSBP with $s = 6$ streams, at different number $l = 1,...,s$ of streams that shift off the target and mean shift size $\Delta p = -2.0\sigma, -1.0\sigma, -0.5\sigma, 0.5\sigma, 1.0\sigma, 2.0\sigma$.

	NN based control charting procedure Δp						Chi-squared control chart Δp					
	-2.0σ	-1.0σ	-0.5σ	0.5σ	1.0σ	2.0σ	-2.0σ	-1.0σ	-0.5σ	0.5σ	1.0σ	2.0σ
l												
1	36.21	143.88	276.24	249.38	123.46	26.08	22.68	138.5	290.7	288.18	146.84	25.07
2	6.2	52.97	186.57	164.20	43.16	4.8	5.91	63.41	190.48	238.1	65.4	6.29
3	2.52	23.66	113.38	99.60	21.43	2.29	2.78	33.59	164.2	176.06	37.79	2.85
4	1.47	11.72	73.86	66.36	10.82	1.4	1.78	20.43	129.07	149.03	23	1.79
5	1.15	6.44	45.56	45.13	6.2	1.23	1.37	13.75	101.83	115.47	15.27	1.38
6	1.05	4.14	33.02	32.77	4.18	1.04	1.17	9.81	82.51	101.21	10.99	1.18

Table 2 Comparison of ARL_1 of the group p control chart [11] and of the proposed NN approach based on 100000 simulations of a MSBP with $s = 6$ streams, at different number $l = 1,...,s$ of streams that shift off the target and mean shift size $\Delta p = -2.0\sigma, -1.0\sigma, -0.5\sigma, 0.5\sigma, 1.0\sigma, 2.0\sigma$.

	NN based control charting procedure Δp						Group p control chart Δp					
	-2.0σ	-1.0σ	-0.5σ	0.5σ	1.0σ	2.0σ	-2.0σ	-1.0σ	-0.5σ	0.5σ	1.0σ	2.0σ
l												
1	36.21	143.88	276.24	249.38	123.46	26.08	12.63	107.76	263.85	374.53	230.95	29.94
2	6.2	52.97	186.57	164.20	43.16	4.8	6.68	59.38	200	389.11	167.5	15.55
3	2.52	23.66	113.38	99.60	21.43	2.29	4.64	41.95	152.44	381.68	131.75	10.4
4	1.47	11.72	73.86	66.36	10.82	1.4	3.63	31.21	124.38	343.64	104.93	7.95
5	1.15	6.44	45.56	45.13	6.2	1.23	3.03	25.63	104.17	343.64	83.96	6.59
6	1.05	4.14	33.02	32.77	4.18	1.04	2.64	22.68	93.02	377.36	77.7	5.6

3 Real-case study

In addition, data mentioned in Section 1, courtesy of Hitachi Rail STS, are employed to demonstrate the effectiveness of the proposed method. Operational data from modern railway HVAC systems installed on-board a passenger 6-coach rail vehicle are collected every two minutes in the summer season and the available variables for each coach are summarised in Table 3. A central unit monitors the HVAC system performance through temperature sensors that activate heating or cooling mode based on changes in the readings of the outside (T_{out}) and interior (T_{in}) temperatures to ensure the passenger thermal comfort required by the European regulations [1]. The HVAC system is designed to maintain T_{in} as close as possible to the target (T_{set}) temperature, that is the desired temperature value at which the HVAC sys-

Table 3 Operational variables available for each of the $s = 6$ train coaches

Variable	Description
T_{in}	Interior temperature
T_{out}	Outdoor Temperature
T_{set}	Target temperature
T_{supply}	Air temperature provided by the HVAC
C_{m}	Cooling mode status

tems attempt to maintain the T_{in} value at each time instant. [1] allows the difference $\Delta T = T_{\text{in}} - T_{\text{set}}$ to be no larger than 2°C, otherwise the train cannot operate. For each of the $s = 6$ coach HVAC streams, the quality variable of interest Y_{tj} is the number of times the cooling mode status (C_{m}) is 1 over $N = 30$ samples (i.e., 1 hour). Specifically, C_{m} is coded by 1 if the HVAC system is cooling at a given subgroup point, 0 otherwise. The parameter $p = 0.52$ is estimated from an IC sample. The idea is to monitor whether or not an HVAC system is cooling properly to ensure the passenger thermal comfort. When a stream j returns a p_j estimate that is significantly different from that of the other streams, then an alarm should be provided. Note that the data used in this short paper refer to the same HVAC systems and time window of [8]. However, in [8], the operational variable monitored was instead the difference $T_{\text{set}} - T_{\text{in}}$, with a sampling period of two minutes. Differently from the number of times ($C_{\text{m}} = 1$) over $N = 30$ samples, $T_{\text{set}} - T_{\text{in}}$ is a continuous random variate and there is no known engineering/physical relationship between them.

4 Conclusions

For comparison purposes, Tables 1 and 2 show the ARL_1 performance at fixed $ARL_0 = 358$ of the Wludyka and Jacobs' Chi-squared and group p MSBP control charts, as well as that of the proposed one at different number $l = 1, ..., s$ of streams that shift off-target by as much as $\Delta p = -2.0\sigma, -1.0\sigma, -0.5\sigma, 0.5\sigma, 1.0\sigma, 2.0\sigma$. Numerical investigations at different $ARL_0 = 20, 100$ give similar results and thus, are not reported here for conciseness. In particular, the proposed monitoring strategy outperforms the competitor control charts when at least one stream shifts off-target for all severity levels of mean shift. By means of a real-case study in the transportation sector, the proposed method was also demonstrated to be capable of signalling OC conditions in real time, even when more than one streams (from different coaches) are OC.

The computations are performed by using a single core of an Intel Xeon Platinum 8160 node of the ENEA CRESCO6 system (2.10GHz, 192 GB RAM, no GPU) [5] and are numerically implemented by means of the open source software environment Python [9]. The training time of the NN is about 16 seconds and the computational time required to process a new sample is about 1 millisecond. This facilitates

the repeatability and reproducibility of this research and the attained results as well as the practical application of the proposed approach in similar industrial scenarios.

Acknowledgements This work has been done in the framework of the R&D project of the multi-regional investment programme "REINForce: REsearch to INspire the Future" (CDS000609) with Hitachi Rail Italy, supported by the Italian Ministry for Economic Development (MISE) through the Invitalia agency. The computing resources and the related technical support used for this work have been provided by CRESCO/ENEAGRID High Performance Computing infrastructure and its staff [5]. CRESCO/ENEAGRID High Performance Computing infrastructure is funded by ENEA, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development and by Italian and European research programmes, see http://www.cresco.enea.it/english for information.

References

1. EN 14750, I.: Railway applications. Air conditioning for urban and suburban rolling stock. Comfort parameters (2006)
2. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters **27**, 861–874 (2006)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
4. Hwarng, H.B.: Neural networks in statistical process control. Wiley StatsRef: Statistics Reference Online (2014)
5. Iannone, F., Ambrosino, F., Bracco, G., De Rosa, M., Funel, A., Guarnieri, G., Migliori, S., Palombi, F., Ponti, G., Santomauro, G., Procacci, P.: Cresco enea hpc clusters: a working example of a multifabric gpfs spectrum scale layout. In: 2019 International Conference on High Performance Computing Simulation (HPCS), pp. 1051–1052 (2019)
6. Montgomery, D.C.: Introduction to statistical quality control. John Wiley & Sons (2007)
7. Psarakis, S.: The use of neural networks in statistical process control charts. Quality and Reliability Engineering International **27**, 641–650 (2011)
8. Sposito, G., Lepore, A., Palumbo, B., Giannini, G.: Neural network for statistical process control of a multiple stream process with an application to HVAC systems in passenger rail vehicles. Preface XIX 1 Plenary Sessions p. 602 (2021)
9. Van Rossum, G., Drake Jr, F.L.: Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam (1995)
10. Wludyka, P.S., Jacobs, S.L.: Controlling homogeneous multistream binomial processes with a chi-squared control chart. Proceedings of the 33rd Annual Meeting of the Decision Sciences Institute pp. 2254–2263 (2002)
11. Wludyka, P.S., Jacobs, S.L.: Runs rules and p-charts for multistream binomial processes. Communications in Statistics-Simulation and Computation **31**, 97–142 (2002)
12. Zorriassatine, F., Tannock, J.: A review of neural networks for statistical process control. Journal of Intelligent Manufacturing **9**, 209–224 (1998)

Sparse Signal Extraction via Variational SVM

Estrazione di Segnali Sparsi via SVM variationali

Cristian Castiglione and Mauro Bernardi

Abstract This paper presents a new semiparametric variational Bayes algorithm for estimating a wide class of regularized support vector machine models. We focus in particular on robust sparse signal extraction via Bayesian support vector regression, taking advantages from a Bayesian formulation for handling both the inclusion of inducing shrinkage priors and automatic learning of the associated penalization parameters. A simulation study is then presented to assess the benefits of the proposed method.

Abstract *In questo articolo viene presentato un nuovo algoritmo variazionale Bayesiano per stimare un gran numero di modelli appartenenti alla classe support vector machine. Un'attenzione particolare viene posta nell'estrazione di segnali sparsi attraverso modelli di regressione robusta, sviluppati all'interno di una cornice inferenziale Bayesiana. Quest'ultima permette di gestire in forma unificata sia l'utilizzo di specifiche informazioni a priori, che una selezione automatica dei parametri di penalizzazione associati alla priori. Infine, viene proposto uno studio di simulazione comparativo per verificare le proprietà del metodo.*

Key words: Support vector regression; Variational Bayes; Sparse signal extraction

1 Introduction

Support vector machines (SVM) [11] are a family of robust prediction models often used in machine learning applications with high dimensional covariates. They are

Cristian Castiglione
Department of Statistical Sciences, University of Padova, via Cesare Battisti 241, Padova, Italy
e-mail: cristian.castiglione@phd.unipd.it

Mauro Bernardi
Department of Statistical Sciences, University of Padova, via Cesare Battisti 241, Padova, Italy
e-mail: mauro.bernardi@unipd.it

typically regularized with an inducing sparsity penalty to manage the bias-variance trade-off and to perform variable selection.

The historical estimation method for SVM relies on dual constrained optimization, which is extremely efficient, but does not permits to quantify the variability of the point estimates and, more in general, to do inference on the model parameters. Here, instead, we take a variational Bayes perspective [5] in order to perform approximate posterior inference and uncertainty quantification. Variational inference does not require computationally expensive posterior simulation and can be performed via efficient optimization methods, allowing for a fast processing of high dimensional data.

Our aim is then to present a semiparametric variational Bayes (SVB) scheme to perform sparse SVM regression. Differently from [6] and [4], our approach does not need to exploit complicate data augmentation strategies and, instead, approximates directly the original posterior density employing the so-called Knowles-Minka-Wand formula [10, 9] for variational Gaussian approximations.

2 Bayesian model

Regularized SVM regression predict the response variable $y_i \in \mathbb{R}$, $i = 1,\dots,n$, as the linear combination $x_i^\top \beta$, where $x_i \in \mathbb{R}^p$ is a vector of covariates and $\beta \in \mathbb{R}^p$ is a parameter vector. The unknown coefficients are then estimated by minimizing a penalized objective function $L_p(\beta;y) = L(\beta;y) + P_\lambda(\beta)$, balancing a goodness of fit criteria, i.e. the loss function $L(\beta;y)$, with a model complexity criteria, i.e. the penalization function $P_\lambda(\beta)$, which depends on the regularization parameter λ. The loss function characterizing the SVR optimization problem is the so-called ε-insensitive loss [11], being defined as

$$L(\beta;y) = \sum_{i=1}^{n} \max(0, |y_i - x_i^\top \beta| - \varepsilon).$$

This makes the SVR model robust to outlier contamination and insensitive to small prediction errors, since $L(\beta;y)$ assigns no loss to residual values $y_i - x_i^\top \beta$ within the insensitivity band $[-\varepsilon, +\varepsilon]$, for any $\varepsilon > 0$.

Here, we assume for the model a Bayesian formulation where

$$\pi(y|\beta) = \exp\{-L(\beta;y)\}, \quad \pi_\lambda(\beta) = \exp\{-P_\lambda(\beta)\},$$

are, respectively, the pseudo-likelihood function and the prior distribution on β. Thus, our subjective beliefs about β can be updated to the general posterior [2] through the Bayes formula: $\pi_\lambda(\beta|y) \propto \pi(y|\beta)\pi_\lambda(\beta)$.

As pointed out by [6, 4], a Bayesian formulation provides a unified approach to overcome many shortcomings of classical SVR, allowing for the inclusion of random effects, structured prior distributions, spatial and temporal processes and automatic tuning of the penalty parameter λ. Here, we focus on the estimation of

sparse signals assuming for β an inducing shrinkage prior belonging to the class of global-local distributions [1]. In particular, we consider an adaptive Bayesian Lasso penalty [3], that can be expressed via the hierarchical representation

$$\beta_j|\tau^2,\psi_j^2 \sim DE(0,\tau^2\psi_j^2), \quad \tau^2 \sim IG(A_\tau,B_\tau), \quad \psi_j^2 \sim IG(A_\psi,B_\psi).$$

where $\tau^2 > 0$ is the so-called global parameter and $\psi_j^2 > 0$ is the element-specific local parameter; $A_\tau, B_\tau, A_\psi, B_\psi > 0$ are fixed user-specified hyperparameters. We denote with $DE(\mu,\sigma)$ the Double-Exponential, or Laplace, distribution with location $\mu \in \mathbb{R}$ and scale $\sigma > 0$, while $IG(A,B)$ is the Inverse-Gamma distribution with shape $A > 0$ and rate $B > 0$. Within this model specification the regularization term λ corresponds to the vector $(\tau^2,\psi_1^2,\ldots,\psi_p^2)$.

A common feature of all global-local priors is to operate an aggressive shrinkage on the parameters corresponding to non-significant variables, while introducing a negligible amount of bias on the signal components. To this end, more involved prior models can be considered [1], such as the Horseshoe, or the Dirichlet-Laplace, however no relevant differences arise in the derivation of our method, that is agnostic to any prior choice.

3 Variational inference

Let us denote with $\theta = (\beta,\psi^2,\tau^2)$ and $\Theta = \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+$ the set of all parameters in the model and the corresponding parameter space, respectively.

Bayesian inference on the SVR model is then carried out via variational approximation [5]. This proceeds by replacing the true posterior $\pi(\theta|y)$ with a simpler density function $q(\theta)$ belonging to a specific functional space. The optimal approximating density for θ is then estimated maximizing the lower bound on the marginal log-likelihood, also called evidence lower bound: $\log \underline{\pi}(y;q) = \mathbb{E}_q\big[\log\{\pi(y,\theta)/q(\theta)\}\big]$, where $\mathbb{E}_q(\cdot)$, or $\mu_{q(\cdot)}$, is the expectation calculated with respect to $q(\theta)$. Assuming the mean field restriction $q(\theta) = q(\theta_1)\ldots q(\theta_K)$, for any partition of θ, there exists an exact coordinatewise solution for $q(\theta_k)$, being

$$q^*(\theta_k) \propto \exp\{\mathbb{E}_{-k}(\log\pi(\theta_k|\text{rest}))\}, \quad k = 1,\ldots,K, \tag{1}$$

where $\mathbb{E}_{-k}(\cdot)$ is the expectation taken with respect to $\prod_{j\neq k} q(\theta_j)$ and $\pi(\theta_k|\text{rest})$ is the full-conditional distribution of θ_k. Whenever θ_k has a conjugate prior distribution, formula (1) can be derived in closed form, otherwise a parametric restriction can be imposed, in order to project the exact solution (1) into the appropriate functional space.

Here, we both assume the factorization $q(\theta) = q(\beta)q(\psi^2)q(\tau^2)$ and a parametric restriction for $q(\beta)$. In particular, we impose $q(\beta) = q(\beta;\mu,\Sigma)$ to be a multivariate Gaussian density with mean vector $\mu \in \mathbb{R}^p$ and variance-covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. Defining $f(\mu,\Sigma) = \mathbb{E}_q[\log\pi(\beta|\text{rest})]$, $g(\mu,\Sigma) = \nabla_\mu f(\mu,\Sigma)$ and $H(\mu,\Sigma) =$

$\nabla^2_\mu f(\mu,\Sigma)$ and employing the so-called Knowles-Minka-Wand method [10, 9] for updating $q(\beta)$ and formula (1) for updating $q(\psi^2)$ and $q(\tau^2)$, we obtain an iterative optimization algorithm that cycles over the following steps:

- $q^*(\beta) \sim N_p(\hat{\mu},\hat{\Sigma})$, where $\hat{\mu} \leftarrow \hat{\mu} - H(\hat{\mu},\hat{\Sigma})^{-1} g(\hat{\mu},\hat{\Sigma})$ and $\hat{\Sigma} \leftarrow -H(\hat{\mu},\hat{\Sigma})^{-1}$;
- $q^*(\psi^2) \sim IG(\hat{A}_\psi,\hat{B}_\psi)$, where $\hat{A}_\psi \leftarrow A_\psi + 1$ and $\hat{B}_\psi \leftarrow B_\psi + \mu_{q(1/\tau^2)}\mu_{q(|\beta|)}$;
- $q^*(\tau^2) \sim IG(\hat{A}_\tau,\hat{B}_\tau)$, where $\hat{A}_\tau \leftarrow A_\tau + p$ and $\hat{B}_\tau \leftarrow B_\tau + \mu^\top_{q(1/\psi^2_j)}\mu_{q(|\beta|)}$.

Here all the involved expectations can be calculated analytically. In particular, the gradient vector and Hessian matrix needed for the Newton update of $\hat{\mu}$ and $\hat{\Sigma}$ are

$$
\begin{aligned}
g(\mu,\Sigma) &= -Dv^{(1)} - X^\top w^{(1)} \\
H(\mu,\Sigma) &= -DV^{(2)} - X^\top W^{(2)} X
\end{aligned}
$$

where $D = \mathrm{diag}[\mu_{q(1/\tau^2)}\mu_{q(1/\psi^2)}]$, $V^{(2)} = \mathrm{diag}[v^{(2)}]$, $W^{(2)} = \mathrm{diag}[w^{(2)}]$, and

$$
\begin{aligned}
v^{(1)}_j &= 2\Phi(0;-\mu_j,\Sigma_{jj}) - 1, \\
v^{(2)}_j &= 2\phi(0;-\mu_j,\Sigma_{jj}), \\
w^{(1)}_i &= 2\big[\phi(+\varepsilon;y_i - x_i^\top\mu, x_i^\top\Sigma x_i) - \phi(-\varepsilon;y_i - x_i^\top\mu, x_i^\top\Sigma x_i) - 1\big], \\
w^{(2)}_i &= 2\big[\phi(+\varepsilon;y_i - x_i^\top\mu, x_i^\top\Sigma x_i) + \phi(-\varepsilon;y_i - x_i^\top\mu, x_i^\top\Sigma x_i)\big],
\end{aligned}
$$

for $j = 1,\ldots,p$ and $i = 1,\ldots,n$. The symbols $\phi(\cdot;\mu,\sigma^2)$ and $\Phi(\cdot;\mu,\sigma^2)$ denote the probability density function and the cumulative density function of a $N(\mu,\sigma^2)$ distribution.

4 Simulation study

The performances of the proposed approximation are then assessed by means of a simulation study, where the responses $y_1,\ldots,y_n$ are independently generated according with the location-scale model

$$
y_i \sim t(f(z_i),\sigma,\nu), \quad z_i \sim U(0,1), \quad i = 1,\ldots,n, \quad n = 300,
$$

with $t(\mu,\sigma,\nu)$ denoting a t-distribution with location $\mu \in \mathbb{R}$, scale $\sigma > 0$ and degrees of freedom $\nu > 0$ and $U(0,1)$ being the uniform distribution over the interval $[0,1]$. Further, $f(z) = -1.02z + 0.018z^2 + 0.4\phi(z;0.38,0.08) + 0.08\phi(z;0.75,0.03)$ is a non-linear function defined over $z \in [0,1]$, $\sigma = 0.31$ and $\nu = 3$.

In this context, we try to reconstruct the true signal, i.e. the function $f(x)$, via semiparametric regression. A set of 40 orthogonal O'Sullivan spline basis [8] $B_1(\cdot),\ldots,B_{40}(\cdot)$ are displaced over the interval $[0,1]$; the i-th row of our design matrix is then equal to $x_i^\top = [1, z_i, B_1(z_1),\ldots,B_{40}(z_i)]$. The rest of the model is specified as discussed is section 2. The parameters are hence estimated using the Gibbs

sampler (MCMC) of [6], the mean field variational Bayes (MFVB) of [4] and the proposed semiparametric variational Bayes (SVB) algorithm.

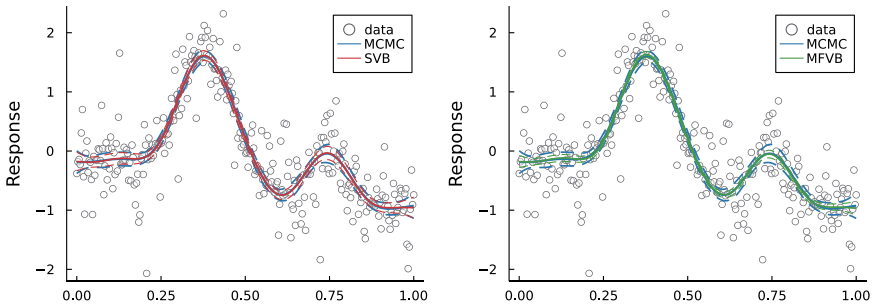


Fig. 1 Fitted values for the MCMC, MFVB and SVB methods. The solid lines are the pointwise predictions, while the dashed lines correspond to the 95% HPD posterior credibility bands.

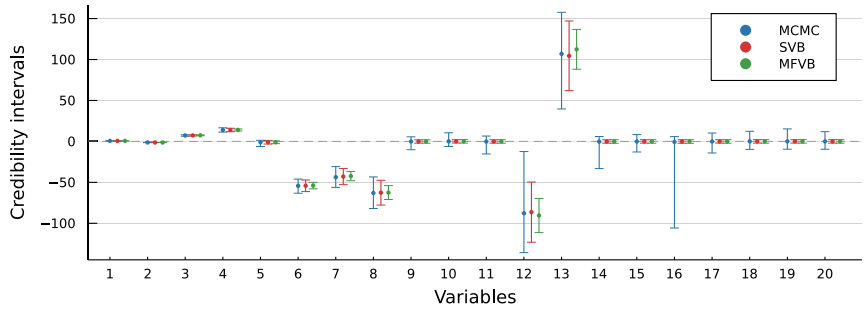


Fig. 2 Point estimates and 95% HPD posterior credibility intervals for the first 20 parameters estimated using the MCMC, MFVB and SVB methods.

Figure 1 shows the predicted values obtained with the methods mentioned so far. MFVB and SVB are comparable in terms of point prediction, but not in terms of variability quantification. Indeed, the prediction bands corresponding to SVB are far closer to MCMC than MFVB, which instead tends to produce narrower bands. The same phenomenon is highlighted by figure 2, which shows the posterior point estimates (posterior mean) and 95% HPD credibility intervals for the first 20 regression parameters.

On the other side, all the three approaches encourage the non-relevant parameters to be overshrunk toward 0, but the deterministic approximation methods, i.e. MFVB and SVB, collapse in a far more aggressive way the irrelevant signals, shrinking not only the posterior mean, but also the posterior variability. Here we use the signal adaptive variable selection (SAVS) method proposed by [7] to distinguish between signal and noise. With all the three approaches we found 32 non-relevant variables and 10 signals, which approximately correspond to the parameters with a posterior HPD interval not including 0.

	RMSE	Accuracy				
		10%	25%	50%	75%	90%
MCMC	0.0897					
MFVB	0.0875	0.6621	0.7205	0.7523	0.7683	0.7761
SVB	0.0848	0.7291	0.7430	0.7592	0.7734	0.9511

Table 1 Posterior goodness of fit measures for MCMC, MFVB and SVB. RMSE denotes the root mean squared error calculated between the estimated curve and the true function $f(\cdot)$. Accuracy refers to the score defined as $\text{Acc}_j(q)$, for $j = 1,\ldots,42$. The percentages correspond to 5 quantiles of the marginal empirical distribution of the accuracy scores, calculated for the 42 parameters in the model.

The ability of a deterministic variational density to approximate the true posterior, i.e. the MCMC one, can be estimated by calculating the marginal accuracy index $\text{Acc}_j(q) = 1 - 0.5 \int_{\mathbb{R}} |q(\beta_j) - \pi(\beta_j|y)|\, d\beta_j$, $j = 1,\ldots,p$, that is a real quantity defined over $[0,1]$. An optimal approximation is reached for values close to 1, whereas values close to 0 suggest a poor approximation. The accuracy scores are then calculated for all the parameters estimated with MFVB and SVB and the results are shown in table 1. For all the 5 considered quantiles, SVB presents a higher level of precision in estimating the true posterior than MFVB, confirming the evidences obtained looking at figure 1 and 2.

References

1. Bhadra, A., Datta, J., Polson, N.G., Willard, B.: Lasso meets horseshoe: a survey. Statist. Sci. **34**, 405–427 (2019)
2. Bissiri, P.G., Holmes, C.C., Walker, S.G.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B. Stat. Methodol. **78**, 1103–1130 (2016)
3. Leng, C., Tran, M.N., Nott, D.: Bayesian adaptive Lasso. Ann. Inst. Statist. Math. **66**, 221–244 (2014)
4. Luts, J., Ormerod, J.T.: Mean field variational Bayesian inference for support vector machine classification. Comput. Statist. Data Anal. **73**, 163–176 (2014)
5. Ormerod, J. T., Wand, M. P.: Explaining variational approximations. Am. Stat. **64**, 140–153 (2010)
6. Polson, N.G., Scott, Steven L.: Data augmentation for support vector machines. Bayesian Anal. **6**, 1–23 (2011)
7. Ray, P., Bhattacharya, A.: Signal adaptive variable selector for the horseshoe prior. arXiv preprint arXiv:1810.09004 (2018)
8. Wand, M.P., Ormerod, J.T.: On semiparametric regression with O'Sullivan penalized splines. Aust. N. Z. J. Stat. **50**, 179–198 (2008)
9. Wand M. P.: Fully simplified multivariate normal updates in non-conjugate variational message passing. J. Mach. Learn. Res. **15**, 1351–1369 (2014)
10. Knowles, D., Minka, T.: Non-conjugate variational message passing for multinomial and binary regression. Adv. Neural Inf. Process. Syst. **24**, 1701–1709 (2011)
11. Vapnik, V.N.: Statistical learning theory. John Wiley & Sons, Inc., New York (1998)

Bayesian modelling and inference 1

Bayesian Inference for the Multinomial Probit Model under Gaussian Prior Distribution

Inferenza Bayesiana per il Modello Probit Multinomiale con Distribuzione a Priori Normale

Augusto Fasano, Giovanni Rebaudo and Niccoló Anceschi

Abstract Multinomial probit (MNP) models are fundamental and widely-applied regression models for categorical data. [1] proved that the class of unified skew-normal distributions is conjugate to several MNP sampling models. This allows to develop Monte Carlo samplers and accurate variational methods to perform Bayesian inference. In this paper, we adapt the above-mentioned results for a popular special case: the discrete-choice MNP model under zero-mean and independent Gaussian priors. This allows to obtain simplified expressions for the parameters of the posterior distribution and an alternative derivation for the variational algorithm that gives a novel understanding of the fundamental results in [1] as well as computational advantages in our special settings.

Abstract *I modelli probit multinomiali (*MNP*) sono una classe popolare di modelli di regressione per dati categoriali. [1] hanno dimostrato che la classe delle distribuzioni normali asimmetriche è coniugata rispetto a modelli* MNP*. Questo ha permesso di sviluppare schemi di campionamento ed accurati metodi variazionali per l'inferenza Bayesiana. Nel presente articolo, adattiamo tali risultati ad un caso particolarmente rilevante: il* MNP *a scelta discreta con distribuzioni a priori normali a media zero. Questo permette di otterenere espressioni semplificate per i parametri della distribuzione a posteriori ed una derivazione alternativa dell'approccio variazionale che fornisce una nuova compresione del risultato fondamentale in [1] oltre a vantaggi computazionali nel nostro caso particolare.*

Key words: Multinomial Probit Model, Variational Inference, Unified Skew-Normal Distribution, Bayesian inference, Categorical Data, Classification

Augusto Fasano
European Commission, Joint Research Centre (JRC), Ispra, Italy and Collegio Carlo Alberto.
e-mail: augusto.fasano@ec.europa.eu

Giovanni Rebaudo
Department of Statistics and Data Sciences, the University of Texas at Austin.
e-mail: giovanni.rebaudo@austin.utexas.edu

Niccoló Anceschi
Department of Decision Sciences, Bocconi University
e-mail: niccolo.anceschi@unibocconi.it

1 Introduction

Multinomial probit (MNP) models constitute a fundamental tool for categorical data regression, thanks to their interpretability and flexibility [2]. Originally introduced by [3] to avoid the restrictive assumption of the *independence of irrelevant alternatives* typical of multinomial logit models, such models have faced the growth of many different specifications. Among them, we consider the Bayesian formulation of the discrete choice MNP model with class-specific effects [4], under a zero-mean and independent Gaussian prior for the parameters, adapting the results of [1] obtained under more general prior specifications and for a wider range of models. In such a construction, originally developed in the econometrics literature, to each possible choice (or class) $\ell = 1, \dots, L$ that individual $i = 1, \dots, n$ faces, a corresponding random latent utility $z_{i\ell} = \mathbf{x}_i^\intercal \boldsymbol{\beta}_\ell + \varepsilon_{i\ell}$ is associated, where $\mathbf{x}_i \in \mathbb{R}^p$ is the covariate vector for observation i, $\boldsymbol{\beta}_\ell \in \mathbb{R}^p$ is the class-specific vector of the covariate effects and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iL})^\intercal \sim \mathrm{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$, independently across units $i = 1, \dots, n$. Note that the error terms for different alternatives can be correlated, since $\boldsymbol{\Sigma}$, which is assumed to be known, is not necessarily diagonal. Among all the possible choices $1, \dots, L$, individual i chooses the one giving her the maximal utility, meaning that $y_i = \ell$ if and only if $z_\ell = \max\{z_1, \dots, z_L\}$. Thus, for each $i = 1, \dots, n$, independently

$$\begin{aligned} \Pr(y_i = \ell \mid \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_L, \mathbf{x}_i) &= \Pr(z_{i\ell} > z_{ik}, \forall k \neq \ell) \\ &= \Pr(\mathbf{x}_i^\intercal \boldsymbol{\beta}_\ell + \varepsilon_{i\ell} > \mathbf{x}_i^\intercal \boldsymbol{\beta}_k + \varepsilon_{ik}, \forall k \neq \ell). \end{aligned} \tag{1}$$

Since in (1) only pairwise differences between the parameters matter, we set $\boldsymbol{\beta}_L = \mathbf{0}$ for identifiability purposes. In order to perform Bayesian inference, we complete the model by specifying a multivariate Gaussian prior distribution for the vector of parameters $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\intercal, \dots, \boldsymbol{\beta}_{L-1}^\intercal\right)^\intercal$ centered in zero, with independent and homoscedastic components. Thus,

$$\boldsymbol{\beta} \sim \mathrm{N}_{p(L-1)}\left(\mathbf{0}, \nu^2 \mathbf{I}_{p(L-1)}\right). \tag{2}$$

The more general results under a unified skew-normal (SUN) prior have been developed in [1] for a broader class of MNP models. In fact, it is shown that in that case the posterior belongs again to the class of SUN distributions, with updated dimensionalities and parameters. This allows to perform posterior inference via i.i.d. samples in small-to-moderate n settings, while a blocked variational procedure is developed to avoid the computational bottlenecks that one may encounter in large n scenarios. The prior specification (2), however, constitutes a popular choice in case there are no reasons to *a priori* assume any dependence between the parameters or asymmetry in their distribution, and it is worth a separate treatment as it allows the simplification of the expression of some important parameters in the posterior distribution and the derivation of an alternative proof for the variational algorithm that can have computational advantages. These two important aspects represent the main focus of the present article and will be the focus of Sections 2 and 3 below, while Section 4 is dedicated to the discussion of possible future research directions.

2 Posterior inference via Monte Carlo samples

Recently, [1] showed that for a broad class of MNP models, a SUN prior distribution leads to a SUN posterior distribution, extending the previous conjugacy results derived by [5] for the classical binary probit model. We specify here results in Section 2.2 in [1] under the particular Gaussian prior distribution (2), for the peculiar advantages explained in Section 1. Before doing that, we briefly recap the definition and the main properties of the SUN distribution. Further details can be found, for instance, in [6]. A random vector $\boldsymbol{\beta} \in \mathbb{R}^q$ has SUN distribution, $\boldsymbol{\beta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, if its density function $p(\boldsymbol{\beta})$ can be expressed as

$$p(\boldsymbol{\beta}) = \phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_h[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma}; \boldsymbol{\Gamma})},$$

where the covariance matrix $\boldsymbol{\Omega}$ of the Gaussian density $\phi_q(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega})$ can be decomposed as $\boldsymbol{\Omega} = \boldsymbol{\omega} \bar{\boldsymbol{\Omega}} \boldsymbol{\omega}$, i.e. by rescaling the correlation matrix $\bar{\boldsymbol{\Omega}}$ via the diagonal scale matrix $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$, with $\odot$ denoting the element-wise Hadamard product. Moreover, $\Phi_h(\mathbf{u}; \mathbf{W})$ denotes the cumulative distribution function of a $\text{N}_h(\mathbf{0}, \mathbf{W})$ evaluated at $\mathbf{u}$. The following additive characterization constitutes a fundamental property to further understand the role of the parameters and to develop an i.i.d. sampler. If $\boldsymbol{\beta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, then $\boldsymbol{\beta} \overset{\text{d}}{=} \boldsymbol{\xi} + \boldsymbol{\omega}(\mathbf{V}_0 + \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\mathbf{V}_1)$, with

$$\mathbf{V}_0 \sim \text{N}_q(\mathbf{0}, \bar{\boldsymbol{\Omega}} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^\top), \ \mathbf{V}_1 \sim \text{TN}_h(\mathbf{0}, \boldsymbol{\Gamma}; A_{-\boldsymbol{\gamma}}),$$

where $A_{-\boldsymbol{\gamma}} = \left\{\mathbf{a} \in \mathbb{R}^h \colon a_i \geq -\gamma_i \ \forall i\right\}$ and $\text{TN}_h(\mathbf{m}, \mathbf{W}; A)$ denotes the h-variate normal distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{W}$, truncated in the region A.

In order to derive the SUN posterior distribution of $\boldsymbol{\beta}$ for model (1)-(2), we explicitly write the likelihood expression for the observed responses $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Proposition 1 (Proposition 2 in [1]). *For each $\ell = 1, \ldots, L$, denote with $\mathbf{v}_\ell$ the $L \times 1$ vector with value 1 in position l and 0 elsewhere, and with $\mathbf{V}_{[-\ell]}$ the $(L-1) \times L$ matrix whose rows are obtained by stacking vectors $(\mathbf{v}_k - \mathbf{v}_\ell)^\top$, for $k \neq \ell$. Finally, define $\mathbf{X}_{i[-\ell]} = -\bar{\mathbf{V}}_{[-\ell]} \otimes \mathbf{x}_i^\top$, where $\bar{\mathbf{V}}_{[-\ell]}$ is the $(L-1) \times (L-1)$ matrix obtained by removing the L-th column from $\mathbf{V}_{[-\ell]}$ and $\otimes$ denotes the Kronecker product.*

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} \Phi_{L-1}(\mathbf{X}_{i[-y_i]}\boldsymbol{\beta}; \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^\top) = \Phi_{n(L-1)}(\bar{\mathbf{X}}\boldsymbol{\beta}; \boldsymbol{\Lambda}),$$

where $\bar{\mathbf{X}}$ is an $n(L-1) \times p(L-1)$ block matrix with $(L-1) \times p(L-1)$ row blocks $\bar{\mathbf{X}}_{[i]} = \mathbf{X}_{i[-y_i]}$, whereas $\boldsymbol{\Lambda}$ denotes an $n(L-1) \times n(L-1)$ block-diagonal covariance matrix with $(L-1) \times (L-1)$ diagonal blocks $\boldsymbol{\Lambda}_{[ii]} = \mathbf{V}_{[-y_i]}\boldsymbol{\Sigma}\mathbf{V}_{[-y_i]}^\top$, for $i = 1, \ldots, n$.

By combining the prior specification (2) with the likelihood (1), the following closed-form expression for the posterior distribution of $\boldsymbol{\beta}$ is obtained as a direct consequence of Theorem 1 in [1] adapted to the special case described in Section 1.

Algorithm 1: Strategy to sample from the SUN posterior in Theorem 1

for *i=1,...,N* **do**

- **[1]** Sample $\mathbf{V}_0^{(i)} \sim \mathrm{N}_{p(L-1)}(\mathbf{0}, \mathbf{I}_{p(L-1)} - \boldsymbol{\Delta}_{\text{pst}}\boldsymbol{\Gamma}_{\text{pst}}^{-1}\boldsymbol{\Delta}_{\text{pst}}^{\intercal})$ [in R use the function `rmvnorm`]
- **[2]** Sample $\mathbf{V}_1^{(i)} \sim \mathrm{TN}_{n(L-1)}(\mathbf{0}, \boldsymbol{\Gamma}_{\text{pst}}; [0,\infty)^{n(L-1)})$ [in R use the function `rtmvnorm`]
- **[3]** Compute $\boldsymbol{\beta}^{(i)} = \nu(\mathbf{V}_0^{(i)} + \boldsymbol{\Delta}_{\text{pst}}\boldsymbol{\Gamma}_{\text{pst}}^{-1}\mathbf{V}_1^{(i)})$

Output: i.i.d. samples $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}$ from the posterior distribution in Theorem 1.

Theorem 1 (from Theorem 1 in [1]). *Under model* (1)-(2)*, the posterior density is*

$$(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \sim \mathrm{SUN}_{p(L-1),n(L-1)}(\mathbf{0}, \boldsymbol{\Omega}_{\text{pst}}, \boldsymbol{\Delta}_{\text{pst}}, \mathbf{0}, \boldsymbol{\Gamma}_{\text{pst}}) \tag{3}$$

with $\boldsymbol{\Omega}_{\text{pst}} = \nu^2\mathbf{I}_{p(L-1)}$*,* $\boldsymbol{\Delta}_{\text{pst}} = \nu\bar{\mathbf{X}}^{\intercal}\mathbf{s}^{-1}$ *and* $\boldsymbol{\Gamma}_{\text{pst}}$ *is an* $n(L-1) \times n(L-1)$ *correlation matrix* $\boldsymbol{\Gamma}_{\text{pst}} = \mathbf{s}^{-1}(\nu^2\bar{\mathbf{X}}\bar{\mathbf{X}}^{\intercal} + \boldsymbol{\Lambda})\mathbf{s}^{-1}$*, where* $\mathbf{s} = [(\nu^2\bar{\mathbf{X}}\bar{\mathbf{X}}^{\intercal} + \boldsymbol{\Lambda}) \odot \mathbf{I}_{n(L-1)}]^{1/2}$.

Theorem 1 and the additive representation of the SUN (2) allow to develop an i.i.d. sampler of the posterior of $\boldsymbol{\beta}$ as described in Algorithm 1.

3 Partially-factorized blocked mean field approximation (PFM-B)

Basic manipulations of the posterior (3) show that $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}) \cdot \Pr[\bar{\mathbf{z}} > \mathbf{0} \mid \boldsymbol{\beta}, \bar{\mathbf{X}}]$, where $\bar{\mathbf{z}} = (\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_n)^{\intercal} \in \mathbb{R}^{n(L-1)}$ and $\bar{\mathbf{z}} \mid \boldsymbol{\beta}, \bar{\mathbf{X}} \sim \mathrm{N}_{n(L-1)}\left(\bar{\mathbf{X}}\boldsymbol{\beta}, \boldsymbol{\Lambda}\right)$. Thus, $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ can be seen as the marginal posterior distribution of the dual model

$$\boldsymbol{\beta} \sim \mathrm{N}_{p(L-1)}\left(\mathbf{0}, \nu^2\mathbf{I}_{p(L-1)}\right) \tag{4a}$$

$$\bar{\mathbf{z}}_i \mid \boldsymbol{\beta}, \bar{\mathbf{X}} \overset{ind}{\sim} \mathrm{N}_{L-1}\left(\bar{\mathbf{X}}_{[i]}\boldsymbol{\beta}, \boldsymbol{\Lambda}_{[ii]}\right), \quad i = 1, \ldots, n, \tag{4b}$$

$$\bar{\mathbf{y}}_i = \mathbb{1}\left[\bar{\mathbf{z}}_i > \mathbf{0}\right], \qquad i = 1, \ldots, n, \tag{4c}$$

in which one observes $\bar{\mathbf{y}}_i = (1, \ldots, 1)^{\intercal} \in \mathbb{R}^{L-1}$ for $i = 1, \ldots, n$ and $\mathbb{1}[\cdot]$ in (4c) is intended component-wise. Thus, it holds $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \int_{\mathbb{R}^{n(L-1)}} p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}}) d\bar{\mathbf{z}}$, with $\bar{\mathbf{y}} = (\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_n)^{\intercal}$ an $n(L-1)$ vector of ones. Since direct sampling from $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$ can run into computational issues when $n(L-1)$ is large due to step 2 in Algorithm 1, one can resort to variational methods, see, e.g., [7], to compute the 'best' possible approximating joint density $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$ in a given class of tractable density functions $\mathscr{Q}$. This optimal solution is the minimizer of the Kullback-Leibler divergence [8] $\mathrm{KL}\left[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) \,||\, p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}})\right] = \mathbb{E}_{q(\boldsymbol{\beta}, \bar{\mathbf{z}})}\left[\log\left(q(\boldsymbol{\beta}, \bar{\mathbf{z}}) / p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}})\right)\right]$.

In order to get a tractable approximation, but retain as much structure of the posterior distribution as possible, following [1], we take $\mathscr{Q} = \mathscr{Q}_{\text{PFM-B}} = \{q(\boldsymbol{\beta}, \bar{\mathbf{z}}) : q(\boldsymbol{\beta}, \bar{\mathbf{z}}) = q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{i=1}^{N} q(\bar{\mathbf{z}}_i)\}$. This class of densities leverages on the partially-factorized mean field approximation developed for the classical probit model in [9], but allows to maintain the intra-correlations between the components of the $\bar{\mathbf{z}}_i$'s, $i = 1, \ldots, n$, while enforcing the inter-correlations between them to be zero. A first result about the optimal approximating density is obtained by the chain rule for the KL divergence: $\mathrm{KL}[q(\boldsymbol{\beta}, \bar{\mathbf{z}}) || p(\boldsymbol{\beta}, \bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}})] = \mathrm{KL}[q(\bar{\mathbf{z}}) || p(\bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}})] + \mathbb{E}_{q(\bar{\mathbf{z}})}\{\mathrm{KL}[q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) || p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \bar{\mathbf{y}}, \bar{\mathbf{X}})]\}$. Thus, whatever is the value for $q(\bar{\mathbf{z}})$, the second summand is

Algorithm 2: CAVI algorithm for $q^*(\boldsymbol{\beta},\bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{i=1}^n q^*(\bar{\mathbf{z}}_i)$

[1] For each $i = 1,\dots,n$, set $\boldsymbol{\Sigma}_i^* = \left(\boldsymbol{\Lambda}_{[i,i]}^{-1} - \mathbf{H}_{[i,i]}\right)^{-1}$ and initialize $\mathbb{E}_{q^0(\bar{\mathbf{z}}_i)}[\bar{\mathbf{z}}_i] \in \mathbb{R}^{(L-1)}$

[2] **for** *t from* 1 *until convergence* **do**

  for *i from* 1 *to n* **do**

    [2.1] Set $\boldsymbol{\mu}_i^{(t)} = \boldsymbol{\Sigma}_i^* \mathbf{H}_{[i,-i]} (\mathbb{E}_{q^{(t)}}[\bar{\mathbf{z}}_1]^\intercal, \dots, \mathbb{E}_{q^{(t)}}[\bar{\mathbf{z}}_{i-1}]^\intercal, \mathbb{E}_{q^{(t-1)}}[\bar{\mathbf{z}}_{i+1}]^\intercal, \dots, \mathbb{E}_{q^{(t-1)}}[\bar{\mathbf{z}}_n]^\intercal)^\intercal$

    [2.2] Compute $\mathbb{E}_{q^{(t)}}[\bar{\mathbf{z}}_i]$ with $\bar{\mathbf{z}}_i \sim \text{TN}_{L-1}\left(\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^*; [0,\infty)^{L-1}\right)$ [in R use `MomTrunc()`]

[3] Set $q^*(\bar{\mathbf{z}}_i) = q^{(t)}(\bar{\mathbf{z}}_i)$ for $i = 1,\dots,n$

[4] Set $q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = \phi_{p(L-1)}(\boldsymbol{\beta} - \mathbf{V}\bar{\mathbf{X}}^\intercal \boldsymbol{\Lambda}^{-1}\bar{\mathbf{z}}; \mathbf{V})$

Output: $q^*(\boldsymbol{\beta},\bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{i=1}^N q^*(\bar{\mathbf{z}}_i)$

zero, and hence minimal, if and only if, calling $\mathbf{V} = (\nu^{-2}\mathbf{I}_{p(L-1)} + \bar{\mathbf{X}}^\intercal \boldsymbol{\Lambda}^{-1}\bar{\mathbf{X}})^{-1}$,

$$q(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) = p(\boldsymbol{\beta} \mid \bar{\mathbf{z}}, \bar{\mathbf{y}}, \bar{\mathbf{X}}) = \phi_{p(L-1)}(\boldsymbol{\beta} - \mathbf{V}\bar{\mathbf{X}}^\intercal \boldsymbol{\Lambda}^{-1}\bar{\mathbf{z}}; \mathbf{V}),$$

which follows by standard properties of multivariate normals applied to model (4). This means that in order to find the KL minimizer $q^*(\boldsymbol{\beta},\bar{\mathbf{z}}) = q^*(\boldsymbol{\beta} \mid \bar{\mathbf{z}}) \prod_{i=1}^N q^*(\bar{\mathbf{z}}_i)$ in $\mathscr{Q}_{\text{PFM-B}}$, we just have to find $q^*(\bar{\mathbf{z}}) = \prod_{i=1}^N q^*(\bar{\mathbf{z}}_i)$ minimizing $\text{KL}[q(\bar{\mathbf{z}}) || p(\bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}})]$. See [1] for further details. Due to the imposed factorization for $q^*(\bar{\mathbf{z}})$, which takes the name of 'mean-field approximation', the desired solution is known to satisfy the following mean field equations (see [7] for additional details):

$$\log q^*(\bar{\mathbf{z}}_i) \propto \mathbb{E}_{q^*(\bar{\mathbf{z}}_{-i})}[\log p(\bar{\mathbf{z}}_i \mid \bar{\mathbf{z}}_{-i}, \bar{\mathbf{y}}, \bar{\mathbf{X}})], \quad i = 1,\dots,n,$$

where the expectation is taken with respect to the distribution of all the $\bar{\mathbf{z}}_j$ other than $\bar{\mathbf{z}}_i$. Here, we show how this solution can be obtained by exploiting the dual hierarchical model (4). This gives further intuition and a broader understanding of the procedure. Moreover, besides specifying to the current setting the results derived in [1] under more general constructions, it allows computational gains due to the simplification of certain parameters, above all the covariance matrices in $q^*(\bar{\mathbf{z}}_i)$.

First, by marginalizing out $\boldsymbol{\beta}$ in (4b), we get $\bar{\mathbf{z}} \mid \bar{\mathbf{X}} \sim \text{N}_{n(L-1)}\left(\mathbf{0}, \boldsymbol{\Lambda} + \nu^2 \bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal\right)$, thus $\bar{\mathbf{z}} \mid \bar{\mathbf{y}}, \bar{\mathbf{X}} \sim \text{TN}_{n(L-1)}\left(\mathbf{0}, \boldsymbol{\Lambda} + \nu^2 \bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal; [0,\infty)^{n(L-1)}\right)$. Taking $\mathbf{V}$ as above and $\mathbf{H} = \boldsymbol{\Lambda}^{-1}\bar{\mathbf{X}}\mathbf{V}\bar{\mathbf{X}}^\intercal\boldsymbol{\Lambda}^{-1}$, by Woodbury's identity it holds $(\boldsymbol{\Lambda} + \nu^2\bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal)^{-1} = \boldsymbol{\Lambda}^{-1} - \mathbf{H}$ and so

$$\begin{aligned} p(\bar{\mathbf{z}}_i \mid \bar{\mathbf{z}}_{-i}, \bar{\mathbf{y}}) &\propto \exp\left\{-0.5\,\bar{\mathbf{z}}^\intercal(\boldsymbol{\Lambda} + \nu^2\bar{\mathbf{X}}\bar{\mathbf{X}}^\intercal)^{-1}\bar{\mathbf{z}}\right\} \mathbb{1}[\bar{\mathbf{z}}_i > \mathbf{0}] \\ &\propto \exp\left\{-0.5\,\bar{\mathbf{z}}^\intercal(\boldsymbol{\Lambda}^{-1} - \mathbf{H})\bar{\mathbf{z}}\right\} \mathbb{1}[\bar{\mathbf{z}}_i > \mathbf{0}] \\ &\propto \exp\left\{-0.5\,\bar{\mathbf{z}}^\intercal(\boldsymbol{\Lambda}_{[i,i]}^{-1} - \mathbf{H}_{[i,i]})\bar{\mathbf{z}} + \bar{\mathbf{z}}_i^\intercal \mathbf{H}_{[i,-i]}\bar{\mathbf{z}}_{-i}\right\} \mathbb{1}[\bar{\mathbf{z}}_i > \mathbf{0}], \end{aligned}$$

from which we get

$$q^*(\bar{\mathbf{z}}_i) \propto \exp\left\{-\frac{1}{2}\bar{\mathbf{z}}^\intercal\left(\boldsymbol{\Lambda}_{[i,i]}^{-1} - \mathbf{H}_{[i,i]}\right)\bar{\mathbf{z}} + \bar{\mathbf{z}}_i^\intercal \mathbf{H}_{[i,-i]}\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-i})}[\bar{\mathbf{z}}_{-i}]\right\} \mathbb{1}[\bar{\mathbf{z}}_i > \mathbf{0}],$$

which shows that $q^*(\bar{\mathbf{z}}_i)$ is the density of a normal random variable with parameters $\boldsymbol{\mu}_i^* = \boldsymbol{\Sigma}_i^* \mathbf{H}_{[i,-i]}\mathbb{E}_{q^*(\bar{\mathbf{z}}_{-i})}[\bar{\mathbf{z}}_{-i}]$ and $\boldsymbol{\Sigma}_i^* = \left(\boldsymbol{\Lambda}_{[i,i]}^{-1} - \mathbf{H}_{[i,i]}\right)^{-1}$, truncated above zero. In

order to obtain in practice the optimal PFM-B variational solution, one can resort to standard CAVI algorithms (see, e.g., [7]), as shown in detail in Algorithm 2. It is worth noting that, differently from Algorithm 1, in Algorithm 2 we only have to deal with expectations of $(L-1)$-variate truncated normals, significantly reducing the computational burden. After $q^*(\boldsymbol{\beta}, \bar{\mathbf{z}})$ has been computed, approximate posterior moments for $\boldsymbol{\beta}$ can be easily obtained leveraging on the law of total expectation as

$$\mathbb{E}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbf{V}\bar{\mathbf{X}}^\intercal \boldsymbol{\Lambda}^{-1}\mathbb{E}_{q^*(\bar{\mathbf{z}})}[\bar{\mathbf{z}}], \quad \text{var}_{q^*(\boldsymbol{\beta})}(\boldsymbol{\beta}) = \mathbf{V} + \mathbf{V}\bar{\mathbf{X}}^\intercal \boldsymbol{\Lambda}^{-1}\text{var}_{q^*(\bar{\mathbf{z}})}(\bar{\mathbf{z}})\boldsymbol{\Lambda}^{-1}\bar{\mathbf{X}}\mathbf{V},$$

while more complicated functionals can be computed with i.i.d. sampling, which, due to the particular block-diagonal structure of the resulting covariance matrix of the multivariate truncated normal distribution $q^*(\bar{\mathbf{z}})$, would require sampling only from multivariate truncated normals of dimension $L-1$.

4 Discussion

This article provides a novel derivation for the PFM-B variational method for a relevant class of MNP models with independent Gaussian priors. As shown in Section 3, this provides a novel understanding of the results in [1] as well as computational advantages in our special settings. Future works include deriving results for the posterior also for the dynamic MNP, which allows to model sequential decisions in the time series framework. In such a way, we plan to extend closed-form expressions, and the related samplers, for the filtering, predictive and smoothing distributions of multivariate dynamic probit models for binary time series in [10]. Finally, the variational Bayes approach in Section 3 can be generalized to perform inference for the smoothing distribution of the dynamic MNP extending the results in [11].

References

1. Fasano, A. and Durante, D.: A class of conjugate priors for multinomial probit models which includes the multivariate normal one. J. Mach. Learn. Res., **23**(30):1-16 (2022)
2. Greene, W. H.: Econometric analysis. Pearson Education India, (2003)
3. Hausman, J. A. and Wise, D. A.: Conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. Econometrica, **46**, 403–426 (1978)
4. Stern, S.: A method for smoothing simulated moments of discrete probabilities in multinomial probit models. Econometrica, **60**, 943–952 (1992)
5. Durante, D.: Conjugate Bayes for probit regression via unified skew-normal distributions. Biometrika, **106**, 765–779 (2019)
6. Azzalini, A. and Capitanio, A.: The skew-normal and related families. Cambridge University Press (2013)
7. Blei, D. M., Kucukelbir, A. and McAuliffe, J. D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc., **112**, 859–877 (2017)
8. Kullback, S. and Leibler, R. A.: On information and sufficiency. Ann. Stat., **22**, 79–86 (1951)
9. Fasano, A., Durante, D. and Zanella, G.: Scalable and accurate variational Bayes for high-dimensional binary regression models. Biometrika, in press (2022)
10. Fasano, A., Rebaudo, G., Durante D. and Petrone S.: A closed-form filter for binary time series. Stat. Comput., **31**:47 (2021)
11. Fasano, A. and Rebaudo G.: Variational inference for the smoothing distribution in dynamic probit models. Book of short papers - SIS 2021 (2021)

Mapping Indicators on the Unit Interval: the tipsae Shiny App

Mappare Indicatori sull'Intervallo Unitario: la tipsae Shiny App

Silvia De Nicolò and Aldo Gardini

Abstract The tipsae shiny app is a dedicated R-based tool for mapping proportions and indicators defined on the unit interval, working in the framework of small area estimation. It implements Beta-based Bayesian Hierarchical models defined at area level through Stan. A set of diagnostics, exploratory analysis and complementary tools complement the application.

Abstract *La shiny app tipsae si presenta come uno strumento, costruito in R, utile a mappare proporzioni o, più in generale, indicatori definiti nell'intervallo unitario nel contesto della stima per piccole aree. Permette di stimare, attraverso Stan, modelli gerarchici Bayesiani basati sulla distribuzione Beta a livello di area. L'applicazione fornisce diagnostiche specifiche, analisi esplorative ed altri strumenti complementari all'analisi.*

Key words: Bayesian Inference, Beta Regression Models, Small Area Estimation, R Tools

1 Introduction

Timely and reliable statistical estimates at a great level of disaggregation are increasingly in demand and require an extensive exploitation of survey data. Nonetheless, domains or areas of study are often different from the ones for which the survey was originally planned, leading to unreliable survey estimates due to observations-poor and possibly non-representative samples.

Silvia De Nicolò
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Cesare Battisti 241, 35121 Padova; e-mail: silvia.denicolo@phd.unipd.it

Aldo Gardini
Dipartimento di Scienze Statistiche "P.Fortunati", Alma Mater Studiorum Università di Bologna, Via Belle Arti 41, 40126 Bologna; e-mail: aldo.gardini2@unibo.it

Small Area Estimation (SAE) techniques exploit auxiliary information to borrow strength across areas and produce estimates of interest with an acceptable level of uncertainty. Specifically, the area-level class of SAE models maps survey estimators of target quantities to areas-specific covariates, generally measured without error (e.g. census data), via an explicit regression model.

We focus on unit interval responses, common in SAE modelling because of the high presence of rates and proportions releases in official statistics. Two different bodies of literature relate with this field, revolving around linear mixed models with suitable transformations [11] and Beta regression models [8]. Although several routines for SAE have been released by developer teams in `R`, only the `emdi` package [9] directly accounts for unit interval responses at area-level via Gaussian models with proper transformations. Furthermore, Beta-based small area models lack of a proper implementation and the `tipsae` package, available on CRAN [4], aims at filling this gap.

We implement area-level models based on the Beta likelihood comprising the standard Beta regression model, Zero and/or One Inflated Beta [14] and Flexible Beta [3] models. Moreover, particular dependence structures can be modelled, including spatial and/or temporal random effects. We decided to operate in a Bayesian fashion via `Stan` routine [2] to easily manage non-Gaussian assumptions, ease the out-of-sample areas treatment, and capture the uncertainty about target parameters through posterior inference. Within this framework, we developed a Shiny application to further facilitate the workflow for non-expert users.

Our application assists the user in carrying out a complete SAE analysis, starting from data loading step, tracking the entire process of data exploration, model estimation and validation, presentation and exportation of results. This allows to straightforwardly use Bayesian models and complex SAE methods.

2 Beta-Based Small Area Models

A classical Beta small area model for $y_d \in (0,1)$, denoting the direct estimator of a generic target quantity and $\boldsymbol{x}_d$ being a set of P covariates for domain d, constitutes as a hierarchical model with two levels. The sampling level models the conditional distribution of the direct estimator as

$$y_d|\theta_d, \phi_d \overset{ind}{\sim} Beta(\theta_d\phi_d, (1-\theta_d)\phi_d), \quad \forall d.$$

where $\mathbb{E}[y_d|\theta_d, \phi_d] = \theta_d$ is the target parameter and is estimated via a logit regression at the linking level, i.e. $\text{logit}(\theta_d)|\boldsymbol{\beta}, v_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + v_d$, with $v_d|\sigma_v^2 \overset{ind}{\sim} \mathcal{N}(0, \sigma_v^2)$ an area specific random effect. Generally, a small area Beta model assumes the dispersion parameter ϕ_d as known in order to allow identifiability. In our package, we contemplate alternative likelihood assumptions at the

sampling level. The Zero/One Inflated Beta model extends the support of y_d to zero/one values, by assuming a mixture of a Beta and Dirac Delta components. Moreover, the Flexible Beta distribution, defined as a mixture of two Beta components, allow to improve the modelling of target quantities with skewed and heavy tailed distributions [10].

3 Workflow

The application can be launched without any preliminary action by running the following commands.

```
R> library(tipsae)
R> runShiny_tipsae()
```

A browser window pops-up, allowing users to navigate on the application. The interface is organized into 5 main sections briefly described in what follows.

The **Home** page comprises a schematic description of the application with some relevant references.

The **Data** page is divided into 4 subsections: the first 3 tabs concern data entry steps and data pre-treatment, while the latter provides graphical exploratory tools. In the *Loading Data* tab, a CSV file can be loaded and the data input may be visualized through the button "View loaded dataset". Additional information about data must be filled, such as the nature of the inserted variables, labeling the response, the covariates and the dispersion values, specifying also if it is the variance or the effective sample size. Here, it is also possible to set up the smoothing procedure, if needed. Other information can be specified, such as a possible time variable, domain ids and sample sizes, useful for following visual diagnostics. When a smoothing procedure is set up, the *Smoothing* subsection enables to change settings and graphically visualize its output. The procedure is a Generalized Variance Function smoothing technique [6]; for a methodological explanation, refer to the package vignette. Note that only a simplified smoothing procedure is provided in the Shiny app, if compared to the package function: it automatically assumes the response variable as a proportion, employing the related variance function. The user can choose between `"ols"` and `"gls"` regression type for smoothing. In subsection *Load Shapefile*, a spatial structure can be incorporated, loading either a SHP file either an RDS file containing a `SpatialPolygonsDataFrame` object. Such object would enable to account for spatial dependencies in the model and/or plot maps with relevant quantities. The last subsection, called *Data Summary*, provides an accurate data exploration before moving to the modelling step, depicting the distribution of the response variable and its relationship with the covariates and the dispersion measure.

The **Model Fitting** section implements the estimation of mentioned Bayesian models, via Stan routine [2]. It provides an efficient Hamiltonian Monte Carlo (HMC) fitting algorithm and customized parallel computing.

The *Model Specification* tab enables to set the model likelihood: our application automatically constrains the model choice among those allowed by the input data. When any response observation $y_i \in (0, 1)$, $\forall i$ the allowed options are Beta and Flexible Beta models. If some observations are recorded as zeros and/or ones, the application automatically restricts the choice to the Zero and/or One Inflated Beta models. Secondly, the prior for the random effect can be chosen among the standard Gaussian one, a robust alternative following a Student's t distribution [5] and a shrinkage distribution, the variance gamma [7], for a more flexible handling. When the data incorporates a time variable, therefore having a panel nature, the model automatically incorporates a temporal dependency structure, adding to the linear predictor an additional random effect with a random walk prior. Moreover, if a `SpatialPolygonsDataFrame` object related the to areas of interest has been loaded, the application asks whether to include a spatial dependency structure into the model or not. The dependecy is incorporated via an additional random effect in the linear predictor with a intrinsic conditional autoregressive (ICAR) prior.

Some algorithm settings can also be modified in the *Settings about the MCMC Algorithm* tab, such as the number of iterations per chain, including warm-up (half of total iterations), the parallel computation to switch on with a proper tick, the number of chains, and the number of cores. Additional HMC options are the maximum allowed tree depth (Maximum treedepth) and the target average proposal acceptance probability (adapt_delta): refer to the Stan documentation for further details. The "Fit Model" button allows the user to start the estimation, whose progress is depicted by an iterative printed output.

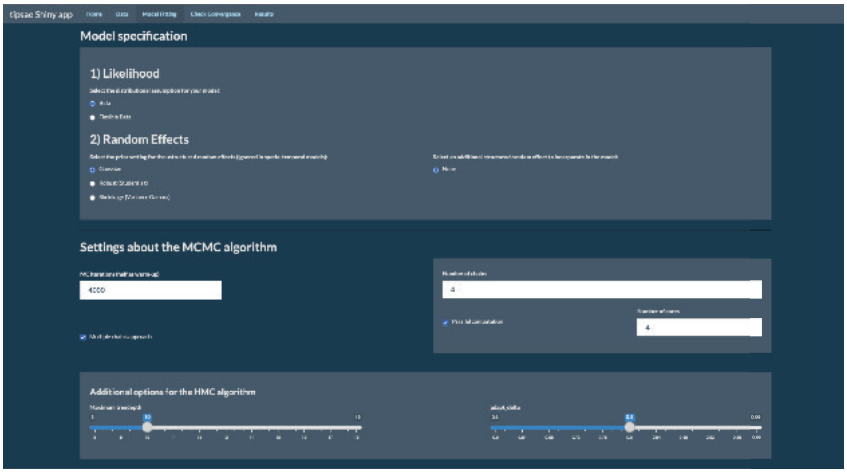


Fig. 1 Model Fitting Section of tipsae Shiny App

Once computations are completed, the mixing of the MCMC algorithm can be checked through graphical tools within the **Check Convergence** section. The posterior densities, chains trace-plots, autocorrelation functions and rank plots can be visualized for any linking level parameter.

Eventually, the model results can be accessed in the **Results** section. The *Model Summaries* subsection provides posterior syntheses of regression coefficients and random effects. Furthermore, a summary of residuals is also reported, including an histogram, and the LOO Information Criterion can be computed through the button "Click to compute LOOIC", enabling for model selection. This criterion is based on approximate leave-one-out cross-validation computed using Pareto-smoothed importance sampling [13]. The *Posterior Predictive Check* subsection displays the sample data kernel density versus those of the datasets generated from the posterior predictive distribution, denoted with $Y_d^\bullet|\boldsymbol{y},\ d = 1, \ldots, D$, in order to assess the goodness of fit. Here, a specific tab focuses on area-specific Bayesian p-values, defined as $BP_d = \mathbb{P}\left[Y_d^\bullet > y_d|\boldsymbol{y}\right]$ [6]. In absence of systematic deviations, the expected Bayesian p-value is 0.5, whereas values near 0 or 1 highlight issues of over-estimation and under-estimation, respectively.

Small area specific diagnostics have a proper subsection (*SAE diagnostics*), visually illustrating the shrinking process induced on model-based estimates and comparing direct and model-based estimates. The standard deviations of both type of estimates are compared and summaries of a measure of standard deviation reduction are provided. The *Random Effect* subsection compares the standardized effects density versus the one of a standard normal and a caterpillar plot, comparing their posterior distributions for each area. Lastly, the *Model Estimates* subsection displays a table with direct and model-based estimates, including relevant posterior summaries of target parameters. Such object can be downloaded in CSV format via a proper button. A caterpillar plot of the target parameter posteriors is also provided.

4 Conclusions

The tipsae shiny app is a dedicated tool for mapping proportions and indicators defined on the unit interval, widely used to measure, for instance, unemployment or disease prevalence. Additional features to be integrated in future releases may be the implementation of shrinking priors for the regression coefficients, allowing for variable selection when several covariates are employed. Other directions may focus on model extensions for variance shrinking [12], able to relax the assumption of known dispersion parameter, and for covariates measured with error [1]. Notice that the `Shiny` app does not include all the package tools. Specifically, the benchmarking procedure has not been implemented and the smoothing procedure does not include as option the `"kish"` method. However, such features may be included in future releases of the package.

References

[1] Arima, S., Datta, G.S., Liseo, B.: Bayesian estimators for small area models when auxiliary information is measured with error. Scandinavian Journal of Statistics **42**(2), 518–529 (2015)

[2] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1), 1–32 (2017)

[3] De Nicolò, S., Ferrante, M.R., Pacei, S.: Put Inequality on the Map: Small Area Estimation of Inequality Measures using a Beta Mixture (2021). Working Paper

[4] De Nicolò, S., Gardini, A.: tipsae: Tools for Handling Indices and Proportions in Small Area Estimation (2022). URL `https://CRAN.R-project.org/package=tipsae`. R package version 0.0.4

[5] Fabrizi, E., Ferrante, M., Trivisano, C.: Hierarchical beta regression models for the estimation of poverty and inequality parameters in small areas. Analysis of Poverty Data by Small Area Methods. John Wiley and Sons pp. 299–314 (2016)

[6] Fabrizi, E., Ferrante, M.R., Pacei, S., Trivisano, C.: Hierarchical bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. Computational Statistics & Data Analysis **55**(4), 1736–1747 (2011)

[7] Fabrizi, E., Ferrante, M.R., Trivisano, C.: Bayesian small area estimation for skewed business survey variables. Journal of the Royal Statistical Society: Series C (Applied Statistics) **67**(4), 861–879 (2018)

[8] Janicki, R.: Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. Communications in Statistics-Theory and Methods **49**(9), 2264–2284 (2020)

[9] Kreutzmann, A.K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., Tzavidis, N.: The R package emdi for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software **91**(7), 1–33 (2019). DOI 10.18637/jss.v091.i07

[10] Migliorati, S., Di Brisco, A.M., Ongaro, A.: A new regression model for bounded responses. Bayesian Analysis **13**(3), 845–872 (2018)

[11] Rao, J.N., Molina, I.: Small-area estimation. Wiley Series in Survey Methodology (2015)

[12] Sugasawa, S., Tamae, H., Kubokawa, T.: Bayesian estimators for small area models shrinking both means and variances. Scandinavian Journal of Statistics **44**(1), 150–167 (2017)

[13] Vehtari, A., Gelman, A., Gabry, J.: Practical bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and Computing **27**(5), 1413–1432 (2017)

[14] Wieczorek, J., Nugent, C., Hawala, S.: A bayesian zero-one inflated beta model for small area shrinkage estimation. In: Proceedings of the 2012 Joint Statistical Meetings, American Statistical Association, Alexandria, VA (2012)

A Bayesian spatio–temporal model of PM_{10} pollutant in the Po Valley

Un modello bayesiano spazio–temporale per l'inquinante PM_{10} nella Pianura Padana

Matteo Gianella, Alessandra Guglielmi and Giovanni Lonati

Abstract Po Valley is well known to be one of the most polluted areas in Italy, due to its population density, its shape and climate. Thus, there is obvious interest in monitoring the air quality in several stations scattered in the whole territory. In this work, we develop a Bayesian spatio–temporal model for the analysis of time series of PM_{10} concentration in Po Valley to assess whether the station features have an effect in explaining the spatial time series. In particular, we focus on the Emilia Romagna region and we rely on *Stan* for posterior inference.

Abstract *La Pianura Padana è nota per essere una delle regioni più inquinate d'Italia, a causa della elevata densità abitativa, la sua forma e clima. Vi è quindi un ovvio interesse nel monitorare la qualità dell'aria attraverso numerose stazioni disseminate per tutto il territorio. In questo lavoro proponiamo un modello bayesiano spazio–temporale per l'analisi di serie temporali della concentrazione di PM_{10}. Ci concentriamo, in particolare, sull'Emilia Romagna e ci affidiamo a Stan per il calcolo dell'inferenza a posteriori.*

Key words: Bayesian inference, Gaussian processes, harmonic regression

1 Introduction

Pollution is defined as the introduction of contaminants into the natural environment which cause adverse change. Pollution can take the form of any substance (solid, liquid, or gas) or energy (such as radioactivity, heat, sound, or light). By *air pollution*, in particular, we mean the release by human activities of gases and particulates

Matteo Gianella[1], Alessandra Guglielmi[1] and Giovanni Lonati[2]

[1] Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: {matteo.gianella,alessandra.guglielmi}@polimi.it

[2] Department of Civil and Environmental Engineering, Politecnico di Milano, Milano, Italy
e-mail: giovanni.lonati@polimi.it

into the atmosphere. Common gaseous pollutants include carbon monoxide, sulfur dioxide, chlorofluorocarbons (CFCs) and nitrogen oxides produced by industry and motor vehicles. Photo–chemical ozone and smog are created as nitrogen oxides and hydrocarbons react to sunlight. Particulate matter, or fine dust is characterized by their micrometre size PM_{10} to $PM_{2.5}$.

Po Valley is a well-known hotspot for PM pollution in Europe [4]. Many factors might be responsible: an elevated population density, together with a high level of urban and industrial areas, its geographic shape and climate. In particular, the population density and the high number of industries are the main contributors to the high level of emissions in this area, while its shape and climate prevent an adequate dispersion of the emitted pollutants. In this work, we consider air quality data of the Po Valley, which includes hourly and daily averaged concentrations for several pollutants. Such data are collected via fixed monitoring stations at 280 sites distributed over four regions. The management and data collection policies are specific to every region. Data of this type are certainly *spatial*, since they have an intrinsic spatial dependence one may want to investigate, whilst being *multiple time series*, as for each location we have a multi–dimensional time series, one for each pollutant.

Spatial models for point–referenced data are usually modelled via Gaussian Processes (GPs). Since learning a GP is computationally intensive as the number of data points increases, a recent interest to provide alternatives that simplify the computational costs has soared up. Another approach comes from Gaussian Random Markov Fields and its link with Gaussian processes [6]. (Multiple) Time Series analysis in a Bayesian setting (see [5]) is also an interesting research field, for which several different approaches are available: CAR models of various orders, non–parametric dynamic AR models or models built on the frequency domain (spectral models).

In this paper, we propose a Bayesian model for the analysis of spatial time series of air pollutants in the Po Valley. We focus only on PM_{10} pollutant to understand if particular features of the monitoring site affect the estimation of the time series and spatial association of the daily averaged concentrations. The latent spatial and temporal processes are modelled separately via Gaussian Processes with Matérn covariance functions and harmonic regression models, respectively. The paper is organized as follows: Section 2 provides a short description of the dataset; then, in Section 3 we describe the Bayesian model we assume for our data and in Section 4 we give initial findings based on posterior inference along with possible future research directions and developments.

2 The dataset

The dataset we consider for the analysis comes from the air quality monitoring network of the Po Valley, the area of interest. This network is managed autonomously by the Regional Agencies of Environmental Protection (ARPA) of Emilia Romagna, Lombardy, Piedmont and Veneto. Data about concentrations of pollutants are col-

lected through fixed monitoring stations, distributed one the whole valley. Each station collects, through time, the concentration of many pollutants, in particular: nitrogen dioxide (NO_2), benzene (C_6H_6), ammonia (NH_3) and particulate matter (PM_{10}, $PM_{2.5}$).

Since each regional agency is independent, we may observe the same data collected in different time scales. For instance, data about NO_2 are available as hourly averages, except for Piedmont, where we have daily averaged concentrations. Particulate matter is collected as daily averages in all regions. The dataset also provides some details on the station recording the concentration time series. For each station, we have information about its location (lat–long coordinates), the main emission source to which the given station is exposed to (e.g., emissions from traffic, industry, etc.) and the level of urbanization of the surroundings (e.g., urban, suburban, rural areas). Data also take into account specific features of the territory of interest, thus distinguishing different zones, specific for each region, due to the intrinsic diversity among the Po Valley. Exploratory data analysis of the whole dataset is in [3], where the authors investigate the effects of COVID–19 lockdown policies in the first semester of 2020.

3 The Bayesian model

Monitoring processes in natural and environmental sciences often results in the collection of spatial or spatio–temporal data [1]. In this context, the objectives of the analysis include the estimation over spatial locations and over time. As mentioned before, we consider time series for a single concentration of pollutant (PM_{10}) in all stations in Emilia Romagna. We denote by $y(\boldsymbol{s}, t)$ the measurement of a given quantity collected at location $\boldsymbol{s}$ at time t, for $\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^d$, with $d = 2$ and for $t \in [0, T] \subset \mathbb{R}^+$. A spatio–temporal regression model for point–referenced data on a spatial domain $\mathcal{D}$ and temporal domain $[0, T]$ can be assigned, in the most general case, as follows:

$$y(\boldsymbol{s}, t) = \boldsymbol{x}(\boldsymbol{s}, t)^T \boldsymbol{\beta}(\boldsymbol{s}, t) + w(\boldsymbol{s}, t) + \epsilon(s, t),$$

where $\boldsymbol{x}(\boldsymbol{s}, t)$ is a $p \times 1$ vector of predictors, $\boldsymbol{\beta}(\boldsymbol{s}, t)$ is the $p \times 1$ vector of regression coefficients, $w(\boldsymbol{s}, t)$ is a spatio–temporal process and $\epsilon(\boldsymbol{s}, t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$ is the random effect, usually in charge of capturing local variability, which can not be detected by the process w neither in space nor in time. As a result, we also need a model for $w(\boldsymbol{s}, t)$. For this purpose, we denote by α the denote temporal effects and w the spatial effect. We consider the following separable model (see [1]) for $w(\boldsymbol{s}, t)$:

$$w(\boldsymbol{s}, t) = \alpha(t) + w(\boldsymbol{s}). \tag{1}$$

Both $\alpha(t)$ and $w(\boldsymbol{s})$ are assumed to be independent from $\epsilon(\boldsymbol{s}, t)$. Equation (1) provides a separable (additive) assumption in the spatio–temporal effect. Non–separable spatio–temporal modeling requires the choice of a specific relationship to connect

space and time scales, which is non trivial and beyond the scopes of this work. Hence, providing a model for our spatio–temporal dataset boils down to the model specification for the temporal process $\alpha(t)$ and the spatial process $w(\boldsymbol{s})$.

Model for the temporal process The concentration of air pollutants often exhibit some periodicity, at least seasonal, mainly due to the different dispersion features of the atmosphere, but also to the sources' activity (e.g., domestic heating). It is then straightforward, at least for an initial modelling, characterize the process using harmonic regression. *Harmonic regression* refers to (usually linear) models describing periodicities in data by means of sinusoidal functions. Standard Fourier representation of periodic functions, in fact, proves useful in modeling series with persistent periodic patterns, especially in dealing with seasonal phenomena. Given an integer period p, a general time series has an exact representation, obtained as super–position of $m = \lfloor p/2 \rfloor$ Fourier harmonics. Thus, for a zero–mean time series as $\alpha(t)$, an appropriate model is given by

$$\alpha(t) = \sum_{k=1}^{m} a_k \cos\left(\frac{2\pi k}{T} t\right) + b_k \sin\left(\frac{2\pi k}{T} t\right).$$

In a Bayesian setting, we need to specify the marginal prior for $\{(a_k, b_k)\}_k$ parameters to assess the contribution of each harmonic in the regression.

Model for the spatial process The most common choice for the spatial process $\{w(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{D}\}$ in (1) is the Gaussian Process. Gaussian Processes are specified with a covariance function $Cov[w(\boldsymbol{s}), w(\boldsymbol{s}')] = \Sigma_\theta(\boldsymbol{s}, \boldsymbol{s}')$ for any two points $\boldsymbol{s}$ and $\boldsymbol{s}'$ in $\mathcal{D}$. If $\mathcal{U} = \{u_i\}_{i \in 1,\dots,n}$ and $\mathcal{V} = \{v_j\}_{j \in 1,\dots,m}$ are two finite sets, we define $\Sigma_\theta(\mathcal{U}, \mathcal{V})$ the $n \times m$ matrix such that $[\Sigma_\theta(\mathcal{U}, \mathcal{V})]_{ij} = \Sigma_\theta(u_i, v_j)$. For our spatial process $w(\boldsymbol{s})$ we assume a centered Gaussian Process of variance Σ_θ, i.e.

$$w(\boldsymbol{s}_i) \overset{\text{iid}}{\sim} GP(\mathbf{0}, \Sigma_\theta) \quad \Longleftrightarrow \quad (w(\boldsymbol{s}_1), \dots, w(\boldsymbol{s}_n)) \sim \mathcal{N}_n(\mathbf{0}, \Sigma_\theta) \quad \text{for all } n.$$

In this definition, we require Σ_θ to be a valid spatial covariance function for the process. Validity is attained ensuring that $\Sigma_\theta(\mathcal{U}, \mathcal{U})$ is positive definite for any finite set $\mathcal{U}$. As a consequence, it is not straightforward to define a valid covariance function, so we rely on standard choices. Further details can be found, for instance, in [1] and references therein. In particular, for our process, we choose the *exponentiated quadratic kernel* (i.e., the limit of a Matérn kernel $C_\nu(\boldsymbol{s}_i, \boldsymbol{s}_j)$ when ν tends to $+\infty$), defined as

$$\Sigma_{(\sigma^2,\, \rho)}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma^2 \exp\left\{-\frac{1}{2\rho^2} \|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2\right\}$$

In this case, random parameters are σ^2 and ρ^2, both required to be greater than 0.

Prior The model just described is then completed with suitable marginal prior distributions for each parameter involved in the spatial and temporal process just described. We assume independence between blocks of parameters. We use a rather vague prior for each component of the regression coefficients' vector $\boldsymbol{\beta}$ and for

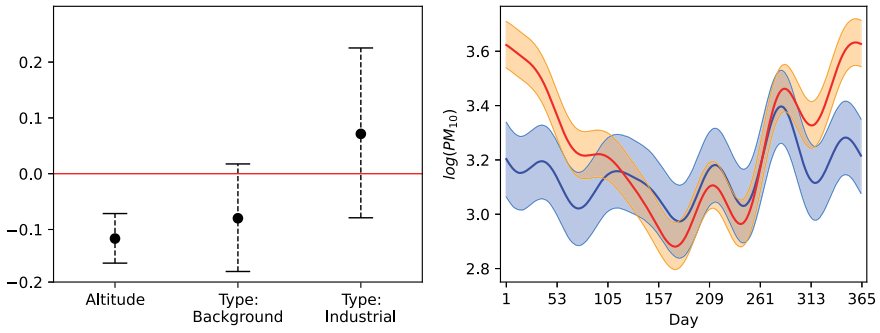


Fig. 1 95% credible intervals for regression coefficients $\boldsymbol{\beta}$ (left); 95% credible bands for the estimate of the temporal process via harmonic regression for rural (blue) and non–rural (red) stations (right)

the harmonic regression pairs $\{(a_k, b_k)\}_k$, assuming a $\mathcal{N}(0, 10^2)$ distribution; the variance of the random effect τ^2 is set in order to avoid infinite variances, thus setting $InvGamma(3, 2)$ as its marginal prior distribution; finally, for the parameters involved in the spatial process, i.e., σ^2 and ρ^2, we set a vague prior on σ^2 and a more informative one to ρ^2. This choice is made to avoid well–known model issues that prevent spatial models to be completely identifiable (see [1]). In particular, since we are more interested in the spatial range associated with the covariance kernel, we keep more care in defining a prior distribution for this parameter. We then set an $\mathcal{U}(0, 1)$ distribution for σ and a $Beta(2, 5)$ for ρ. The supports of those marginal priors matches empirical estimates based on likelihood maximization whilst ensuring enough flexibility.

4 Posterior Inference

At this stage of research, we focus only on the Emilia Romagna region during 2018. This reduced dataset contains daily averaged PM_{10} concentrations in 49 stations for every day of the year (thus $T = 365$). As regressors of interest, we consider the *Altitude* of the station and the most probable source of pollution (*Traffic*, *Industrial* or *Background* in case the source of pollution can not be easily detected).

As mentioned before, we rely on *Stan* probabilistic programming language [2] for posterior simulation. This language, in fact, is flexible enough to quickly set up the model here discussed and carry out posterior inference in reasonable time. As future development, we will consider *ad–hoc* sampling strategies to speed up computation even with more complex models. We run four MCMC chains in parallel in *Stan*, each of them for a total of $10,000$ iterations, discarding the first $6,000$ as burn–in and thinning the chain every two iterations. Hence, the final sample size is $2,000$.

Figure 1, left panel, displays the credible intervals at 95% for the regression coefficients. We find that all regressors are quite significant at this level, especially the *Altitude* seems to play a relevant role in the linear predictor, as expected: air

quality improves with altitude. Note that the *Type: Industrial* regressor exhibit a wider variance than *Type: Background* since the number of stations marked as *Background* is small. Moreover, since the base category is *Type: Traffic*, posterior inference highlights that air quality worsen in presence of Industrial areas. Figure 1, right panel, shows the posterior estimate of the harmonic regression for the temporal process in case the station is located in a rural area (blue band) and non–rural area (red band). In non–rural areas, usually cities or industrial sites, the curve exhibit an interesting U–shape that highlights higher PM_{10} concentration levels in cold seasons. In rural stations, on the other hand, we observe a flatter curve and those areas are generally characterized by a better air quality. As long as the spatial process $w(\boldsymbol{s})$ is concerned, we report the 95% marginal posterior credible interval for ρ, which is a indicative of the level of spatial correlation estimated by our model, that is $[0.001, 0.011]$ with posterior median 0.005. This means that the spatial correlation is weak. This can be motivated by the relatively high regional background of PM_{10} over the whole Po valley and even in rural areas, with lower concentration levels only observed at mountain sites.

Future work includes the assumption of a Bayesian non–parametric model which aims at clustering stations according to the functional shape of the temporal process. Moreover, we plan to further investigate the spatial effect in the model.

Aknowledgements

We would like to thank Camilla Battistini, Michela Frigeri, Francesca Pessina, Michael Ronzulli, Claudia Speranza and Luigi Umana, M.Sc. students in Mathematical Engineering at Politecnico di Milano, for *Stan* code implementation and summary plots in Section 4.

References

1. Banerjee, S., Carlin, B.P., Gelfand, A.E.: *Hierarchical modeling and analysis for spatial data*. Chapman and Hall / CRC (2015)
2. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017)
3. Lonati, G., Riva, F.: Regional Scale Impact of the COVID-19 lockdown on air quality: gaseous pollutants in the Po Valley, Northern Italy. *Atmosphere* **12**(2), 264 (2021)
4. Masiol, M., Squizzato, S., Formenton, G., Harrison, R.M., Agostinelli, C.: Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the Veneto region, NE Italy. Science of The Total Environment **576**, 210–224 (2017)
5. Prado, R., West, M.: *Time series: modeling, computation, and inference*. Chapman and Hall / CRC (2010)
6. Rue, H., Held, L.: *Gaussian Markov random fields: theory and applications*. Chapman and Hall / CRC (2005)

Construction of a proper prior for a Bayesian Envelope Model

Costruzione di una prior propria per un modello Envelope bayesiano

Andrea Mascaretti

Abstract Envelope models are multivariate linear regression techniques that aim at reducing the variance of the estimator. Bayesian envelopes allow to quantify the uncertainty of inference by means of the posterior distribution. In this work, we construct a proper prior distribution and compare it to the existing literature. A prior sensitivity analysis is conducted, yielding similar results.

Abstract *I modelli envelope sono una particolare tipologia di regressione lineare multivariata finalizzata a ridurre la varianze degli stimatori. La formulazione bayesiana di questi modelli consente di quantificare direttamente l'incertezza degli stimatori mediante l'analisi della posterior. In questo lavoro, proponiamo una prior propria per il modello e valutiamo l'impatto della prior sull'inferenza, ottenendo risultati comparabili alle proposte presenti in letteratura.*

Key words: envelope models, bayesian statistics

1 Response Envelopes

Envelopes [2, 1] are a class of models aimed at increasing the efficiency of multivariate regression by exploiting the relations between response and predictors that affect the accuracy of the results and are not taken into account by standard methods. Within the usual multivariate regression setting, the expected value of a random variable $Y \in R^r$ is given a functional form such that we get

$$Y_i = \mu + \beta X_i + \varepsilon_i, \;\; i = 1, \ldots, n, \tag{1}$$

Andrea Mascaretti
University of Padova, Via Cesare Battisti, 241, 35121, Padova (PD), Italy, e-mail: mascaretti@stat.unipd.it

where $\{X_i\}_{i=1}^n$ is a sequence of non-stochastic vectors, with $X_i \in R^p$ for $i = 1,\dots,n$, the errors are independent and identically distributed multivariate normal vectors with zero mean and covariance Σ, $\mu \in R^r$ is an unknown vector of intercepts and $\beta \in M_{(r,p)}$ (where $M_{(a,b)}$ denotes the space of real matrices of dimensions (a,b)) is the unknown matrix of regression coefficients. For simplicity (and without loss of generality), we assume that the predictors are centred, $\sum_{i=1}^n X_i = 0$. Moreover, let Y be the $(n \times r)$ matrix of rows $(Y_i - \bar{Y})^T$, where $\bar{Y}$ is the sample mean, and Y_0 be the non-centred matrix. In a similar fashion, let $X = \left\{X_i^T\right\}$ be the matrix of the predictors, $S_{Y,X} = n^{-1}Y^T X$ and $S_X = n^{-1}X^T X$. The maximum likelihood estimator,

$$\hat{\beta} = S_{Y,X} S_X^{-1}, \tag{2}$$

is incidentally equal to the ordinary least squares estimator. From Eq. 2, we notice that this is akin to performing r separate univariate regressions: one for every element of Y on X. Inference on $\beta_{j,k}$, the (j,k)th element of β is the same we would obtain by constructing a univariate model. The model in Eq. 1 becomes operational when inference is conducted simultaneously on different rows of β or various elements of Y jointly.

The intuition behind envelope models is that there might be linear combinations of the response vectors whose distribution is invariant with respect to the non-stochastic predictors. Explicitly modelling for this property allows to obtain estimator whose variance is reduced. We call such linear combinations of Y X-invariant. Notice that for a linear transformation $G \in M_{(r,q)}$, with $q \leq r$, if $G^T Y$ is invariant, then also $A^T G^T Y$ has the same property for any non-stochastic matrix $A \in M_{(q,q)}$. In other words, only span (G) is identifiable.

From a mathematical point of view, this is equivalent to assuming the existence of two matrices Γ and Γ_0 such that $O = [\Gamma\ \Gamma_0]$ is orthogonal. We obtain

1. $\Gamma_0^T Y | X \sim \Gamma_0^T Y$
2. $\Gamma^T Y \perp \Gamma_0^T Y | X$

The conditions above entail that span $(\beta) \subseteq$ span (Γ) and $\Sigma = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$, where $P_{(\cdot)}$ is the orthogonal projector operation on a space and $Q_{(\cdot)} = I - P_{(\cdot)}$ is the projection on the orthogonal space. In this scenario, span (Γ) is a reducing subspace of Σ ([2]). The Σ-envelope of $\mathscr{B} =$ span (β), $\mathscr{E}_\Sigma(\mathscr{B})$, is the smallest reducing subspace of Σ that contains $\mathscr{B}$.

Model in Eq. 1 can be rewritten as

$$Y_i = \mu + \Gamma \eta X_i + \varepsilon, \tag{3}$$

where $\beta = \Gamma\eta$, $\Gamma \in M_{(r,u)}$ is an orthogonal basis of $\mathscr{E}_\Sigma(\mathscr{B})$ and u is the dimension of the envelope $\mathscr{E}_\Sigma(\mathscr{B})$. Moreover, the variance is $\Sigma = \Sigma_1 + \Sigma_2 = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$, where $\Omega \in M_{(u,u)}$ and $\Omega_0 \in M_{(r-u,r-u)}$ are two diagonal matrices carrying the coordinate information with respect to the basis Γ and Γ_0.

1.1 Bayesian Envelopes

The only contribution, to the best of our knowledge, on Bayesian envelopes models is [3]. The rationale behind Bayesian envelopes is that it allows to quantify the uncertainty of the predictions by computing the posterior distribution (as opposed to bootstrap or asymptotic considerations), as well as extending the model to the cases where $n < r$. Moreover, prior information can be incorporated into the learning process, be it on the values of the parameters or to induce sparsity or other desirable properties. As for the selection of u, the dimension of the envelope, [3] adopt a Deviance Information Criterion to obtain the best value, in lieu of the Likelihood Ratio Tests used within the frequentist framework. The interest in obtaining a proper prior distribution for a Bayesian envelope stems from the fact that this is a prerequisite to extend it to more complex scenarios, such as mixtures or nonparametric formulations. In our work, we build on [3] and change the prior on the mean μ. Notice that we require that the mean component be separated from the regression coefficients because we model our data up to a translation: assuming that a subset of Y is X-invariant does not imply that it has zero mean. We also assume a priori independence, which seems logical for the mean value of Y should not be influenced by the values of X a priori. Whereas [3] employ an improper prior for the data, *i.e.* they set $\pi(\mu) \propto 1$ on the support of μ, our proposal relies on a proper prior distribution.

The prior distribution is defined on the parameters $(\mu, \eta, (\Gamma, \Gamma_0), \Omega, \Omega_0)$. Notice that, for identifiability, we constrain Ω and Ω_0 to be diagonal matrices with entries disposed in decreasing order. This is equivalent to post-multiplying Γ and Γ_0 by the matrices of eigenvectors of the original Ω and Ω_0. From a mathematical point of view, this is equivalent to fix Γ and Γ_0 to be bases of the envelope and, thus, as elements of a subset of a Stiefel manifold restricted to have that the maximum element for each column as positive sign, denoted by $S^+_{(\cdot,\cdot)}$. In this respect, we notice that the Stiefel manifold of arbitrary finite dimensions (a,a) is a compact unimodular group with a unique Haar measure, which induces a measure on $S_{(a,b)}$ and $S^+_{(a,b)}$.

The parameter space is then given by $M_{(r,1)} \times M_{(u,p)} \times S^+_{(r,r)} \times O_u \times O_{r-u}$, where O_a is the set of diagonal matrices of dimension a with entries disposed in decreasing order.

We define the prior on the parameters as follows:

1. μ is set to be independent from the other parameters. We endow it with a multivariate normal distribution, so that $\pi(\mu) = \mathscr{N}_r(\mu_0, \Sigma_0)$,
2. The conditional prior on η is a matrix normal:

$$\pi(\eta | (\Gamma, \Gamma_0, \Omega, \Omega_0)) = \mathscr{N}_{(u,p)}\left(\Gamma^T, \Omega, C^{-1}\right),$$

   where C^{-1} is a positive definite matrix in $M_{(p,p)}$.
3. The prior on $O = (\Gamma, \Gamma_0)$ is a matrix Bingham distribution with parameters G and D, where G is a positive semi-definite matrix in $M_{(r,r)}$ and D is in O_r with positive entries. Thus, $\pi(O) = \mathscr{B}_{(r,r)}\left(G, D^{-1}\right)$. The density is proportional to $\exp\left\{(-1/2)\operatorname{tr}\left(D^{-1}O^T GO\right)\right\}$

4. Denoting by ω and ω_0 the diagonal vectors of, respectively, Ω and Ω_0, we assume that, a priori, they are distributed as order statistics of u and $r-u$ independent and identically distributed observations from Inverse-Gamma distributions of shape and rate parameters α, ψ and α_0, ψ_0.

Notice that the main difference between our work and [3] is the prior on μ. From a computational point of view, this means that the structure of the Gibbs sampler is similar, the only difference being the structure of the full-conditional for μ, which can be easily computed to be of the form

$$\pi\left(\mu|\eta,\left(\Gamma,\Gamma_0\right),\omega,\omega_0,Y\right)=\mathscr{N}_r\left(\mu_c,\Sigma_c\right), \tag{4}$$

where

$$\Sigma_c=\left(\Sigma_0^{-1}+\left(\frac{\Sigma}{n}\right)^{-1}\right)^{-1},$$

and

$$\mu_c=\Sigma_c\left(\Sigma_0^{-1}\mu_0+\left(\frac{\Sigma}{n}\right)^{-1}\bar{Y}\right).$$

Notice that the Harris ergodicity of the chain is also a straightforward extension of [3].

2 Simulation and Data Analysis

We now perform a test for different values of the prior distribution on a synthetic dataset. The aim is to assess the sensitivity with respect to the choice of the hyperparameters. We generated $n=100$ data points from a normal distribution with zero mean and identity matrix as covariance. We set $u=1$, $p=2$, $r=3$. The parameters are defined as follows:

1. $\mu=(12,12,12)$
2. $\omega=6.2$
3. $\omega_0=(3.2,1.4)$
4. $O=I_r$

and Y_i are randomly drawn from a multivariate normal with mean $\mu+\Gamma\eta X_i$ and covariance $\Gamma\Omega\Gamma^T+\Gamma_0\Omega_0\Gamma_0^T$. As for the hyperparamaters, we distinguish between three cases. We focus on μ as it is the most relevant change we make. In the first case, we use a weakly informative proper prior with $\mu_0=(0,0,0)$ and $\Sigma_0=\kappa I_r$, with $\kappa=10$. In the second test case, we set $\mu_0=\bar{Y}$ and $\Sigma=I_r$. Finally, we consider the improper prior as in [3]. The other parameters are set as follows: $C=I_p$, $D=I_r$, $G=I_r$, $\alpha=3$, $\psi=3$, $\alpha_0=3$ and $\psi_0=3$.

For each case study, we run a Gibbs sampler for 1000 iterations, with a burn in of 300. The initialisation for each chain was from the same random point in the parameter space. This was mainly for convenience, as we did not want to find a suitable

initial point (as it is customary for applied Bayesian analysis, *f.i.* by starting from the approximate mode of the posterior), given that we wanted to assess convergence starting from a somewhat arbitrary place.

Results for the three components of μ are reported, respectively in Tables 1, 2, and 3.

We see that even though the empirical and the noninformative priors lead to somewhat closer posterior estimates, the effect of placing a weakly informative prior also yields posterior higher density intervals that are in line with the other two classes of prior distributions. However, the true advantage of a proper prior is that it allows for extending the model to more complex settings. The fact that it yields similar results notwithstanding different hyperparameters is certainly encouraging, although, as always, some care should be put in their refinement.

Table 1 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with a weakly informative prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.67	12.005	13.359
μ_2	12.043	11.715	12.324
μ_3	12.097	11.882	12.313

Table 2 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with an empirical prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.832	12.1	13.49
μ_2	12.071	11.746	12.361
μ_3	12.117	11.903	12.307

Table 3 Posterior inference for $\mu = (\mu_1, \mu_2, \mu_3)$ with a noninformative prior: posterior higher density interval (HDI) are reported.

Parameter	Mean	3% HDI	97% HDI
μ_1	12.85	12.246	13.425
μ_2	12.07	11.74	12.364
μ_3	12.121	11.916	12.286

3 Conclusions

In this work, we have constructed a proper prior distribution for a Bayesian envelope model. We carried out an assessment of the prior sensitivity on a simple test case, obtaining that the choice of the hyperparameters for the parameter μ yields similar results in the three cases studied: a weakly informative, an empirical one and a non-informative prior. The importance of allowing more options stem from different lines of reasoning. First of all, it might be required when prior knowledge is to be incorporated into the model. Moreover, even though the actual impact of any of these prior distributions is somewhat limited whenever the likelihood is dominant, this might not necessarily be the case. In such scenarios, a weakly informative or empirical prior might be preferable. Another advantage of a proper prior is that it allows for an extension of the model to more complex scenarios. For instance, finite mixture models and nonparametric extensions require prior distributions to be proper. This is due to the latent cluster structure they posit on the data: the more the clusters, the more likely they are empty. An improper prior is then ill-suited in this context, as it is not an actual distribution and does not respect the probabilistic properties required by Bayesian inference.

References

1. R. D. Cook. *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 1st edition edition, 2018.
2. R. D. Cook, B. Li, and F. Chiaromonte. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statistica Sinica*, 20(3):927–960, 2010.
3. K. Khare, S. Pal, and Z. Su. A Bayesian approach for envelope models. *The Annals of Statistics*, 45(1):196–222, Feb. 2017.

Hilbert principal component regression for bimodal bounded responses

Regressione con componenti principali in Hilbert per risposte limitate bimodali

Enea G. Bongiorno, Agnese M. Di Brisco, Aldo Goia, and Sonia Migliorati

Abstract The flexible beta regression model is an effective approach to deal with bounded and bimodal responses. The aim of this work is to generalized this regression model to cope with a generic Hilbert covariate, either high-dimensional or functional. The dimensionality reduction procedure is based on principal components and the selection of the significant ones in the regression framework is carried out within a Bayesian rationale. The effectiveness of the proposal is illustrated both on simulated and real data.

Abstract *Il modello di regressione flexible beta è un approccio efficace per gestire risposte limitate e bimodali. Lo scopo di questo lavoro consiste nel generalizzare il suddetto modello di regressione per far fronte a una generica covariata Hilbert, sia ad alta dimensionalità che funzionale. La procedura di riduzione della dimensionalità è basata sulle componenti principali e la selezione di quelle significative nel quadro di regressione è effettuata seguendo una logica bayesiana. L'efficacia della proposta è illustrata sia su dati simulati che reali.*

Key words: flexible beta, bayesian estimation, bayesian variable selection

---

Enea G. Bongiorno
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: enea.bongiorno@uniupo.it

Agnese M. Di Brisco
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: agnese.dibrisco@uniupo.it

Aldo Goia
Università del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, via Perrone, 18, 28100, Novara, Italia e-mail: aldo.goia@uniupo.it

S. Migliorati
University of Milano-Bicocca, Department of Economics, Management and Statistics, Milan, Italy, e-mail: sonia.migliorati@unimib.it

1 Introduction

The issue of modeling a bounded response, further complicated by possibly latent mixture structure inducing a bimodal shape, is common in many applications such as in Kalivas (1997); Reiss et al. (2017); Di Brisco et al. (2022). Standard regression techniques are not feasible in this context, and among many reasons the most evident is the risk of predicting values outside the support. A promising approach, though not the only one possible, relies on modeling the response variable directly on its bounded support by making use of proper distributions. The additional bimodal nature of the response requires even more ad hoc modeling. In particular, here the flexible beta model is considered for which the response is a suitable mixture of two betas (Migliorati et al., 2018). The aim of the work is to describe a Hilbert flexible beta (HFB) regression model, that is when the covariate of interest belongs to a Hilbert space. In the light of the type of covariate, either high-dimentional or functional, a dimensionality reduction procedure based on the principal components (PCs) is considered and a selection of how many and which ones components is carried out with a Bayesian variable selection approach. The outline of this work is as follows. Section 2 illustrates the flexible beta regression model with Hilbert covariate and describes the Bayesian estimation method. Section 3 applies the presented technique to the spectometric example.

2 Hilbert flexible beta regression

Let Y be a real random variable in $(0,1)$, and X a random element, with either high-dimensional or functional nature, in $\mathscr{H}$, a Hilbert space equipped with an inner product $\langle\cdot,\cdot\rangle$. Assume that the conditional pdf of Y given $X=x$ is a flexible beta (FB) (Migliorati et al., 2018), that is

$$f_{FB}(y|\lambda_1,\lambda_2,\phi,p) = pf_B(y|\lambda_1,\phi)+(1-p)f_B(y|\lambda_2,\phi), \qquad 0<y<1, \quad (1)$$

where $f_B(y|\lambda_j,\phi)$ denotes the pdf of a beta random variable with a mean ($0<\lambda_j<1$) and precision ($\phi>0$) parameterization. It is worth noting that λ_1 and λ_2 represent the conditional means given x of the first and second mixture components with the constraint $0<\lambda_2<\lambda_1<1$ to ensure identifiability, $\phi>0$ is the precision parameter, and finally $0<p<1$ is the mixing proportion parameter,

Let $g(\cdot)$ be a monotone and twice differentiable link function and $\mu=\mathbb{E}[Y|x]=p\lambda_1+(1-p)\lambda_2$ be the overall conditional mean of the mixture defined in (1). Assume $g(\mu)=\alpha+\langle\beta,x\rangle$, where $\alpha\in\mathbb{R}$, $\beta\in\mathscr{H}$ is the regression coefficient, either functional or high-dimensional depending on the nature of the covariate. Note that the proposed model does not belong to the GLM family since the related distribution is not in the exponential family. Morever, to explicitly include μ and to obtain a variation independent parameter space as well, the FB distribution can be reparametrized by means of $0<\mu<1$, $0<w<1$, $0<p<1$, and $\phi>0$ where

$w = (\lambda_1 - \lambda_2)/\min\{\mu/p, (1-\mu)/(1-p)\}$ quantifies the distance between the two mixture components.

Let $\{(X_i, Y_i), i = 1, \ldots, n\}$, $n \geq 1$, be a sample of iid copies of (X, Y). To deal with the regression coefficient that belongs to either a high-dimensional or infinite dimensional space, a principal component regression (PCR) strategy is performed leading to

$$\tilde{g}(\mu_i) = \alpha + \sum_{k=1}^{K} b_k \chi_{ik}, \qquad i = 1, \ldots, n, \qquad K \geq 1 \tag{2}$$

where $(\psi_k)_{k=1}^{\infty}$ is the orthonormal basis of principal component analysis, $b_k = \langle \beta, \psi_k \rangle$ are unknown real parameters, and $\chi_{ik} = \langle X_i, \psi_k \rangle$ are real random variables called PCs. To select how many PCs and which ones better explain the response, one takes advantages of Bayesian techniques of variable selection (O'Hara and Sillanp, 2009).

The problem of computing the posterior distribution from which to sample has no analytical solution, and thus one can resort to MCMC techniques which require the definition of the likelihood function and the a priori distributions of the unknown parameters. The complete-data likelihood, computed according to a data augmentation strategy, is given by

$$L(\mathbf{y}, \mathbf{v}|\eta) = \prod_{i=1}^{n} [p f_B(y_i|\lambda_{1i}, \phi)]^{v_i} [(1-p) f_B(y_i|\lambda_{2i}, \phi)]^{1-v_i},$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ is the vector of responses, $\eta = (\alpha, \mathbf{b}, \phi, p, w) \in \mathbb{R}^{K+4}$ is the vector of unknown parameters, and $\mathbf{v}$ is a latent variable with $v_i \sim Bernoulli(p)$, for $i = 1, \ldots, n$.

For what concerns the a priori information, non-informative prior distributions are selected with the scope of inducing the least impact on the posterior distributions.

$$\begin{cases} b_j|I_j, \tau_{b_j} \sim I_j \mathcal{N}(0, \tau_{b_j}^{-1}) + (1-I_j)\delta(0) \\ I_j|\pi_j \sim Bernoulli(\pi_j) \\ \pi_j \sim Unif(0,1) \\ \tau_{b_j} \sim Gamma(\gamma, \gamma) \end{cases} \qquad \begin{cases} \alpha \sim \mathcal{N}(0, \sigma) \\ \phi \sim Gamma(\gamma, \gamma) \\ p \sim Unif(0,1) \\ w \sim Unif(0,1) \end{cases}$$

It is worth noting that the coefficients b_j are associated with spike-and-slab priors showing a mixture structure with Bernoulli weights of a diffuse normal (the slab) and a $\delta(0)$ indicating a discrete measure concentrated at zero (the spike). The hyperparameters γ and σ have been suitably selected to induce non-informativeness.

The joint prior distribution results from the product of the marginal prior distributions since the parametric space is variation-independent. To simulate from the posterior distribution, that is proportional to the joint prior multiplied by the likelihood function, a Gibbs sampling procedure is adopted. In particular, a metropolis-Hasting algorithm within Gibbs is performed for sampling from the full conditionals that have not closed form. The implementation of the estimation algorithm is done

through the software OpenBUGS and R. Numerical and graphical inspections of the simulated chains are performed to assess the stationarity, the absence of autocorrelation, and overall the convergence to the posterior distributions.

3 Numerical example: high-dimensional setting

A first illustration of the proposed method concerns a toy example with a high-dimensional covariate X_i, for $i = 1,\dots,n = 100$, generated from a $D = 200$ dimensional multivariate centered normal with a covariance matrix $\Sigma = \left[0.9^{|i-j|}\right]_{1\leq i,j\leq D}$. As a result, the first 30 PCs explain more than 80% of the total variability and are thus included in the regression model. The response variable is simulated from a flexible beta distribution with $\phi = 100$, $p = 0.8$, and $w = 0.7$, thus being characterized by both boundedness and bimodality. Its conditional mean is as in (2) with $\tilde{g}(\cdot) = logit(\cdot)$, $\alpha = 0.5$, $b_1 = 3$, $b_3 = -3$, $b_6 = 1.5$, and all remaining b_k equal to zero. Estimation is performed by adopting the Bayesian approach described in Section 2. The main results, based on 100 Monte Carlo replications, are reported in Table 1. In all Monte Carlo replications the coefficients b_k different from zero are identified as significant, meaning that the posterior probability of inclusion are (by far) greater than 0.5. The remaining coefficients are therefore shrunk to 0 in almost all replications. Overall, the HFB provides very precise estimates with low associated MSEs for both the regression coefficients and the additional parameters of the FB.

Table 1 Monte Carlo posterior means and MSEs

	$\alpha = 0.5$	$\beta_1 = 3$	$\beta_3 = -3$	$\beta_6 = 1.5$	$log(\phi) = 4.6$	$p = 0.8$	$w = 0.7$
Post. Mean	0.491	3.008	-2.973	1.522	4.603	0.796	0.698
MSE	0.01	0.098	0.098	0.082	0.026	0.002	0.001

4 Application: infinite dimensional setting

The proposed application is based on a spectroscopic dataset described in Kalivas (1997). The functional data refers to the observed spectra measuring the intensity of reflection over the near infra-red wavelength of the radiation and collected on 100 wheat samples. Instead, the bounded response variable is the moisture content measured in percentage. Left-hand panel of Figure 1 shows the estimated density curve of the moisture content, which is characterized by an evident bimodality in addition to the bounded support. Right-hand panel of Figure 1 shows the once-differenced spectra of the wheat samples collected on a regular grid of 701 points ranging be-

tween 1100nm and 2500 nm. The HFB regression model has been estimated either

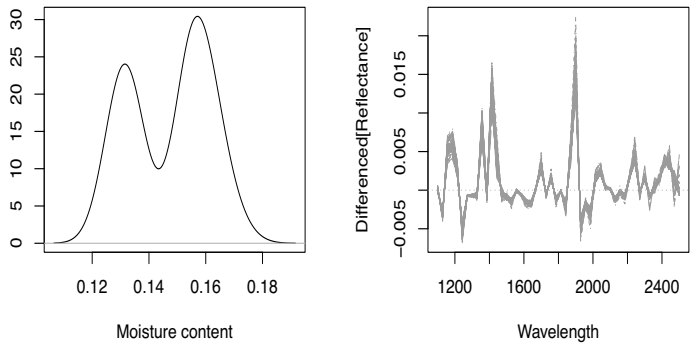


Fig. 1 Left-hand panel: estimated density curve of the moisture content in percentage. Right-hand panel: once-differenced wheat spectra.

by including the observed spectra or the once-differenced spectra as functional covariate with $g(\cdot) = logit(\cdot)$, according to Reiss et al. (2017). In what follows only results concerning the model with the once-differenced spectra are included, since the direct use of the observed spectra leads to consistent results. The posterior means and standard deviations (SD) of the coefficients are provided in Table 2. In particular, only the posterior estimates of the significant coefficient b_k are reported, i.e. those for which the posterior probability of inclusion of that coefficient is greater than 0.5. It is worth noting that only the first six (except for the fourth) components were significant, despite from the third component onwards the percentage of explained variability of the covariate is negligible. Moreover, it is of interest to observe that the model recognises two almost equally weighted groups (posterior mean of p around 0.57) whose component means are well-separated (posterior mean of w around 0.7). For comparative purpose, a beta-type regression model has been implemented too. The estimated functional parameter β is quite similar to that of the proposed model, but the overall fit to data ($WAIC = -915.509$) is worse than that provided by the HFB model ($WAIC$=-923.479).

Finally, a 5-fold cross-validation analysis was performed. The predictive accuracy of the model is measured by computing the RMSE and is compared to the one obtained from a standard generalized linear model (GLM) (Ramsay and Silverman, 2005) and a non parametric (NP) scalar-on-function regression model (Ferraty and Vieu, 2006) evaluated with respect to the $logit$ transformation of the response. As it emerges from Table 3, the proposed model is the best choice both in terms of overall fit on the whole sample and in terms of predictive accuracy in cross-validation.

Table 2 Posterior means and SD of the parameters

	α	β_1	β_2	β_3	β_5	β_6	$log(\phi)$	p	w
Post. Mean	-0.626	-0.038	-0.029	-0.069	-0.010	0.009	9.785	0.567	0.702
SD	0.687	0.023	0.018	0.042	0.008	0.008	0.334	0.238	0.202

Table 3 R^2 computed on the whole sample. RMSE based on 5-fold CV (results have been multiplied by 1000 to improve readability)

	HFB	GLM	NP
R^2	0.972	0.904	0.79
RMSE-CV	2.78	24.857	4.164

References

Di Brisco AM, Bongiorno E, Goia A, Migliorati S (2022) Bayesian Flexible Beta Regression Model with Functional Covariate. PrePrint

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer Series in Statistics, Springer, New York

Kalivas JH (1997) Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems 37(2):255–259

Migliorati S, Di Brisco AM, Ongaro A (2018) A New Regression Model for Bounded Responses. Bayesian Analysis 13(3):845 – 872

O'Hara RB, Sillanp MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian Analysis 4(1):85 – 117

Ramsay J, Silverman BW (2005) Functional Data Analysis. Springer Series in Statistics, Springer, New York

Reiss PT, Goldsmith J, Shang HL, Ogden RT (2017) Methods for scalar-on-function regression. International Statistical Review 85(2):228–249

Methods of causal inference

Bayesian causal mediation analysis through linear mixed-effect models

Analisi di mediazione causale Bayesiana mediante modelli lineari a effetti misti

Chiara Di Maria, Antonino Abbruzzo and Gianfranco Lovison

Abstract In mediational settings, the main focus is on the estimation of the indirect effect of an exposure on an outcome through a third variable called mediator. The traditional maximum likelihood estimation method presents several problems in the estimation of the standard error and the confidence interval of the indirect effect. In this paper, we propose a Bayesian approach to obtain the posterior distribution of the indirect effect through MCMC, in the context of mediational mixed models for longitudinal data. A simulation study shows that our method outperforms the traditional maximum likelihood approach in terms of bias and coverage rates.
Abstract *In contesti in cui è presente una mediazione, ci si concentra sulla stima dell'effetto indiretto di un'esposizione su una risposta tramite una terza variabile detta mediatore. Il tradizionale metodo di stima basato sulla massima verosimiglianza presenta diversi problemi nella stima dell'errore standard e dell'intervallo di confidenza dell'effetto indiretto. In questo articolo, proponiamo un approccio bayesiano per ottenere la distribuzione a posteriori dell'effetto indiretto tramite MCMC, applicato a modelli di mediazione a effetti misti per dati longitudinali. Uno studio di simulazione mostra che il nostro metodo è superiore alla stima con massima verosimiglianza in termini di distorsione e copertura.*

Key words: longitudinal mediation analysis, mixed effect models, Bayesian methods, causal inference

1 Introduction

In many fields it may be crucial to identify the causal mechanisms underlying phenomena of interest, such as the ones regulating the development of cancer or

Chiara Di Maria, Antonino Abbruzzo, Gianfranco Lovison
University of Palermo, Viale delle Scienze Building 13, Palermo (PA), Italy, e-mail: chiara.dimaria@unipa.it, antonino.abbruzzo@unipa.it, gianfranco.lovison@unipa.it

those related to the insurgence of psychological issues like depression or violent behaviours. Causal mediation analysis is a statistical technique widely used for this scope, since it allows researchers to understand how an exposure X exerts its effect on a response Y, if only directly, or also indirectly though intermediate variables called *mediators*. Longitudinal data are particularly suited for mediation analysis, since they ensure the correct time ordering of the variables (exposure has to precede the mediator, which in turn has to precede the outcome) and the necessary time span for the causal mediational effects to unfold [1, 6]. Both in single-time and longitudinal frameworks, if mediation models are linear, the indirect effect is generally obtained as a product of regression coefficients. Even if the estimators of such regression coefficients are assumed to be normally distributed, their product may not [2, 4]; hence, the estimation of the standard error and the confidence intervals of indirect effects is not straightforward using maximum likelihood (ML).

In this paper, we address the problem of estimating the indirect effect in a longitudinal setting, modelling data through linear mixed models (LMM) within a Bayesian framework, adopting a Markov Chain Monte Carlo (MCMC) approach. This allows us to draw values from the posterior distribution of the indirect effect, and derive point estimates and highest credibility intervals (HDI).

We will first introduce the modeling framework, then we move to the description of a simulation study comparing the performance of ML and the Bayesian approach in terms of standard errors, bias and coverage rates, the latter using bootstrap confidence intervals and HDIs, respectively. Some conclusions will follow.

2 The counterfactual mediation framework

Defining mediation in a causal framework requires specific notation and assumptions. Causality is indeed a concept difficult to address, especially in observational studies, where the randomisation of the exposure is not guaranteed and there is plenty of potential unobserved confounders.

One of the most widely accepted (although not free from controversies) framework to deal with causality is the counterfactual one [7, 8]. Imagine to observe a patient dying of stroke after taking a drug. Has it caused the stroke and led the patient to death or would he/she have died in any case because of other reasons? To answer this question one should observe the same patient not taking the medication and see if he/she dies anyway or not. This is clearly impossible, but it is the kind of reasoning at the very basis of the counterfactual framework. The definition of *causal effect* of X on Y relies on the comparison of the two counterfactual values $Y(x)$ and $Y(x')$ which the outcome would assume if X were set to x and x', respectively. In the very common case of binary exposure, the *total causal effect* is defined as the expected difference $\mathbb{E}[Y(1) - Y(0)]$.

In the mediational framework the total effect of the exposure on the response can be decomposed into two effects, the *direct effect*, i.e. the effect not passing through the mediator, and the *indirect effect*, i.e. the effect conveyed through the

mediator. There exist several definitions of mediational effects, and the most popular is that of natural direct and indirect effects [9]. The former is defined as $NDE = \mathbb{E}[Y(x, M(x)) - Y(x', M(x))]$, that is, the expected difference between the values of the response if X were set to x and x', respectively, keeping the mediator fixed at the natural value it would assume if X were x. The latter is defined as $NIE = \mathbb{E}[Y(x, M(x)) - Y(x, M(x'))]$, that is, the expected difference in the outcome if X were fixed at x and the mediator at the value it would naturally assume if X were set to x and x', respectively (see [9] for an exhaustive introduction on the topic).

Making inference on the direct and indirect effects is quite hard, especially in complex settings like the longitudinal one, since, in order to make mediational effects identifiable from observed data, strict assumptions on unmeasured confounders have to be made. In particular, it is assumed the absence of unmeasured exposure-mediator confounders, exposure-outcome confounders and mediator-outcome confounders (in the last case, either affected and unaffected by the exposure). See [9] for a discussion on these assumptions in the general setting and [3] for those in the longitudinal one.

3 Model specification and Bayesian approach to longitudinal mediation

In this paper, we focus on a longitudinal mediation setting, where the exposure, the mediator and the outcome are measured repeatedly over a certain time interval. One of the modeling framework proposed to address such a setting is that of mixed-effect models [3].

Let X_{it}, M_{it} and Y_{it} denote the exposure, the mediator and the outcome, respectively, for the i-th subject at the t-th measurement occasion, for $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. Although the framework proposed by [3] is that of generalised mixed models, here we focus only on the case of normal mediator and outcome with identity link functions, i.e. on LMMs. The expected values of the mediator and the outcome can then be written as

$$\mathbb{E}[M_{it}|X_{it} = x, \mathbf{b}_i] = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x \tag{1}$$

$$\mathbb{E}[Y_{it}|X_{it} = x, M_{it} = m, \mathbf{g}_i] = (\gamma_0 + g_{0i}) + (\gamma_1 + g_{1i})x + (\gamma_2 + g_{2i})m \tag{2}$$

where $\mathbf{b}_i = (b_{0i}, b_{1i})^T$ and $\mathbf{g}_i = (g_{0i}, g_{1i}, g_{2i})^T$ are vectors of subject-specific random effects. It is assumed that $(\mathbf{b}_i, \mathbf{g}_i)^T$ follows a multivariate normal distribution with null mean vector and variance-covariance matrix allowed to be non-diagonal.

[3] prove that, under the aforementioned traditional assumptions concerning the absence of unobserved confounders, and assuming the absence of time-varying confounders, the natural mediational effects, defined as

$$NDE = \mathbb{E}[Y_{it}(x, M_{it}(x)) - Y_{it}(x', M_{it}(x))|\,\mathbf{b}_i, \mathbf{g}_i] \tag{3}$$
$$NIE = \mathbb{E}[Y_{it}(x, M_{it}(x)) - Y_{it}(x, M_{it}(x'))|\,\mathbf{b}_i, \mathbf{g}_i], \tag{4}$$

are identifiable from observed data. In particular, they derive closed forms of these effects for some combinations of distributions of the mediator and the outcome, and link functions of the corresponding models. For example, in the normal-normal/identity-identity link case the mediational effects, marginalising over the random effects are:

$$NDE = \gamma_1(x - x') \tag{5}$$
$$NIE = (\beta_1\gamma_2 + \sigma_{b_1 g_2})(x - x'), \tag{6}$$

where $\beta_1, \gamma_1, \gamma_2$ are as in Eq. (1)-(2) and $\sigma_{b_1 g_2}$ is the covariance between b_1 and g_2.

It can be noticed that the expression for the NIE includes a product of coefficients. As already mentioned, this makes inference on the indirect effect not straightforward, since, even assuming that the two regression coefficients estimators are normally distributed random variables, generally their product is not. In fact, the distribution of the product may be highly skewed and, most importantly, unknown and this complicates the estimation of confidence intervals for the mediated effect [5]. [3] use non parametric bootstrap to obtain standard errors and confidence intervals, but it is well know that this technique is computationally intensive.

Bayesian MCMC methods allow to obtain the posterior distribution of the indirect effect and a highest density credibility interval, which, broadly speaking, is more interpretable and accurate than its frequentist version, since it takes into account the asymmetry of the distribution and the information deriving from the data.

4 Simulation study

We carried out a simulation study in order to compare the performance of the Bayesian approach to the traditional ML estimation method. The simulations were conducted making use of R and JAGS, by considering six scenarios obtained by varying the number of subjects $n = 100, 1000$ and the number of measurement occasions $T_i = T = 5, 7, 10$. Abiding by the setting described by [3], the exposure at time 1 was randomly drawn from a Bernoulli and its subsequent measurements were obtained by fitting an autoregressive model so that $\mathbb{E}[X_{it}] = \text{logit}^{-1}(\alpha_0 + \alpha_1 X_{it-1})$, where α_0 and α_1 were drawn independently from a Uniform(-2,2).

The mediator and the outcome were generated from a normal distribution with expectations as in Eq. (1) and (2), respectively, and standard deviations are set to 1. As regards the random factors, we drew $(\mathbf{b}_i, \mathbf{g}_i)^T \sim MVN(\mathbf{0}, \Theta)$, where Θ is drawn from a Wishart with five degrees of freedom and scale matrix given by a Toeplitz matrix. Fixed coefficients do not vary over iterations, while $\sigma_{b_1 g_2}$ changes at each, as well as the true value of the indirect effect.

Table 1 Results of the simulation study showing the bias of indirect effect, the percent relative bias, standard errors and the coverage rate of bootstrap confidence intervals and HDI.

		ML				Bayes			
	T	S.E.	Cov. Rate	Bias	% Bias	S.E.	Cov. Rate	Bias	% Bias
n = 100	5	0.289	75%	-0.13	18.75	0.419	95.5%	0.05	1.44
	7	0.289	75%	-0.11	18.36	0.410	98%	0.051	0.048
	10	0.283	76.5%	-0.13	13.90	0.394	96%	0.006	0.678
n = 1000	5	0.090	19%	-0.28	27.71	0.145	94%	-0.013	0.284
	7	0.090	36.5%	-0.17	17.7	0.133	94%	-0.004	0.009
	10	0.089	44%	-0.11	14.09	0.126	94.5%	0.001	1.11

In the Bayesian estimation, we used non-informative priors. We modeled the mediator and the outcome as $M_{it} \sim N(\mu_{it}, \tau_m)$, $Y_{it} \sim N(\nu_{it}, \tau_y)$, where μ_{it} and ν_{it} are as in Eq. (1) and (2), respectively, while τ_m and τ_y are precision parameters[1] following a Gamma$(0.01, 0.01)$. All regression coefficients were assumed to be drawn from a normal distribution with null mean and precision 0.01. Finally, the random effects were modeled as draws from a multivariate normal with zero mean and precision matrix from an Inverse Wishart. The Bayesian estimation procedure relied on three chains of length 10,000 and adaptation 5,000.

We compared the performance of the ML estimation method, combined with non-parametric bootstrap for estimating the standard error and the confidence interval of the indirect effect, and the Bayesian approach using 95% highest density credibility intervals. We ran 200 iterations, and for each one we computed the bias, the relative bias and the standard error of the estimated indirect effect and the coverage rates of its bootstrap-based confidence intervals and HDIs. In the Bayesian case, the standard error was obtained by averaging standard errors of the NIE mean estimated in the 200 iterations.

Table 1 shows the results of the simulation study. It is evident that the Bayesian approach performs better than ML in terms of bias and coverage rate, while standard errors are smaller in the ML approach. It is also worth remarking that the Bayesian approach is advantageous even from a computational perspective, since estimates are obtained much faster than those obtained from bootstrap. Figure 1 shows twenty randomly selected indirect effects and corresponding ML/Bayesian estimates in the scenario with $n = 100$ and $T = 10$. It is easy to see that Bayesian posterior means are generally closer to the true effect and HDIs generally include it, although they are often wider than bootstrap CI. These differences attenuate in the scenario with 1000 observations.

5 Conclusions

This work addresses the issue of making inference on the indirect effect in a longitudinal mediational setting through the Bayesian approach, whose main strength is the

[1] JAGS requires precision instead of variance.

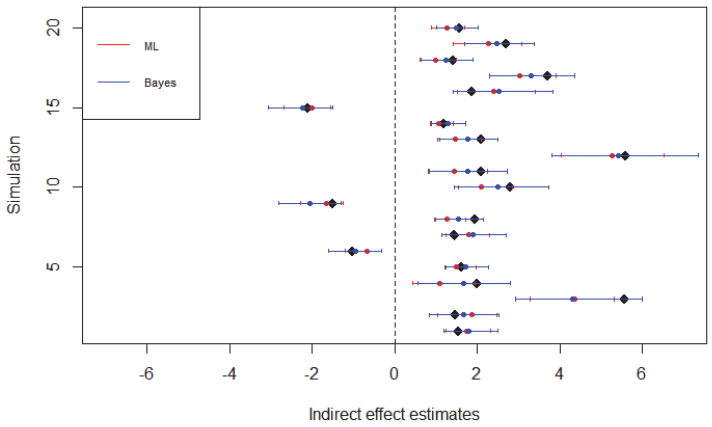


Fig. 1 Simulation results in the scenario $n = 100, T = 10$ for 20 randomly selected iterations. Black diamonds is the true indirect effect, blue and red dots/lines correspond to posterior means/HDI and ML point estimates/bootstrap confidence intervals, respectively.

possibility of obtaining the whole distribution for the indirect effect and, as a consequence, more reliable credibility intervals. The simulation study shows excellent results in the case of normal mediator and outcome modeled through linear mixed models. Further research is required to evaluate the performance of this approach when the mediator and/or the outcome are not normally distributed and the link functions differ from the identity, and in the presence of covariates, either baseline and time-varying.

References

1. Aalen, O. O. and Frigessi, A.: What can Statistics Contribute to a Causal Understanding? Scand. J. Stat. **34**(1), 155—168 (2007)
2. Aroian, L.A.: The probability function of a product of two normal distributed variables. Ann. Math. Stat. **18**, 256–271 (1947)
3. Bind, M.-A. C., VanderWeele, T. J., Coull, B. A., Schwartz, J. D.: Causal mediation analysis for longitudinal data with exogenous exposure. Biostat. **17**(1), 122–134 (2016)
4. Craig, C. C.: On the frequency function of xy. Ann. Math. Soc. **7**, 1–15 (1936)
5. MacKinnon, D. P., Lockwood, C. M., Williams, J.: Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivar. Behav. Res. **39**(1) (2004)
6. Maxwell, S. E., Cole, D. A.: Bias in Cross-Sectional Analyses of Longitudinal Mediation. Psychol. Methods **12**(1), 23–44 (2007)
7. Pearl, J.: Causality, Cambridge: Cambridge University Press (2009)
8. Rubin, D. B.: Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. J. Am. Stat. Assoc. **100**(469), 322–331 (2005)
9. Vanderweele, T. J.: Explanation in Causal Inference. New York: Oxford University Press (2015)

Bootstrap-aggregated Adjustment Set Selection

Selezione di un Adjustment Set tramite Bootstrap-aggregating

Lorenzo Giammei

Abstract Causal effects can be estimated from observational data within the Causal Graph framework. When the true causal graph is unknown, data can be used to learn the graph and then select a sufficient set of covariates for adjustment according to its structure. Graph learning is a crucial step of the process since misspecifications of the graph can result in biased causal estimates. We propose a procedure that resorts to bootstrap-aggregating to select the adjustment set based on its stability across bootstrap resamples. The performance of the novel method is tested on graphs of different complexity and compared to three alternative procedures. The results show that the relative accuracy of the proposed methodology improves as the complexity of the graph increases.

Abstract *Gli effetti causali possono essere stimati tramite dati non sperimentali nel framework dei Grafi Causali. Quando il grafo causale non è noto, i dati posso essere impiegati per apprendere il grafo e selezionare un adjustment set di variabili che consenta di ottenere stimatori non distorti degli effetti causali. Proponiamo una procedura che seleziona un adjustment set sulla base della sua stabilità nelle repliche bootstrap. La performance del nuovo metodo è testata su grafi di diversa complessità e confrontata con tre procedure alternative. I risultati mostrano che l'accuratezza relativa della procedura proposta migliora all'aumentare della complessità del grafo.*

Key words: Causal effect estimation, Causal bayesian networks, Backdoor criterion, Bootstrap.

Lorenzo Giammei
Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, Italy
e-mail: lorenzo.giammei@uniroma1.it

1 Introduction

Causal inference investigates causal relations between variables. One of the possible frameworks to investigate causal claims, when dealing with observational data, are Causal Graphs [7], also called Causal Diagrams or Causal Bayesian Networks. Causal diagrams represent variable interactions clearly through a graph, and, if identifiable, estimate the treatment's causal effect on an outcome variable.

When treatment assignment is not randomized, the main concern is confounding: the situation where there is at least one variable that has a direct causal effect on both the treatment and the outcome. This kind of configuration can generate biased estimates unless adequately accounted for. However, if the causal relations between variables are known a priori, it is possible to select a sufficient set of covariates for confounding adjustment, if such a set exists. In the Causal Graphs framework, sufficient adjustment sets can be derived from the structure of the graph, following the back-door criterion [7]. Once the set is selected, adjustment can be performed through various methods, including matching, weighting, regression adjustment or doubly robust methods.

Unfortunately, in real settings the true graph is often unknown or just partially known. In this case the graph can be learnt from data through causal discovery methods [7]. Once the graph is learnt, it is usually taken as a good approximation of the true graph and then a sufficient adjustment set can be selected to estimate a chosen treatment effect. However, if the structure of the graph is misspecified, the selected adjustment set could be not sufficient for confounding adjustment. Here a procedure to select the adjustment set from multiple graphs, employing bootstrap is proposed. This procedure is tested on three graphs of increasing complexity and then compared to other adjustment set selection techniques.

The paper is organized as follows. Section 2 introduces some basic causal graph terminology and notation. The proposed method is described in detail in Section 3, while Section 4 contains simulations and results. The findings are discussed in Section 5.

2 Causal Graphs

A graph $G = (V, E)$ is a collection of nodes or vertices V and edges E. A *causal graph* is a graph where nodes represent random variables and edges describe the causal relations between these nodes. If two nodes X_i and X_j are connected by an edge the nodes are *adjacent*. A *path* between two nodes X_i and X_j is a sequence of nodes beginning with X_i and ending with X_j where all the nodes are connected to the next. Edges can be *directed* or *undirected*. An edge is directed if it goes out from one node into another and undirected without such orientation. A graph where all the edges are directed is a *directed graph*. If a directed edge goes from X_i to X_j then X_i is a *parent* node of X_j, and X_j is a *child* of X_i. We will denote parents of node X_j with the notation pa_{X_j}. A *directed path* between two nodes is a path where no node

has two edges on the path directed into it, or two edges directed out of it. Given a directed path, the first node is an *ancestor* of every node of the path, and every node of the path is a *descendant* of the first node. A directed path that begins and ends with the same node is a *cycle*. A directed graph containing no cycles is called a *directed acyclic graph* (DAG). If the underlying dependence model of a problem is unknown, the graph can be estimated from a dataset containing the variables of interest through a *structural learning algorithm*.

The transition from an observational context to an interventional one requires some further discussion. Pearl in [7] introduces the notation $do(X_i = x_i)$ to indicate that the variable X_i is set to the value x_i by intervention. Hence $P(X_j|do(X_i = x_i))$ denotes the distribution of X_j given that X_i is forced to take value x_i. In an observational context the quantity $P(X_j|do(X_i = x_i))$ is not measured but it can be derived employing the back-door criterion [7]. Given a graph G with set of nodes $\mathbf{X}$, a treatment variable $T \in \mathbf{X}$ and an outcome variable $Y \in \mathbf{X}$, the back-door criterion is a graphical test that allows the selection of a set of covariates $\mathbf{S} \subseteq \mathbf{X}$, ensuring unbiased causal estimates for the effect of T on Y. A set that satisfies the back-door criterion is called a *sufficient* adjustment set. If a sufficient adjustment set exists, then interventional distribution can be expressed in observational terms:

$$P(X_j|do(X_i = x_i)) = \sum_{\boldsymbol{S}} P(X_j|\mathbf{S} = \mathbf{s}; X_i = x_i)P(\mathbf{S} = \mathbf{s}) \tag{1}$$

3 Bootstrap-aggregated Adjustment Set

Bootstrap [8] is a resampling technique that creates replicates of the original dataset by sampling from it with replacement. The procedure is repeated M times to produce M bootstrap replicates of the same size as the original dataset. This work builds on a particular implementation of bootstrap called bootstrap-aggregating or bagging [3]. Bagging is a machine learning method that generates multiple versions of a predictor through bootstrap and uses these to get an aggregated predictor. A plurality vote is performed if the predicted element is a class. Here, the logic of bagging is applied to graph structural learning and adjustment set selection. The pseudocode of the procedure is contained in Algorithm 1.

Consider a dataset $D = (\mathbf{X}; T; Y)$ of size N sampled from an unknown graph G containing a set of covariates $\mathbf{X}$, an outcome Y and a treatment T. In the first step, D is used to produce M bootstrap replicates of the same size. In step 2, the multiset Θ is initialized as an empty set, which will be populated in the following iterations. A graph $\hat{G}_i$ is learnt with an algorithm $\mathscr{L}$ on each generated bootstrap sample B_i, given the constraint that the treatment T is a parent of the outcome Y. This assumption encodes in the graph the a priori knowledge that T directly causes Y. Then an adjustment set selection method $\mathscr{A}$ for estimating the effect of T on Y is applied to each graph $\hat{G}_i$. The selected sets are then added to Θ until steps 4 and 5 have been iterated over all the bootstrap resamples. The last step of the algorithm

Algorithm 1: Bootstrap-aggregated adjustment set

Input: A sample $D = (\mathbf{X}, T, Y)$ from an unknown graphical model G, a number of bootstrap samples M, a structural learning algorithm $\mathscr{L}$, an adjustment set selection procedure $\mathscr{A}$.

Output: A bootstrapped adjustment set $\theta^\star$

Generate M bootstrap samples $B_1, ..., B_M$ from D;
$\Theta = \emptyset$;
for $i = 1$ *to* M **do**
    $\hat{G}_i = \mathscr{L}(B_i | T \in pa_Y)$;
    add $\mathscr{A}(\hat{G}_i, T, Y)$ to Θ;
end
$\theta^\star = \max_{\theta_j \in \Theta} \{v(\theta_j)\}$

defines the bagged adjustment set $\theta^\star$ as the set which maximizes $v(\theta_j)$ among all sets $\theta_j \in \Theta$, where $v(\cdot)$ denotes the multiplicity of a given element of a multiset. The bagged adjustment set $\theta^\star$ is thus the set, among all $\theta \in \Theta$, which satisfies the back-door criterion in the highest number of learnt graphs.

4 Simulation

In this section, we assess the performance of the proposed method by comparing it to three alternative approaches. The first benchmark procedure consists in a straightforward implementation of structural learning and adjustment set selection; the second resorts to bootstrap to build an average graph[4] and then selects an adjustment set according to its structure; the third employs bootstrap and selects the adjustment set on the graph with the best score across the replicated samples. The main difference between the benchmark methods and the proposed technique is that the latter aims at selecting a stable adjustment set, whereas the other methods focus on retrieving the best graph and in a separate phase select a sufficient adjustment set. The techniques will be tested on graphs of increasing complexity, and the accuracy of the obtained results will be compared.

Given a known graph $G(\mathbf{X}, T, Y)$, where $T \in pa_Y$, we sample K datasets $D_1, ..., D_K$ of size N from it. Then we apply the methods on each sample D to obtain K adjustment sets $\theta_i^\star$ for the estimation of the causal effect of T on Y. Finally, we check if $\theta^\star$ is a sufficient adjustment set for calculating the effect of T on Y in the true graph G. The results are summarized by the quantity

$$R = \frac{\sum_{i=1}^{K} I_G(\theta_i^\star)}{K}, \tag{2}$$

where $I_G(\theta_i^\star)$ is an indicator function taking value 1 if $\theta_i^\star$ is a sufficient adjustment set for (T, Y) in G and 0 otherwise.

R is calculated for each pair (T,Y), such that $T \in pa_Y$ according to the structure of G. The total number of pairs (T,Y) is equal to $\sum_{v \in \mathscr{V}} \#pa_v$ where $\mathscr{V}$ is the ensemble of vertices of graph G and $\#pa_v$ is the number of parents of vertex $v \in \mathscr{V}$. Once R has been calculated for every possible pair (T,Y), an average of the results is computed to obtain a summary of the accuracy measure for the whole graph. If we denote $M = \sum_{v \in \mathscr{V}} \#pa_v$, we can write the average of the accuracy measures for a given graph as

$$\overline{R} = \frac{\sum_{m \in M} R_m}{M}. \tag{3}$$

Equation 3 thus gives an account of the performance of an adjustment set selection procedure over the set of nodes of a causal graph.

The simulations are performed on three discrete networks of increasing dimension and complexity that have been frequently used in the literature: Asia [6], Alarm [1] and Insurance [2]. We set the number of samples $K = 10$, and the number of bootstrap replicates $B = 200$. The same number of replicates is used in the bootstrap-aggregated procedure and in the benchmark methods that implement bootstrap. Different sample sizes are used in the simulations, according to the complexity of the network. The structural learning algorithm $\mathscr{L}$ used to learn the graphs $\hat{G}$ is the *Tabu Search* algorithm [5] with a BIC score. Tabu Search belongs to the family of score-based algorithms, and it has been chosen because it is more accurate and faster than most other learning algorithms, for both small and large sample sizes. The chosen adjustment set selection procedure $\mathscr{A}$ is the *minimal adjustment set*. The method selects the adjustment set with the smallest cardinality between all possible adjustment sets. The results are summarized in Figure 1, where the bootstrap-aggregated procedure is represented by a dotted line. The x-axis represents sample size n, while the y-axis describes the average accuracy measure $\overline{R}$.

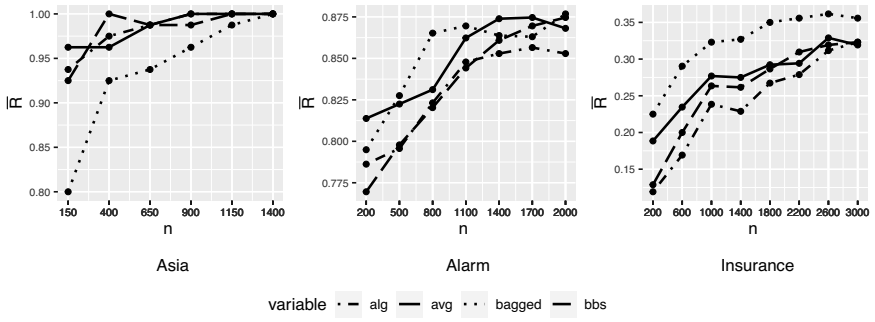


Fig. 1: Accuracy of the methods resulting from simulations (alg: adjustment set selection on a graph learnt through a single algorithm; avg: adjustment set selection on an average graph; bagged: bootstrap-aggregated adjustment set selection; bbs: adjustment set selection on the best-scoring graph among bootstrap replicates).

5 Discussion

This work proposed a novel procedure to select an adjustment set for causal effect estimation when the true causal graph is unknown. The paper contributes to the literature by proposing a procedure that directly aims at selecting a sufficient adjustment set with the use of bootstrap. Bootstrapping has already been used in the graph learning phase to measure confidence toward the presence of an edge between two nodes. However, this implementation is still oriented on recovering the most reliable graph structure and then selecting the adjustment set in a different step. Instead, the proposed procedure uses bootstrap to obtain a multiset of adjustment sets from multiple estimated graphs. This way of proceeding detaches the adjustment set selection from a single learnt graph's structure, thus aiming to achieve higher accuracy.

The technique is tested on different networks, and its results are compared to those obtained with three alternative methods. All the chosen procedures show similar levels of accuracy, which generally decrease as the complexity of the considered structure increases. On the simplest graph, at small sample sizes, the bagged adjustment has a lower accuracy than other benchmark methods. However, as the sample size increases, the performance of bagging improves. When considering more complex graphs, the relative performance of the novel procedure stands out. In particular, the results on the most complex diagram show that even if all the methods achieve low accuracy levels, bagging produces the most accurate results at both low and high sample sizes. Note that these findings remain tied to the tested causal graphs and the assumptions made in the simulations. Further analysis is required to assess how the procedure behaves with different graph configurations and assumptions.

References

1. Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F.: The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In: AIME 89, pp. 247–256. Springer (1989)
2. Binder, J., Koller, D., Russell, S., Kanazawa, K.: Adaptive probabilistic networks with hidden variables. Machine Learning **29**(2), 213–244 (1997)
3. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
4. Friedman, N., Goldszmidt, M., Wyner, A.J.: On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks. In: AISTATS (1999)
5. Glover, F.: Future paths for integer programming and links to artificial intelligence. Computers & operations research **13**(5), 533–549 (1986)
6. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society: Series B (Methodological) **50**(2), 157–194 (1988)
7. Pearl, J.: Causal inference in statistics: An overview. Statistics Surveys **3** (2009). DOI 10.1214/09-SS057
8. Tibshirani, R.J., Efron, B.: An introduction to the bootstrap. Monographs on statistics and applied probability **57**, 1–436 (1993)

Exploiting partial knowledge to evaluate the average causal effect via an ABC perspective

L'introduzione di conoscenze parziali per valutare l'effetto medio causale tramite una prospettiva ABC

Giulia Cereda, Fabio Corradi, Cecilia Viscardi

Abstract To evaluate the average causal effect, we propose a method to be used when the variables required to remove confounding are not observed. We consider the case in which the causal structure is known, and there is also some additional partial knowledge. The evaluation is probabilistic, and the partial knowledge may consist of an ordering of some relevant conditional probabilities and/or of the marginal distributions of some variables. A simple way to integrate such additional information is to adopt an approximate Bayesian computation perspective to derive an approximate posterior distribution for the average causal effect. We experimentally evaluate our methodology through an example from the literature, and we compare the results with the exact evaluation of the average causal effect.
Abstract *Al fine di valutare l'effetto causale medio, proponiamo un metodo che possa essere usato quando le variabili richieste per eliminare il confondimento non sono osservabili. Consideriamo il caso in cui la struttura causale del modello è nota e sono disponibili alcune informazioni aggiuntive. La valutazione fornita è probabilistica e la conoscenza aggiuntiva può consistere in un ordinamento di alcune rilevanti probabilità condizionate e/o nella conoscenza di distribuzioni marginali. Un modo semplice per incorporare tali informazioni è quello di adottare una prospettiva di calcolo Bayesiano approssimato per derivare una distribuzione a posteriori approssimata dell'effetto causale medio. La metodologia proposta è stata valutata su un esempio noto in letteratura e i risultati sono stati confrontati con una valutazione esatta.*

Key words: ABC, ACE, Adjustment formula, Causal graph, Simpson's Paradox.

1 Introduction and preliminaries

In causal inference, the evaluation of the Average Causal Effect (ACE) [5], measuring the difference between the expected values of the outcome under different scenarios, has significant relevance. A starting point to determine ACE is to assume the knowledge of a causal mechanism represented by a graph, more specifically a causal direct acyclic graph

Cecilia Viscardi
Dipartimento di Statistica, Informatica, Applicazioni e-mail: cecilia.viscardi@unifi.it

Fabio Corradi
Dipartimento di Statistica, Informatica, Applicazioni e-mail: fabio.corradi@unifi.it

Giulia Cereda
Dipartimento di Statistica, Informatica, Applicazioni e-mail: giulia.cereda@unifi.it

(DAG), defined by two sets $\{V,E\}$. V is the set of nodes representing the model variables, and E is the set of direct edges connecting nodes. Two connected nodes are linked by a direct causal relation, according to the orientation of the edge.

In a simple setting, two nodes of the causal graph are mainly of interest: X, the cause (often a treatment), and Y, the outcome. ACE is evaluated by measuring the difference in the expectation of Y for different interventions on X. An intervention on X, expressed by the so-called $do(\cdot)$ operator, consists of assigning the same x values to quantify changes in the system. This action is purely fictitious (observed x values correspond to past events that cannot be modified), but it has been proved that $do(X = x)$ corresponds to conditioning to $X = x$ in the model obtained as a surgery modified graph removing all the edges incident on X [4].

In the case of boolean Y and dichotomous X, the $do(\cdot)$ operator can be used to calculate the (adjusted) ACE according to:

$$ACE(X \to Y) = p(Y|do(X = x)) - p(Y|do(X = x')) \tag{1}$$

The discrepancy between the estimation of the causal effect obtained by (1) and

$$p(Y|X = x) - p(Y|X = x') \tag{2}$$

is due to confounding. Its detection may be obtained by using some topological properties of the causal graph, such as the well-known Back-door or Front-door criteria [4].

In a simple causal model with only one identified but unobserved confounder, we propose to incorporate some additional knowledge for the probabilistic evaluation of ACE via an algorithm based on the ABC type of reasoning. We experimentally evaluate our methodology through a running example, and we compare our results with a situation where the confounder is observed and the intervention can be evaluated correctly.

2 The running example

As a simple running example, we refer to a well-known causal mechanism with treatment and an outcome both influenced by a third variable, as represented in the causal graph in Fig. 2 (left). This graph represents the causal mechanism behind Simpson's paradox which occurs in many applications.

Fig. 1 (Left) Direct acyclic graph representing the causal mechanism assumed in the running example. (Right) Surgery modified causal graph.

To be concrete, we refer to an example provided by a real-life medical study concerning kidney stone treatments [3]. Let $Y \in \{0,1\}$ represent the no-recovery/recovery of a patient; $X \in \{A,B\}$ stand for two possible treatments; $Z \in \{S,L\}$ represent the size (Small or Large) of the stones of the treated individuals. The study gathers data from 700 patients with kidney stones undergoing the two treatments. Data are summarised in Table 1.

The causal story behind the graph tells us that both treatment and recovery are influenced by the size of the kidney stones of the exposed individual. Actually, doctors prefer

			Treatment (X)	
		Outcome (Y)	$X = A$	$X = B$
Size (Z)	Small	$Y = 1$	81	234
		$Y = 0$	6	36
	Large	$Y = 1$	192	55
		$Y = 0$	71	25
	Total	$Y = 1$	273	289
		$Y = 0$	77	61

Table 1 Complete data for the kidney stones example.

to assign the most effective and expensive treatment (A) to patients with the most severe kidney large stones pathology which experienced a lower rate of recovery. The opposite happens for patients affected by small kidney stones, which are more likely to receive treatment B and are more prone to recover.

If the leftmost graph in Fig. 2 correctly represents the causal mechanism, Z acts as a confounder, and the average causal effect can be evaluated as the difference in the probability of recovery under the two treatment regimes evaluated by using propagation of evidence in the surgery modified causal graph in Fig. 2 (right) or, directly and equivalently, by the so-called Adjustment formula [2]:

$$p(Y = 1|do(X = i)) = \sum_{j\in\{S,L\}} p(Y = 1|X = i, Z = j)p(Z = j) \qquad i \in \{A,B\}. \tag{3}$$

This can be used to calculate ACE according to (1):

$$ACE = \sum_{j\in\{S,L\}} \big(p(Y = 1|X = A, Z = j) - p(Y = 1|X = B, Z = j)\big)p(Z = j). \tag{4}$$

To estimate the components of (3) we need to observe the triplet (x, y, z) over a set of n observations; for the case at hand, the availability of the triplets, including observations on Z, is also the condition established by the Back-door criterion to remove confounding.

3 Dealing with unobserved confounders

Consider the causal graph in Fig. 2 again and assume that we are convinced of the validity of the depicted causal relations, but observations on Z are not accessible to us so that only the totals of Table 1 are available. The empirical joint distribution of X and Y leads to the conclusion that recovery is more likely undergoing treatment B (0.82) than undergoing treatment A (0.78), and the difference in predictions (2) amounts to -0.04.

Actually, the correct evaluation of ACE according to (4) amounts at 0.05, showing that treatment A is better than treatment B. This means that evaluating the causal effect relying on the probability of Y conditionally to X would be misleading.

A way to mitigate the effect of our lack of knowledge on Z is to exploit some additional pieces of information we are willing to assume. The main issue is finding a way to incorporate the available information in estimating the adjustment formula. In our running example we considered a set of three available pieces of information, $\mathscr{I} = \{\mathscr{I}_1, \mathscr{I}_2, \mathscr{I}_3\}$, defined as follows:

- $\mathscr{I}_1$ is the marginal distribution of Z, corresponding to assume that the distribution of small and large kidney stones among people affected by the disease is known in the population of interest;
- $\mathscr{I}_2$ is an ordering of $p(X|Z)$, in particular $p(X=A|Z=L) > p(X=B|Z=L)$ and $p(X=B|Z=S) > p(X=A|Z=S)$, corresponding to assume that Treatment A is more likely assigned to people with large kidney stones, while treatment B is assigned preferentially to people with small kidney stones;
- $\mathscr{I}_3$ is an ordering of $p(Y|X,Z)$, in particular

$$p(Y=1|Z=S,X=A) > p(Y=1|Z=S,X=B) > p(Y=1|Z=L,X=A) > p(Y=1|Z=L,X=B), \tag{5}$$

corresponding to assume that Treatment A is better than treatment B in both small and large kidney stones groups, and that the recovery with small stones is more probable than the recovery with large stones. Notice that this assumption implies a positive ACE.

In principle, a possible solution to estimate (4) is to build a complete database which require to impute Z values by sampling from the predictive

$$p(Z|X,Y) \propto p(Z)p(X,Y|Z) \tag{6}$$

which is based on the unavailable $p(X,Y|Z)$ in (3).

To deal with this issue and to incorporate $\mathscr{I}$ in the estimation, we adopt the same type of reasoning as ABC methods [6]. In particular, we consider the following distribution defined on an augmented space where X' and Y' are two auxiliary random variables:

$$p(Z,X',Y'|X,Y) \propto p(Z)p(X',Y'|Z)p(X,Y|X',Y',Z). \tag{7}$$

Generally speaking, samples from the above distribution can be obtained by implementing an ABC sampling scheme: 1) sample Z from its marginal $p(Z)$; 2) sample pairs (X',Y') from $p(X',Y'|Z)$; 3) approximate $p(X,Y|X',Y',Z) \approx \mathbb{1}\{(X,Y)=(X',Y')\}$. In practice, a two-fold approximation is usually introduced relying on $p(X,Y|X',Y',Z) \approx \mathbb{1}\{d\big(s(X,Y),s(X',Y')\big) \leq \varepsilon\}$, where $d(\cdot,\cdot)$ is a proper distance function, $s(\cdot)$ are summary statistics reducing the size of the data, and ε is a positive tolerance threshold.

Note that in the case at hand, the tuple (X',Y') represents pseudo-data to be generated from a *simulator*. The simulator would be a computer program producing samples from $p(X',Y'|Z)$ in the ABC framework. Here, due to the unavailability of such a generative mechanism, we consider an approximation $\hat{p}_{\mathscr{I}}(X',Y'|Z)$ that complies with $\mathscr{I}_2$ and $\mathscr{I}_3$. Such a mechanism may be based on the decomposition $\hat{p}_{\mathscr{I}}(X',Y'|Z) = \hat{p}_{\mathscr{I}_2}(X'|Z)\hat{p}_{\mathscr{I}_3}(Y'|X',Z)$ and the fact that $X'|Z$ and $Y'|X',Z$ are Bernoulli random variables whose parameters, denoted as $\theta_{X=A|Z}$ and $\theta_{Y=1|X,Z}$ for $X \in \{A,B\}$ and $Z \in \{S,L\}$, satisfy constraints implied by $\mathscr{I}_2$ and $\mathscr{I}_3$. More precisely, we get samples from $\hat{p}_{\mathscr{I}}(X',Y'|Z)$ as detailed in Algorithm 1.

Algorithm 1 SIMULATOR

Draw $\theta_{X=A|L} \sim Uniform(0.5,1)$; $\theta_{X=A|S} \sim Uniform(0,0.5)$ and $\theta_{Y=1|B,L} \sim Uniform(0,1)$
Draw $\theta_{Y=1|A,L} \sim Uniform(\theta_{Y=1|B,L},1)$
Draw $\theta_{Y=1|B,S} \sim Uniform(\theta_{Y=1|A,L},1)$
Draw $\theta_{Y=1|A,S} \sim Uniform(\theta_{Y=1|B,S},1)$

Once defined a simulator, we can exploit the knowledge deriving from observed data and get samples from an approximate distribution $\hat{p}_{\mathscr{I}}(Z,X',Y'|X,Y)$ following the ABC scheme in Algorithm 2.

Algorithm 2 ABC SCHEME

Draw $Z^{(t)} \sim p(Z)$ according to $\mathscr{I}_1 \quad \forall t \in \{1,...,T\}$
Generate $(X',Y')^{(t)} \sim \hat{p}^{(t)}_{\mathscr{I}}(X',Y'|Z)$ simulated according to Alg.1 $\quad \forall t \in \{1,...,T\}$
Accept $((X',Y')^{(t)},Z^{(t)})$ if $d(\widehat{p}^{(t)}_{X',Y'},\widehat{p}_{X,Y}) \leq \varepsilon \quad \forall t \in \{1,...,T\}$

Note that here observed and simulated data are compared by means of sufficient summary statistics, the empirical distributions $\hat{p}_{X,Y}$ and $\hat{p}_{X',Y'}$.

Looking at Alg. 2, it is apparent that the outlined sampling scheme can incorporate whatever state of knowledge simply by replacing Alg. 1 with a proper simulator based on different information $\mathscr{I}$. In any case, the output of the algorithm is a sample from an approximate posterior distribution $\hat{p}_{\mathscr{I}}(X',Y',Z|X,Y)$ which can be used to evaluate $p(Y=1|do(X=i))$ as

$$\sum_{j\in\{S,L\}} \hat{p}_{\mathscr{I}}(Y'^{(t)}=1|X'^{(t)}=i,Z^{(t)}=j)p(Z^{(t)}=j) = \sum_{j\in\{S,L\}} \theta^{(t)}_{Y=1|i,j}\cdot p(Z^{(t)}=j) \qquad i\in\{A,B\}$$

for each retained sample t. It follows that we came up with a posterior distribution of ACE based on the accepted triplets $((X',Y')^{(t)},Z^{(t)})$.

4 Experiments

We consider the example described in the previous section. Using the frequencies of Tab. 1, and formula (4) we obtain an ACE equal to 0.05.

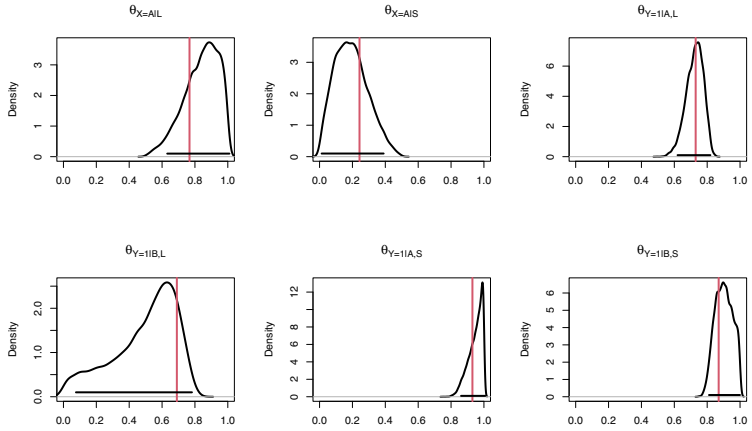


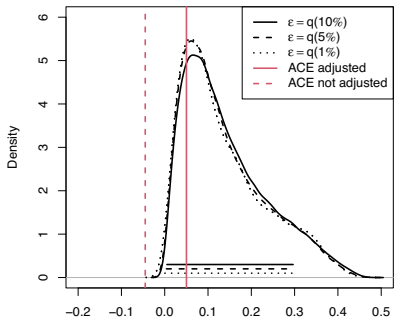
Fig. 2 Posterior distribution of the six θ parameters corresponding to the conditional probabilities of Algorithm 1.

To evaluate the performance of our method, we pretend that the values of Z are not observable and compute a posterior distribution for ACE by using $800,000$ samples from $\hat{p}_{\mathscr{I}}(Z,X',Y'|X,Y)$, obtained with the procedure described in Alg. 1 and Alg. 2.

Figure 2 shows the posterior distributions for $\theta_{X=A|S}$, $\theta_{X=A|L}$, $\theta_{Y=1|A,S}$, $\theta_{Y=1|B,S}$, $\theta_{Y=1|A,L}$ and $\theta_{Y=1|B,L}$, $\varepsilon = 0.076$, corresponding to the 1% quantile of the distance empirical dis-

tribution. Considering that before the analysis we had a uniform prior belief about their values, we can conclude that the procedure allows to learn appreciably about the six parameters. Figure 3 shows the ACE posterior distributions obtained using three different values of the ABC threshold, ε, selected through the quantile method [1].

Note that the three posterior modes are almost coincident with the true ACE, and they get closer and closer as the threshold ε decreases.



quantile	epsilon	HPD	MAP
1%	0.076	(0.000, 0.299)	0.057
5%	0.128	(0.004, 0.293)	0.059
10%	0.164	(0.005, 0.297)	0.062

Fig. 3 Posterior distributions for the adjusted ACE obtained using three different values of ε, along with the corresponding 90% HPD intervals. Table contains the specification of the ε values along with corresponding 90% HPD intervals and MAP.

5 Conclusions

In causal inference, this work provides a flexible method to evaluate ACE when observations of the variable(s) required to remove confounding are not available. The research stage is at its infancy, and we plan to consider more complex causal stories to verify the ability of the method to exploit qualitative information about the conditional probabilities beyond the causal DAG to provide useful probabilistic information about the required ACE. As a final remark, we want to stress that the method measures the confounding only under the assumption of its existence.

References

1. M. Beaumont, "Approximate Bayesian computation" *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 379–403, 2019.
2. S. Greenlan, "Relation of probability of causation, relative risk,and doubling dose: a methodlogical error that has become a social problem" *American Journal of Pubblic Healt*, vol. 89, no. 8, pp. 1166–1169, 1999.
3. S. A. Julious and M. A. Mullee, "Confounding and Simpson's paradox," *British Medical Journal*, vol. 309, pp. 1480–1481, 1994.
4. J. Pearl, "Causality" *Cambridge University Press*, 2009.
5. D. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies" *J. Educ. Psychol.*, vol. 66, no. 5, pp. 6688–701, 1995.
6. S. A. Sisson, Y. Fan, and M. Beaumont, *Handbook of approximate Bayesian computation.* Chapman and Hall\CRC., 2018.

Intertemporal propensity score matching for casual inference: an application to covid-19 lockdowns and air pollution in Northern Italy

Propensity score matching intertemporale per inferenza causale: un'applicazione su covid-19 lockdown e inquinamento atmosferico nel Nord Italia

Daniele Bondonio and Paolo Chirico

Abstract This paper develops an intertemporal propensity score matching (PSM) approach for estimating the impact of covid-19 lockdowns on air pollution. While PSM has been exclusively applied in the context of matching cross-sectional units, this paper shows that, under specific circumstances, PSM can be also applied for estimating causal inference by means of matching across different temporal units in the context of multivariate time series data. We apply our intertemporal PSM model to the data collected from a large number of air-pollution-measurement stations in Northern Italy, estimating the casual effect of the March-May-2020 lockdown on air-pollution without resorting to the more stringent functional form assumptions of the existing literature.

Abstract *Questo paper sviluppa una procedura intertemporale di propensity score matching (PSM) per stimare l'impatto dei lockdown covid-19 sull'inquinamento atmosferico. Sebbene il PSM sia stato applicato esclusivamente nel contesto dell'abbinamento di unità cross-sezionali, questo paper mostra che, in circostanze specifiche, il PSM può essere applicato anche per stimare l'inferenza causale mediante l'abbinamento tra diverse unità temporali nel contesto di serie temporali multivariate. Applichiamo il nostro modello di PSM intertemporale ai dati raccolti da un ampio numero di stazioni di rilevamento dell'inquinamento atmosferico nel Nord Italia, stimando l'effetto casuale del lockdown di marzo-maggio-2020 senza ricorrere alle più stringenti ipotesi di forma funzionale dei modelli regressivi utilizzati nella letteratura esistente.*

Key words: propensity score matching, air pollution, coronavirus lockdown

Daniele Bondonio
University of Piemonte Orientale (UPO)
e-mail: daniele.bondonio@uniupo.it

Paolo Chirico
University of Piemonte Orientale (UPO)
e-mail: paolo.chirico@uniupo.it

1 Introduction

Statistical matching is widely used as tool for estimating casual inference in the context of quasi-experimental designs in which cross-sectional units are sorted into treated and untreated units based on the exposure of the program intervention of interest. For all known statistical matching estimations, including propensity score matching (PSM), the temporal periods in which the outcome variables are measured have to remain equal across all treated and untreated units. This is a strict requirement for all quasi-experimental designs, in which the existence of the comparison group of untreated units is aimed at controlling for the secular trends that may affect Y independently from the treatment. For this reason, intertemporal PSM matching (i.e. matching across units with Y measured at different times) is not usually considered a viable impact identification strategy. The stringent covid-19 lockdowns of the first half of 2020 provide an exceptional source of an abrupt exogenous decrease on human activities (transportations and productions) that offers a unique opportunity to quantify the effect of human activities on air pollution in urbanized areas. What is unique about the impact identification conditions related to the covid-19 lockdowns, however, is also the fact that the available data are in terms of a number of multivariate time-series, one for each air-pollution measurement station, with units of observations represented by the days in which air pollution and weather characteristic are measured. These features of the data represent a scenario in which the abrupt change in the secular trend (i.e. the lockdown-induced sharp decrease in traffic and production activities) is the treatment of interest, while all the major confounding factors to be controlled for (i.e. the weather characteristics) are observable in the data and they can be safely assumed of not being subject to unobserved secular trend (within the short pre- post-treatment periods of times considered in the analysis). This paper aims at showing that in such a data scenario, an intertemporal statistical matching in which, separately for each air-pollution measurement station, the treated days (corresponding to the lockdown periods) are matched to previous non-lockdown days (untreated units) that share the same weather features, is a viable empirical option that overcomes the limitations of the recently-emerging literature on the air-quality impact of the covid-19 lockdowns.

2 Existing Literature and Impact Identification Properties of Intertemporal PSM

Since spring 2020, an increasingly large number of studies have been using covid-19 lockdowns data on a variety of locations around the world in order to estimate the reduction of air pollution caused by an abrupt decrease of traffic and human activities. A number of these papers fail to properly address the challenges of a reliable causal inference that requires, above all, a suitable controlling of the possible differences in the distribution of the weather characteristics of the pre-lockdown days

(untreated units) and the lockdown days (treated units) [1, 3, 4, 14]. These papers identify the causal effect of the lockdowns only under the very stringent assumption of a perfectly equal distribution of the weather characteristics between the lockdown and the pre-lockdown periods. Other studies make use of panel-data regression approaches that have to face the issue of error autocorrelation, and that require strong functional form assumptions with respect to the way in which the observable cofounding factors affect air-pollution [7, 5, 15, 16].

The intertemporal PSM matching approach developed in this paper can be described as follows. Let $\{Y_t\}$ be a stochastic process on which an intervention lasting for a time interval L_1 produces an effect. The causal effect of the intervention should be determined as the difference:

$$\mathrm{E}[Y_t^{(1)}|t \in L_1] - \mathrm{E}[Y_t^{(0)}|t \in L_1] \tag{1}$$

where $Y_t^{(1)}$ is the observable outcome at time t if time t were affected by the intervention, and $Y_t^{(0)}$ if not affected. In the case of the lockdown studies, the temporal units t are represented by the days in which air-pollution and weather characteristics are measured. The difference (1) can be considered as an intertemporal version of the standard *Average Treatment effects on the Treated* (ATT). As $\mathrm{E}[Y_t^{(0)}|t \in L_1]$ is not observable, the intertemporal ATT (1) can be identified in terms of

$$E_{X_1}\left[\mathrm{E}[Y_t|t \in L_1, X_1] - \mathrm{E}[Y_t|t \in L_0, X_1]\right] \quad X_1 = X_t|t \in L_1 \tag{2}$$

when two conditions apply: i) the period of observation $L = L_0 \cup L_1$, that includes also a previous time-interval L_0, is short enough so that the process Y_t can be assumed to have no underlying trend (besides the break in the observed outcome caused by lockdown period that represents the treatment under consideration); ii) the existing multivariate process $\{X_t\}$ of observable covariates (i.e. the weather controls) makes the expected value of $Y_t^{(0)}$ independent of $t \in L$, i.e.:

$$\mathrm{E}[Y_t^{(0)}|t \in L_1, X_t] = \mathrm{E}[Y_t^{(0)}|t \in L_0, X_t] \tag{3}$$

Similarly to cross-sectional PSM, if assumption (3) is satisfied, the intertemporal ATT can be also identified when conditioning on propensity score such as $\varphi_t = P[t \in L_1|X_t]$:

$$E_{\varphi_1}\left[\mathrm{E}[Y_t|t \in L_1, \varphi_1] - \mathrm{E}[Y_t|t \in L_0, \varphi_1]\right] \quad \varphi_1 = \varphi_t|t \in L_1 \tag{4}$$

This leads to the following intertemporal PSM estimation procedure: (i) treated times are matched with the untreated ones that have an equal (similar) propensity score; (ii) the sample mean of Y_t for the matched treated and control times ($M_t = 1$) is computed; (iii) the ATT is estimated as the difference of the sample means ii):

$$(\bar{y}_t|t \in L_1, M_i = 1) - (\bar{y}_t|t \in L_0, M_i = 1) \tag{5}$$

iv) the estimated ATT is validated if the balancing property of PSM is checked [13]; v) in the case of data containing multiple multivariate time-series, the procedure (i)-

(iv) is repeated separately for each time-series, yielding a series of local ATTs that are then averaged-out to a global impact estimate.

Compared to the panel-data regression approaches adopted elsewhere in the literature, this intertemporal PSM approach share the same well-known advantage [2, 8] of the standard cross-sectional PSM in strongly diminishing the sensibility of the estimated impacts to the choice of functional form that links the effects of the controls on the outcome variable.

3 The impact of covid-19 lockdowns on air pollution in northern Italy

We apply the intertemporal PSM model described in the previous section to the air-pollution data from five regions in Northern Italy located along the Po-river Valley, i.e. Piedmont, Lombardy, Emilia-Romagna, Veneto, Friuli Venezia Giulia. These regions represent the most industrialized and densely-populated area of Italy (with an average density of 252 inhab. per km2). Due to these characteristics and the specific orographic features of the Po Valley (that determine a systemic lack of sustained ventilation), this area has some of the highest levels of air-pollution in Europe. For the analysis we collected data on the average daily levels of nitrogen dioxide (NO2) and particulate matter (PM10) detected at 83 ARPA air-pollution measurement stations, located in the main urban areas of the five regions. The observation period covers both the entire length of the nationwide covid-19 lockdown period of March 10 - May 17, 2020 (that posed a very strict stay-at-home mandate and closure of all schools, universities and non-essential services) and the corresponding (untreated) period of the previous year (March 10 – May 17, 2020). The assembled database also includes as controls (X_t) the main weather characteristics that are known in the literature to influence air pollution: average daily temperature (TEMP), rainfall (RAIN), maximum daily wind speed (WIND) [7, 5, 15, 16].

Table 1 shows the average values of the air-pollutants and weather variables during the 2020 lockdown days (treatment) and the corresponding (untreated) period of 2019. These descriptive statistics do not provide a reliable indication of causal effect of the lockdown on air-pollution because the comparison between the treated and non-treated days does not control for the possible differences in the distribution of the weather variables in the two periods. Indeed, the data from Table 1 show, for example, that the lockdown (L_1) period of 2020 experienced less than half of the rain than the corresponding non-treated period of 2019 (L_0). For this reason, any casual inference drawn from a descriptive-statistics comparison between the two periods would be highly biased, particularly for the PM10 outcomes that has a high potential of being influenced by rain and wind conditions.

In order to grant a perfect balancing also of the locational characteristic of the ARPA measuring station (i.e. distance from major roads, highways, airports, surrounding production activities, etc.), we apply to the analysis our intertemporal PSM approach in terms of comparisons between the 2020 lockdown and the 2019

(untreated) period holding constant the same ARPA station. This empirical strategy entails estimating 83 different local impact parameters obtained by applying the intertemporal PSM model (implemented in terms of caliper radius matching) separately to the single multivariate time-series data from each the ARPA stations. The final (global) estimated impact is then obtained as the average of the different local impacts for which the balancing property was successfully tested by means of the Rubin's R and B indices [13].

Table 1 Descriptive statistics of air-pollution and weather characteristics in the 2020 lockdown and corresponding untreated period of 2019

period	N 83x69	NO2 (μg/m^3)	PM10 (μg/m^3)	WIND (m/s)	RAIN (mm)	TEMP (C^o)
10/3 - 17/5, 2019	5727	25.45 (16.08)[a]	20.27 (12.05)	2.28 (1.47)	2.85 (7.73)	12.55 (2.80)
10/3 - 17/5, 2020	5727	15.26 (11.12)	23.90 (17.41)	2.20 (1.49)	1.35 (5.11)	13.74 (4.15)

[a] Std.dev. in parentheses

The main results from the intertemporal PSM analysis are summarized in Table 2. Restricting the focus on the working days only, the covid-19 lockdown is shown to cause a reduction of about 13.1 ($\mu g/m^3$) in the daily average level of NO2 and a reduction of about 3.3 ($\mu g/m^3$) in the level of PM10, compared a counterfactual scenario of no-lockdown (estimated from a comparison group composed by the corresponding calendar days of 2019 with similar weather characteristics). These estimates correspond to 51.3% and a 16.2% decrease in the daily average level of NO2 and PM10, respectively. Taking into account also the week-ends and the festivities ("All days" column), the estimated impacts, albeit slightly smaller, remain of a quite large magnitude: about -12.1 ($\mu g/m^3$), corresponding to -47.7%, and -2.9 ($\mu g/m^3$), corresponding to -14.1%, for the NO2 and the PM10, respectively.

Table 2 Locdown effect on NO2 and PM10 estimated by means of intertemporal PSM

polluttant	Aggregate ATT estimates working days	all week days	Descriptive-statistics Difference $L_1 - L_0$
NO2	-13.07 *** (.221)[a]	-12.12 *** (.206)	-10.27 *** (.745)
PM10	-3.29 *** (.378)	-2.86 *** (.364)	3.60 *** (.220)

Statistical significance: *=at 10% level; **= at 5% level; ***= at 1% level
[a] Std.dev. in parentheses

Due to a higher incidence of unfavourable weather conditions in the lockdown days, compared to previous-year period, the results from a mere descriptive-

statistics comparison ($L_1 - L_0$, last column in Table 2) are indeed of a lower magnitude than the intertemporal PSM estimates. This is particularly relevant for the case of PM10, for which the descriptive-statistics analysis is suggestive of a worsening impact of the lockdown ($+3.6\mu g/m^3$), a result that highlights how misleading could be the evidence produced in the absence of adequate statistical tools for a reliable causal inference.

Acknowledgements This work was supported by funding from the University of Piemonte Orientale. We thank Giuseppe Gruttad'Auria, Andrea Nigido, Giulia Pasculli and Michela Rosselli for their contribution in the data acquisition process.

References

1. Anil, I., Alagha, O.: The impact of COVID-19 lockdown on the air quality of Eastern Province, Saudi Arabia. Air Qual Atmos Health 14: 117–128 (2020)
2. Angrist, J.: Estimating the labor market impact of voluntary military service using Social Security data on military applicants. Econometrica 66, 249–88 (1998)
3. Arregocés, H. A., Rojano, R., Restrepo, G.: Impact of lockdown on particulate matter concentrations in Colombia during the COVID-19 pandemic. Science of The Total Environment, 764, 142874 (2021)
4. Collivignarelli, M. C., Abbà, A., Bertanza, G., Pedrazzani, R., Ricciardi, P., Miino, M. C.: Lockdown for CoViD-2019 in Milan: What are the effects on air quality?. Science of the total environment, 732, 139280 (2020)
5. Dang H.A.H., Trinh, T.A.: Does the covid-19 lockdown improve global air quality? new cross-national evidence on its unintended consequences. Journal of Environmental Economics and Management 105:102–126 (2021)
6. Dehejia, R.H., Wahba, S.: Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 84(1), 151-161 (2002)
7. He G., Pan Y., Tanaka, T.: The short-term impacts of covid-19 lockdown on urban air pollution in china. Nature Sustainability 3(12):1005–1011 (2020)
8. Heckman, J.J., Ichimura, H., Todd, P.: Matching as an econometric evaluation estimator. The review of economic studies, 65(2), 261-294 (1998)
9. Holland, P.W.: Statistics and causal inference. Journal of the American statistical Association 81(396):945–960 (1986)
10. Imbens, G.W., Rubin D.B.: Causal inference in statistics, social, and biomedical sciences. Cambridge University Press (2015)
11. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55 (1983)
12. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66(5):688 (1974)
13. Rubin D.B.: Using propensity scores to help design observational studies: application to the tobacco litigation. Health Services and Outcomes Research Methodology 2(3):169–188 (2001)
14. Seo, J.H., Jeon, H.W., Sung, U.J., Sohn, J.R. (2020). Impact of the COVID-19 outbreak on air quality in Korea. Atmosphere, 11(10), 1137.
15. Song, Y., Li, Z., Liu, J., Yang, T., Zhang, M., Pang, J.: The effect of environmental regulation on air quality in china: A natural experiment during the covid-19 pandemic. Atmospheric Pollution Research 12(4):21–30 (2021)
16. Wang, J., Xu, X., Wang, S., He, S., He, P.: Heterogeneous effects of covid-19 lockdown measures on air quality in northern china. Applied Energy 282:116–179 (2021)

Methods for Spatio-temporal data

Local Spatio-Temporal Log-Gaussian Cox Processes for seismic data analysis

Processi Log-Gaussiani di Cox locali e spazio-temporali per l'analisi di dati sismici

Nicoletta D'Angelo, Giada Adelfio, and Jorge Mateu

Abstract We propose a local version of the spatio-temporal log-Gaussian Cox processes (LGCPs) employing the Local Indicators of Spatio-Temporal Association (LISTA) functions into the minimum contrast procedure to obtain space as well as time-varying parameters. We resort to the joint minimum contrast method fitting method to estimate the set of second-order parameters for the class of Spatio-Temporal LGCPs. We employ the proposed methodology to analyse real seismic data occurred Greece between 2004 and 2015.

Abstract *Proponiamo una versione locale dei processi spazio-temporali Log-Gaussiani di Cox (LGCP) che impiegano le funzioni LISTA (Indicatori spazio-temporali locali di associazione) nella procedura di minimo contrasto per ottenere parametri variabili sia nello spazio che nel tempo. Si ricorre al metodo di adattamento del metodo del minimo contrasto congiunto per stimare l'insieme dei parametri del secondo ordine per la classe degli LGCP spazio-temporali. Vi presentiamo un'analisi dei dati reali dei dati sismici in Grecia.*

Key words: Earthquakes, Second-order characteristics, Spatio-temporal point processes, Local models, Log-Gaussian Cox Processes, Minimum contrast

1 Introduction

Local extensions of spatio-temporal point process models are very welcome in many fields of study, such as epidemiology and seismology. Indeed, one could be interested in identifying the most inhomogeneous locations, both in space and time, to

Nicoletta D'Angelo and Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy e-mail: nicoletta.dangelo@unipa.it; giada.adelfio@unipa.it

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain e-mail: mateu@uji.es

be further examined separately. For spatial point process, Baddeley (2017) presents a general framework based on the local composite likelihood to detect and model gradual spatial variation in any parameter of a spatial stochastic model (such as Poisson, Gibbs and Cox processes). In particular, the parameters in the model that govern the intensity, the dependence of the intensity on the covariates and the spatial interaction between points, are estimated locally. D'Angelo et al. (2022) show that these spatial local models provide good inferential results, if applied to seismic data.

Motivated by this, we propose a local version of the spatio-temporal log-Gaussian Cox processes (LGCPs) employing the Local Indicators of Spatio-Temporal Association (LISTA) functions in the minimum contrast procedure to obtain space as well as time-varying parameters. The underlying hypothesis is that also temporal characteristics vary within the point process, in addition to the spatial ones, as proved in D'Angelo et al. (2022). From a methodological point of view, we extend the joint minimum contrast method (Siino et al., 2018) to the local context, managing to estimate a set of second-order parameters of the spatio-temporal LGCPs for each point.

The structure of the paper is the following. In Section 2 the spatio-temporal log-Gaussian Cox Processes are recalled, as well as the joint minimum contrast procedure. Section 3 contains the proposed method to estimate local parameters, and it is applied in Section 4 to earthquakes occurred in Greece between 2005 and 2014. Conclusions are drawn in Section 5.

2 Spatio-temporal log-Gaussian Cox processes

In the point process theory, log-Gaussian Cox processes are the most prominent clustering models. By specifying the intensity of the process and the moments of the underlying Gaussian Random Field (GRF), it is possible to estimate both the first and second-order characteristics of the process. Following the inhomogeneous specification in Diggle (2013), a log-Gaussian Cox process for a generic point with $\mathbf{u}$ and t coordinates in space and time, has the intensity

$$\Lambda(\mathbf{u},t) = \lambda(\mathbf{u},t)\exp(S(\mathbf{u},t))$$

where S is a Gaussian process with $\mathbb{E}(S(\mathbf{u},t)) = \mu = -0.5\sigma^2$ and so $\mathbb{E}(\exp S(\mathbf{u},t)) = 1$ and with variance and covariance matrix $\mathbb{C}(S(\mathbf{u}_i,t_i),S(\mathbf{u}_j,t_j)) = \sigma^2\gamma(r,h)$, with $\gamma(\cdot)$ the correlation function of the GRF, and r and h some spatial and temporal distances. Following Møller et al. (1998), the first-order product density and the pair correlation function of a log-Gaussian Cox process are $\mathbb{E}(\Lambda(\mathbf{u},t)) = \lambda(\mathbf{u},t)$ and $g(r,h) = \exp(\sigma^2\gamma(r,h))$, respectively.

We consider a separable structure for the covariance function of the GRF that has exponential form for both the spatial and the temporal components,

$$\mathbb{C}(r,h) = \sigma^2 \exp\left(\frac{-r}{\alpha}\right)\exp\left(\frac{-h}{\beta}\right), \tag{1}$$

where σ^2 is the variance, α is the scale parameter for the spatial distance and β is the scale parameter for the temporal one. The exponential form is widely used in this context and nicely reflects the decaying correlation structure with distance or time. Moreover, we can consider a non-separable covariance of the GRF useful to describe more general situations.

In general, the Cox model is estimated by a two-step procedure, involving first the intensity and then the cluster or correlation parameters. First, a Poisson model with the same model formula is fitted to the point pattern data, providing the estimates of the coefficients of all the terms in the model formula that characterize the intensity. Second, the estimated intensity is taken as the true one and the cluster or correlation parameters are estimated by one among the method of minimum contrast, Palm likelihood, or composite likelihood.

The most common technique is the *minimum contrast*. Let the function J represent either the pair correlation function g of the K-function, and $\hat{J}$ stands for the corresponding non-parametric estimate. Siino et al. (2018) propose a new fitting method to estimate the set of second-order parameters for the class of spatio-temporal log-Gaussian Cox point processes with constant first-order intensity function. We denote by θ the vector of (first-order) intensity parameters, and by ψ the cluster parameters, also denoted as correlation or interaction parameters. In the case of a spatio-temporal log-Gaussian Cox process with exponential covariance as the one in Equation (1), the cluster parameters correspond to $\psi = \{\sigma, \alpha, \beta\}$.

The vector of estimates $\hat{\psi}$ is found by minimizing

$$M_{J,i}\{\psi_i\} = \int_{h_0}^{h_{max}} \int_{r_0}^{r_{max}} \phi(r,h)\{\nu[\bar{J}_i(r,h)] - \nu[J(r,h;\psi)]\}^2 \mathrm{d}r\mathrm{d}h,$$

where $\phi(r,h)$ is a weight that depends on the space-time distance and ν is a transformation function. Its main advantage is that it can be used in the case of both separable and non-separable parametric specifications of the correlation function of the underlying GRF, representing a more flexible method with respect to other available methods.

3 Locally weighted spatio-temporal minimum contrast

In the purely spatial context, a localised version of minimum contrast is developed using the local K-functions or local pair correlation functions by Baddeley (2017), bearing a very close resemblance to the local Palm likelihood approach.

Combining the *joint minimum contrast* (Siino et al., 2018) and the *local minimum contrast* (Baddeley, 2017) procedures, we can obtain a vector of parameters $\hat{\psi}_i$ for each point i, by minimizing

$$M_{J,i}\{\psi_i\} = \int_{h_0}^{h_{max}} \int_{r_0}^{r_{max}} \{\bar{J}_i(r,h) - J(r,h;\psi)\}^2 \mathrm{d}r\mathrm{d}h, \tag{2}$$

where $\bar{J}_i(r,h)$ is the average of the local functions $\hat{J}_i(r,h)$, weighted by some pointwise kernel estimates. This procedure not only provides individual estimates, but it does also account for the vicinity of the observed points, and therefore the contribution of their displacement on the estimation procedure. Thus, consider the weights $w_i = w_{i,\sigma_s} w_{i,\sigma_t}$ given by some kernel estimates, where w_{i,σ_s} and w_{i,σ_t} are weight functions and $\sigma_s, \sigma_t > 0$ are the smoothing bandwidths. It is not necessary to assume that w_{σ_s} and w_{σ_t} are probability densities. For simplicity, we shall consider only kernels of fixed bandwidth, even though spatially adaptive kernels could also be used. Then, the averaged weighted local statistics $\bar{J}_i(r,h)$ in Equation (2) are

$$\bar{J}_i(r,h) = \frac{\sum_{i=1}^{n} \hat{J}_i(r,h) w_i}{\sum_{i=1}^{n} w_i},$$

one for each point i. In particular, we consider $\hat{J}_i(\cdot)$ as the local spatio-temporal pair correlation function (in short, pcf-function)

$$\hat{J}_i(r,h) = \hat{g}_i(r,h) = \frac{1}{4\pi r |W \times T| \hat{\lambda}^2} \sum_{j \neq i} \frac{\kappa_{\varepsilon,\delta}(||\mathbf{u}_i - \mathbf{u}_j|| - r, |t_i - t_j| - h)}{\omega(\mathbf{u}_i, \mathbf{u}_j)\omega(t_i, t_j)} \tag{3}$$

where ω is the edge correction factor. The kernel function κ has a multiplicative form $\kappa_{\varepsilon,\delta}(||\mathbf{u}_i - \mathbf{u}_j|| - r, |t_i - t_j| - h) = \kappa_\varepsilon(||\mathbf{u}_i - \mathbf{u}_j|| - r)\kappa_\delta(|t_i - t_j| - h)$ where κ_ε and κ_δ are kernel functions with bandwidths ε and δ, respectively. Both of them are computed using the Epanechnikov kernel and the bandwidths are estimated with a direct plug-in method.

After having properly chosen the bandwidth of the weight in both the minimum contrast procedure and in the pair correlation function, the result of the maximization of Equation (2) provides a set of covariance parameters $\hat{\psi}_i = \{\hat{\sigma}_i^2, \hat{\alpha}_i, \hat{\beta}_i\}$ for each point i, which can be also interpolated along the whole spatio-temporal area under study.

4 Seismic data analysis

The analysed data concern 1111 earthquakes occurred in Greece between 2005 and 2014, and come from the Hellenic Unified Seismic Network (H.U.S.N.). Only seismic events with a magnitude larger than 4 are considered in this study.

In Figure 1 (a) earthquakes are reported, while in Figure 1 (b) the observed pcf-function is represented. The theoretical value indicating Complete Randomness is $pcf = 1$. The pcf statistic of the observed point pattern, with larger values than the theoretical one of a Poisson process especially for shortest distances, suggests that distances among the point of the observed pattern are shorter than the Poisson ones. In other words, events are more clustered than an homogeneous Poisson pattern, as it is also evident in Figure 1 (a).

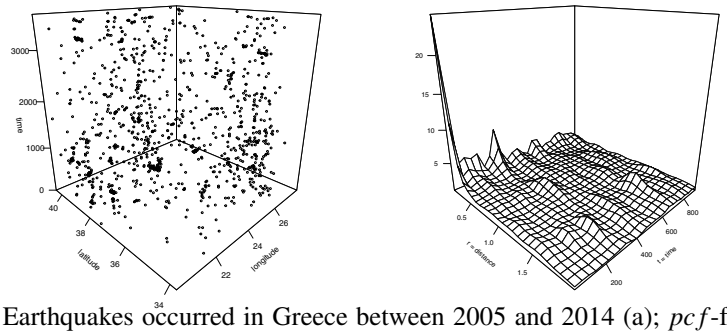


Fig. 1: Earthquakes occurred in Greece between 2005 and 2014 (a); pcf-function of the observed point pattern of the Greek data (b).

A spatio-temporal log-Gaussian Cox process is fitted, following the locally weighted minimun contrast proposed in Section 3 with exponential covariance function as in Equation (1). The bandwidths ε and δ, for the kernel in the pair correlation function are equal to 0.15 and 28.49, for space and time respectively, while the bandwidths $\{\sigma_x, \sigma_y, \sigma_t\}$ are chosen to be $2.58, 1.79, 1140.31$, for x, y, and t coordinates respectively.

The results of the fitting come in Table (1) and in Figure (2). As shown in Table (1), the proposed procedure provides a whole distribution for each parameter. Then, as evident from Figure (2), the estimated parameters allow to clearly distinguish different sub-areas where point behave differently from each other. Indeed, by the inspection of the top panels of Figure (2), we can clearly spot two main spatial regions: the one on the top and on the left with higher $\hat{\sigma}_i^2$ values but lower estimates for $\hat{\alpha}_i$ and $\hat{\beta}_i$, and the area on the bottom-right which displays the opposite situation (low variance but large scale parameters).

This result is particularly appealing because it gives us more insight in the local behaviour of the seismic phenomenon. Indeed, from the global estimates (see Table (1)) we could draw the conclusion that the analysed process is overall clustered (given the high estimated variance, and the relatively small scale parameters). Conversely, our application, shows that earthquakes occurred on the top-left of the analysed area are grouped in smaller clusters, while the ones on the bottom-left have a way more diffuse clustering behaviour.

Furthermore, by the inspection of the bottom panels of Figure (2), we can assess also the time-varying behaviour of the estimates and therefore on the underlying process: the most relevant result concerns the variability of the covariance parameters $\hat{\sigma}_i$ in time, indicating that the number of earthquakes tends to increase in time, and in particular, in the top-left spatial region.

Table 1: Estimates of both the local and global LGCPs fitted to the data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Global
σ^2	1.28	5.19	6.03	5.77	6.60	12.47	6.14
α	0.13	0.25	0.28	0.39	0.35	1.94	0.27
β	42.34	366.43	534.86	535.52	597.59	1411.45	449.55

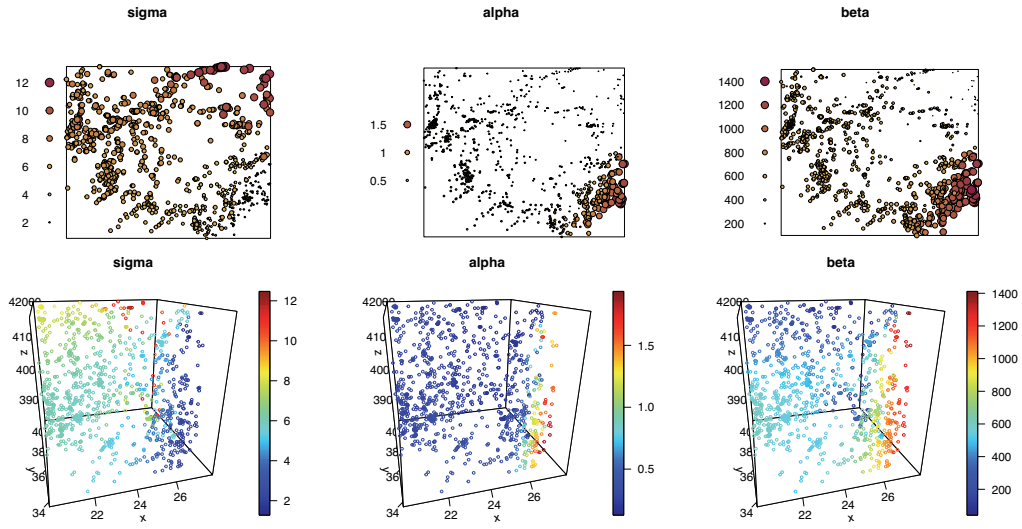


Fig. 2: Local estimates in space (*top panels*) and space-time (*bottom panels*).

5 Conclusions

In this paper, we have introduced a novel local fitting procedure for obtaining space-time varying estimated for a log-Gaussian Cox process fitted to the data. From a methodological point of view, we have resorted to the joint minimum contrast procedure (which is appealing for its flexibility in dealing also with non-separable covariances), extending it to the local context, and therefore allowing to obtain a whole set of covariance parameters for each point of the analysed process.

The motivating problem came from the seismic application, where of course it is of interest studying the characteristics of the process in relation to both the spatial and the temporal occurrence of points. Our proposal could pose the basis for many further investigations and applications, such as in the epidemiological context.

References

Baddeley, A. (2017). Local composite likelihood for spatial point processes. *Spatial Statistics*, 22:261–295.

D'Angelo, N., Siino, M., D'Alessandro, A., and Adelfio, G. (2022). Local spatial log-gaussian cox processes for seismic data. *AStA Adv. Stat. Anal. Accepted*.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Siino, M., Adelfio, G., and Mateu, J. (2018). Joint second-order parameter estimation for spatio-temporal log-gaussian cox processes. *Stochastic environmental research and risk assessment*, 32(12):3525–3539.

Spatial explorative analysis of thyroid cancer in Sicilian volcanic areas

Analisi spaziale esplorativa del tumore tiroideo nelle zone vulcaniche della Sicilia

Francesca Bitonti, Angelo Mazza

Abstract In the most advanced countries the epidemiological transition has led non-communicable diseases to dominate the causes of death (ruling out the recent Covid-19 pandemic). Therefore, the spatial methodologies traditionally implemented to study the infectious diseases have been also exploited to analyse those non-communicable diseases whose causes include environmental risk factors. In this work, we explore the spatial distribution of thyroid cancer (TC) near Mt Etna (Sicily). It is well-known in literature that active volcanoes emit carcinogens substances. Here, we apply the standardised incidence ratio (SIR) and the local Moran I to verify the presence of high-risk pockets, considering both the proximity to the volcano and the demographic heterogeneity in the at-risk population.

Abstract *Nei Paesi avanzati, in seguito alla transizione epidemiologica, le malattie non trasmissibili sono diventate la prima causa di mortalità (escludendo la recente pandemia da Covid-19). Pertanto, le metodologie spaziali tradizionalmente implementate per lo studio delle malattie infettive sono state applicate anche alle malattie non trasmissibili le cui cause includono fattori di rischio ambientale. In questo lavoro, esploriamo la distribuzione spaziale del tumore tiroideo (TT) vicino all'Etna (Sicilia). È noto in letteratura che i vulcani attivi emettono sostanze cancerogene. Qui, applichiamo il rapporto standardizzato di incidenza (RSI) e il Moran I locale per verificare la presenza di aree ad alto rischio, considerando sia la vicinanza al vulcano che l'eterogeneità demografica nella popolazione a rischio.*

Keywords: Spatial demography, Standardised Incidence Ratio, Local Moran's I, Thyroid Cancer

[1] Francesca Bitonti, University of Catania; email: francesca.bitonti@phd.unict.it
Angelo Mazza, University of Catania; email: a.mazza@unict.it

1 Introduction

In the most advanced countries experiencing the third phase of the epidemiologic transition, infectious disease pandemics (ruling out the recent Covid-19 pandemic) are replaced by degenerative diseases as major causes of death [24]. Specific patterns of human behaviour (e.g., violence, accidents, or substance abuse) overtake infectious agents as the major contributor to morbidity and mortality. In these contexts, the use of the spatial analysis methodologies, typical approaches to the analysis of infectious diseases, has also spread within the study of non-communicable diseases, in which environmental risk factors are often included among the possible contributing causes [1,7,12,13,23,25].

Thyroid cancer (TC), the object of this study, is the most widespread endocrine neoplasm, whose incidence has steadily increased worldwide in recent decades [9,11,19]. Epidemiological studies suggest that the aetiology of TC includes iodine consumption, familiarity, and exposure to ionising radiation [16,28]. Kung et al. [18] identified the elements present in volcanic gases as possible further TC etiological agents. Other studies have confirmed a higher TC incidence than national averages in the Hawaiian Islands [14,17], Iceland [3,15], the Philippines [8,10], New Caledonia [29] and French Polynesia [6,9], all regions whose common denominator is the presence of active volcanoes. Goodman et al. [14] compared the standardised TC incidence rates of different ethnic groups residing in Hawaii with those of the same groups residing in other geographic areas. Also in Sicily, a region located in South Italy, higher TC incidence rates than the regional and national averages were detected in the vicinity of an active volcanic areas, [21,22,30,31]. In particular, a double incidence compared to that of the region was observed near Mt Etna [27].

This work is organized as follows: the next section describes the applied methodology and the exploited data sources; the third section discusses the main findings.

2 Methodology and analyzed data

The spatial distribution of TC cases can be described through the inhomogeneous Poisson point process. In this model, the expected number of events $N(u)$ in a generic sub-area U of the region under study R follows a Poisson distribution whose intensity λ varies over space. For a generic point u we will have $\lambda=\lambda(u)$ and the expected number of events $E[N(u)]$ is

$$E[N(U)] = \int_U \lambda(u)du\,. \tag{1}$$

The Standardised Incidence Ratio (SIR) is computed at census tract level and is exploited to produce maps of TC risk. The SIRs were calculated by indirect standardization [32]. The SIR is the ratio between observed O_i and expected TC cases E_i in each census tract i

$$RSI = \frac{O_i}{E_i} \tag{2}$$

The expected number of cases is computed as the product of the population at risk P_i (and therefore the entire resident population) in the given census tract i and the general incidence rate r_+ for the entire investigated area

$$E_i = P_i r_+ \quad . \tag{3}$$

Hence it follows that the SIR of a single census tract is thus computed as

$$RSI = \frac{O_i}{P_i \frac{O_+}{P_+}} \quad . \tag{4}$$

where O_+ corresponds to the total number of observed TC cases and P_+ is the total resident population of eastern Sicily. When the characteristics of the population determine a subdivision into strata with different risk levels (for instance, by sex and age group), a proper weight is associated with each stratum based on the risk of the specific stratum. In this case, the general incidence rate r_+ is replaced by a different rate computed for each stratum j as $r_j = \frac{\sum_i O_{ij}}{\sum_i P_{ij}}$. Hence, the expected number of cases in census tract i is given by $E_i = \sum_j P_{ij} r_j$. Any deviation of the observed SIR distribution from the theoretical one, obtained if the allocation of TC cases over the territory were random, would depend on other variables than the demographic structure of the resident population. Comparing the two distributions enables to verify, for each sub-area, whether potential environmental risk variables determine significant departures in TC risk from the expected one. The inferential test is based on the Poisson distribution and admits that under the null hypothesis, the SIR is equal to 1 and that an observed value of SIR significantly greater than 1 implies an increase in risk [4].

The local Moran's I was proposed by Anselin [2] and it is defined by the following formula

$$I_i = \frac{(y_i - \bar{y})}{S_i^2} \sum_{j=1, j \neq i}^{n} \left(w_{ij} \left(y_j - \bar{y} \right) \right) \tag{5}$$

where n is the number of geographical units, y_i is the value of the variable y in region i, $\bar{y}$ is the sample mean of the variable, y_j is the value of the variable y in unit j, S_i^2 is the sample variance of the variable y, and w_{ij} is a weight that can be defined as the inverse of the distance between the various geographical units. There are several alternatives to define w_{ij} , some contemplate the choice of a limit distance to define the neighbourhood of a given geographical unit: the units falling within the limit distance take on a positive weight, while the external ones take on a weight equal to zero. Positive values of local Moran's I for a given region indicate that the region is surrounded by regions that have mostly similar values of the variable under study. The emerging spatial groups are defined as "high-high" (region with a high value surrounded by regions with high values) or "low-low" (region with low value surrounded by regions with low values). Spatial outliers result in the composition of two cluster types: "high-low", in the presence of a high value surrounded by neighbours with low values, or "low-high", viceversa. To test the statistical significance of the local Moran's statistic we performed conditional permutation proposed by Anselin (1995).

The data available come from the Cancer Registry of Eastern Sicily and refer to TC cases diagnosed in the period 2003-2016 to residents in the four provinces of interest with age between 5-95 years. Overall, 7,085 cases are included. The data

concerning the resident population in the census tracts of the four provinces of interest come from the 15th General Population Census carried out by Istat in 2011 (https://www.istat.it/it/archivio/104317). Google Maps API was exploited to geocode the residential addresses of TC cases (https://cloud.google.com/maps-platform/).

3 Results and discussion

The TC cases distribution is influenced by the spatial arrangement and the age and sex structure of the resident population. It's well-known that adult females are the most affected by TC [26]; furthermore, the population ageing in inner areas, where depopulation dynamics are in place, tends to accentuate the imbalance between sexes, given the higher life expectancy of women compared to men [33]. In an attempt to distinguish the effects of the demographical and environmental components on the TC cases spatial distribution, we propose SIR maps at the census tract level and together with the geographical representation of the statistical significance. Figures 1.a-1.b respectively report the SIR and the results of the one-tailed test with significance level α=0.05 highlighting the census tracts with SIR significantly greater than 1. From the mere SIR calculation (fig. 1.a) different areas of risk emerge. These areas are located in the area nearby Mt Etna, but also in the non-volcanic provinces, especially in those of Enna and Messina. The area south-east of Mt Etna and different sections located mainly in the Messina province prove to be statistically significant from the one-tailed test (fig. 1.b). Finally, whether in the non-volcanic provinces the census

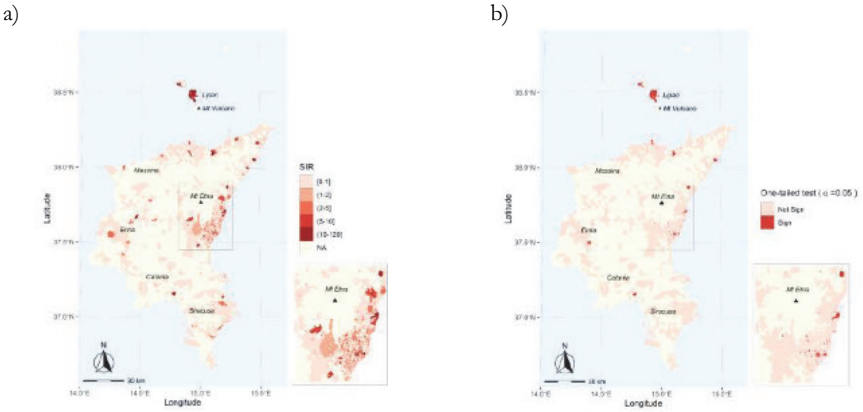


Figure 1: SIR distribution by census tracts (a) and representation of their statistical significance (b). Source: authors' original elaboration.

sections with SIR greater than 1 are randomly arranged, in the province of Catania the high-risk sections mainly concentrate in an area close to Mt Etna, leaving the rest of the province almost empty. To detect the location of the TC cases clusters

over the area and to visually assess their arrangement in relation to the proximity of Mt Etna, the local Moran's I on the SIR was computed. Figures 2.a-b show the local statistic for all TC cases with a bandwidth of 10 km and the pseudo p-values obtained from the conditioned permutation procedure, respectively. Two high-risk areas appear: the first in the south-eastern side of Mt Etna and the other to the north, between the Aeolian Islands and near Barcellona Pozzo di Gotto. The location of the high-risk areas along the NW-SE axis could suggest that the presence of persistent winds toward the SE direction, documented by Boffetta et al. [5], can carry the toxic volcanic substances and therefore pollute the atmosphere of the territories positioned along the axis itself. A similar consideration could be applied to the Aeolian Islands and the sections near Barcellona Pozzo di Gotto which are in proximity of Mt Vulcano, an active volcano presenting limited activity compared to that of Mt Etna.

a) b)

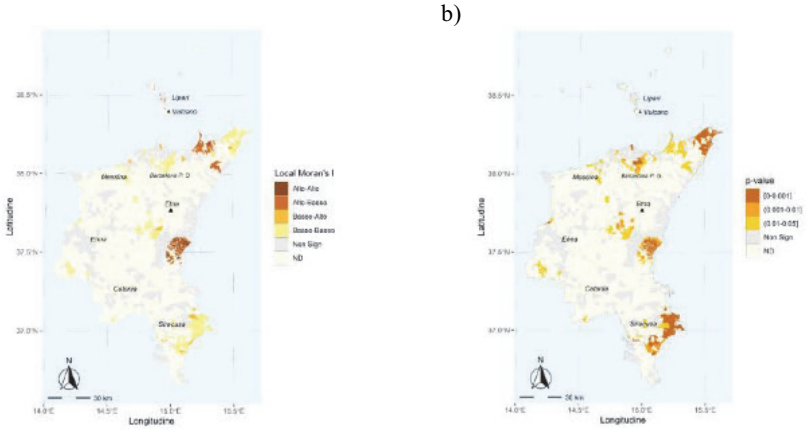


Figure 2: Cluster map based on the Local Moran's I (a) and related pseudo p-values (b) computed by conditional permutation method. Source: authors' original elaboration.

References

1. Anand, S., Staniec, A., Montez-Rath, M., Vlahos, P.: Using GIS Mapping to Track Hot Spots of Kidney Disease in California. New Engl. J. Med., 382, 23, 2265-2267, (2020)
2. Anselin, L.: Local Indicators of Spatial Association-LISA. Geogr. Anal., 27, 93-115, (1995)
3. Arnbjörnsson, E., Arnbiörnsson, A., Ólafsson, A.: Thyroid cancer incidence in relation to volcanic activity. Arch. Environ. Health, 41, 1, 36-40, (1986)
4. Bivand, R.S., Pebesma, E., Gómez-Rubio, V.: Applied Spatial Data Analysis with R. New York: Springer, (2008)
5. Boffetta, P., Memeo, L., … Sciacca, S.: Exposure to emissions from Mount Etna (Sicily, Italy) and incidence of thyroid cancer: a geographic analysis. Sci. Rep., 10, 21298, (2020)
6. Bray, F., Colombet, M., … Ferlay, J. (eds.): Cancer incidence in five continents Vol. XI. Lyon: Int. Agen. Res. Cancer, (2017)
7. Bristow, R.E., Chang, J., … Vieira, V.M.: Spatial analysis of advanced-stage ovarian cancer mortality in California. Am. J. Obstet. Gynecol., 213, 1, 43.e1-43.e8, (2017)
8. Caguioa, P.B., Bebero, K.G.M., Bendebel, M.T.B., Saldana, J.S.: Incidence of thyroid carcinoma in the Philippines: A retrospective study from a tertiary university hospital. Ann. Oncol., 30, (2019)
9. Curado, M.P., Edwards, B., … Boyle, P.: Cancer incidence in five continents, Vol. IX, (2019)

10. Duntas, L.H., Doumas, C.: The “rings of fire” and thyroid cancer. Hormones, \textbf{8}, 4, 249-253, (2009)
11. Fitzmaurice, C., Dicker, D., … Naghavi, M.: The Global Burden of Cancer 2013. JAMA Oncol., 1, 4, 505-527, (2015)
12. Geraghty, E.M., Balsbaugh, T., Nuovo, J., Tandon, S.: Using geographic information systems (GIS) to assess outcome disparities in patients with type 2 diabetes and hyperlipidemia. J. Am. Board Fam. Med., 23, 1, 88-96, (2010)
13. Ghosh, M., Natarajan, K., Waller, L.A., Kim, D.: Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. J. Stat. Plan. Infer. Vol. 75, (1999)
14. Goodman, M.T., Yoshizawa, C.N., Kolonel, L.N.: Descriptive epidemiology of thyroid cancer in Hawaii. Cancer, 61, 1272-1281, (1988)
15. Hrafnkelsson, J., Tulinius, H., … Sigvaldason, H.: Papillary Thyroid Carcinoma in Iceland: A study of the Occurrence in families and the coexistence of other primary tumours. Acta Oncol., 28, 6, 785-788, (1989)
16. Kitahara, C.M., Sosa, J.A.: The changing incidence of thyroid cancer. Nat. Rev. Endocrinol., 12, 11, 646-653, (2016)
17. Kolonel, L.N., Hankin, J.H., … Hinds, M.W.: An epidemiologic study of thyroid cancer in Hawaii. Cancer Cause. Control, 1, 223-234, (1990)
18. Kung, T.M., Ng, W.L., Gibson, J.B.: Volcanoes and carcinoma of the thyroid: A possible association. Arch. Environ. Health, 36, 5, 265-267, (1981)
19. Liu, Y., Su, L., Xiao, H.: Review of Factors Related to the Thyroid Cancer Epidemic. Int. J. of Endocrinol. (2017)
20. Mahdavifar, N., Pakzad, R., … Salehiniya, H.: Spatial analysis of breast cancer incidence in Iran. Asian Pac. J. Cancer P., 17, 59-64, (2016)
21. Malandrino, P., Russo, M., … Vigneri, R.: Increased thyroid cancer incidence in a basaltic volcanic area is associated with non-anthropogenic pollution and biocontamination. Endocrine, 53, 471-479, (2016)
22. Malandrino, P., Scollo, C., … Mian, C.: Descriptive epidemiology of human thyroid cancer: experience from a regional registry and the “volcanic factor.” Front. Endocrinol., 4, 65, 1-7, (2013)
23. Musa, G.J., Chiang, P.H., … Hoven, C.W.: Use of GIS Mapping as a Public Health Tool–-From Cholera to Cancer. Health Serv. Ins., 6, (2013)
24. Omran, A.R.: The epidemiologic transition. A theory of the epidemiology of population change. Milbank. Meml. Fund. Q., 49, 4, 509-538, (1971)
25. Palagiano, C., Pesaresi, C.: La salute nel mondo. Geografia medica e qualità della vita. Roma: Carocci, (2011)
26. Parkin, D.M., Bray, F., Ferlay, J., Pisani, P.: Global Cancer Statistics, 2002. CA-Cancer J. Clin., 55, 2, 74-108, (2005)
27. Pellegriti, G., De Vathaire, F., … Vigneri, R.: Papillary thyroid cancer incidence in the volcanic area of sicily. J. Natl. Cancer I., 101, 1575-1583, (2009)
28. Seib, C.D., Sosa, J.A.: Evolving Understanding of the Epidemiology of Thyroid Cancer. Endocrin. Metab. Clin. N. Am., 48, 1, 23-35, (2019)
29. Truong, T., Rougier, Y., … Guénel, P.: Time trends and geographic variations for thyroid cancer in New Caledonia, a very high incidence area (1985-1999). Eur. J. Cancer Prev., 16, 1, 62-70, (1985).
30. Vigneri, R, Malandrino, P., … Vigneri, P.: Heavy metals in the volcanic environment and thyroid cancer. Mol. Cell. Endocrinol., 457, 73-80, (2017)
31. Vigneri, R., Malandrino, P., Vigneri, P.: The changing epidemiology of thyroid cancer: Why is incidence increasing? Curr. Opin. Oncol., 27, 1-7, (2015)
32. Waller, L.A., Gotway, C.A.: Applied Spatial Statistics for Public Health Data. Hoboken: John Wiley & Sons, (2004)
33. World Population Ageing (2017). Retrieved from https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf

Using geo-spatial topic modelling to understand the public view of Italian Twitter users: a climate change application

Utilizzo della modellazione geospaziale degli argomenti per comprendere la visione pubblica degli utenti italiani su Twitter: un'applicazione sul cambiamento climatico

Yuri Calleo[a] and Francesco Pilla[a]

Abstract The spread of climate change uncertainty among citizens requires the adoption of different methods to understand the public view and to take efficient actions. In the Big Data framework, a large amount of data from location-aware devices are available, which can constitute a valuable information resource. However, in social networks, the lack of geo-spatial data still remains a challenge and for this reason, we implemented a new approach to extract a larger sample, using spatial data mining techniques. Finally, to detect the Italian Twitter users' discussion and the geographical areas of origin, we applied probabilistic topic models, to a dataset of tweets in the climate change context.

Abstract *La diffusione di incertezza sui cambiamenti climatici tra i cittadini richiede l'adozione di diversi metodi per comprendere l'opinione pubblica e per intraprendere azioni efficienti. Nel framework dei Big Data una grande quantità di dati provenienti da dispositivi location-aware è disponibile, i quali possono costituire un'utile risorsa informativa. Tuttavia, nei social network, la mancanza di dati geospaziali rimane*

[a] Yuri Calleo, Spatial Dynamics Lab, School of Architecture, Planning and Environmental policy, University College Dublin, Ireland; email: yuri.calleo@ucdconnect.ie

Francesco Pilla, Spatial Dynamics Lab, School of Architecture, Planning and Environmental policy, University College Dublin, Ireland; email: francesco.pilla@ucd.ie

ancora una sfida e per questo motivo implementiamo un nuovo approccio per estrarre maggiori dati, utilizzando tecniche di data mining spaziale. Per rilevare le discussioni degli utenti Twitter italiani e le aree geografiche di provenienza, applichiamo topic models su un dataset di tweet sul cambiamento climatico.

Key words: geo-spatial data, topic modeling, climate change

1. Introduction

In the current climate change context, our planet is facing numerous environmental challenges that are threatening the future of public health and the well-being of citizens. Studies demonstrate [1] how citizens' health is compromised by different climate impacts, including environmental degradation, urban pollution, rising temperatures, heatwaves, and storms. In this context, climate complications [2] could not only lead to possible environmental disasters but they are spreading anxiety and fear among citizens, unaware of the future perspectives [3]. From what was previously described, we need valid solutions to reduce concern among citizens, starting from the study of their perceptions and visions.

With the exponential increase of data available from the use of social networks, it is possible to extract valuable resources to understand the public view of users. The availability of textual data is increasing in social network platforms, and with the use of location-aware devices (LAS), more geo-spatial data are proliferating (we refer here in a broad meaning, as information useful to link textual data to a geographic object), making possible different methodological implementations. However, due to their sensibility and low presence, the extraction and analysis of geospatial data still remains a challenge.

Normally, on Twitter, the percentage of users attaching their location to textual information (tweets) is low (as we will see below, less than 10% of the total corpus). For these reasons this study developed here a different method to extract and manage geo-spatial data, combining users' profile location to each tweet to then apply the Latent Dirichlet Allocation (LDA), extracting the major topics in the Italian climate change context. This contribution provides an answer to the following research questions (RQ): RQ1: does this approach improve the number of geospatial data obtained? RQ2: what benefit can the combination of spatial and textual techniques provide for understanding the public debate on climate change? And finally: what are the most discussed topics and in which geographical areas? (RQ3).

2. Materials and Methods

In order to understand the public view of the Italian Twitter users, we use unsupervised learning models extracting major topics within a dataset of tweets, in a

time span of 15 years. A dataset of 120.222 tweets was extracted using a Python Scraper script, which allows us to obtain tweets by generating an extraction-loop. We perform it through a cleaning process that provides the tokenization (we use the NLTK Python package and the nltk.word_tokenize function) [4], the conversion in lowercase, and finally the special characters (e.g., "#", "@", "*") and stop words removal. For this preliminary analysis, we use here: "cambiamento climatico" as a principal keyword, obtaining tweets starting from the newest to the oldest (November 19, 2021, to June 26, 2007). Once we obtain the data, two datasets are available: 1) containing users' information (e.g., name, username, location, URL, etc.) and 2) containing textual data. Our approach, in order to obtain only the tweets written by Italian users, consider the user location present in the personal profile excluding the location attached in the singular tweet. At this point, from 31.776 users, only 25.315 profiles contained the geographical position within their Twitter account, and after the removal of non-significant and non-existent locations a total of 18.540 profiles have significant data. After this first analysis, the resulting tweets with geodata number 37.430. Once we obtain a dataset composed by tweet and geographical area of reference, we use the Latent Dirichlet allocation (LDA), analysing the whole corpus of tweets in order to obtain a frequency distribution useful to calculate the discussion rate for each topic and for each Italian region. We can define Latent Dirichlet allocation [5] as a generative Bayesian model of a corpus, composed of three different hierarchical levels [6]. In this case, LDA is based on a Dirichlet distribution definable as the probability density with vector value having the same principle as the multinomial parameter θ with non-zero values, in particular in LDA terms are associated or related with a latent (or hidden) topic, resulting from the following joint probability:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

where $z_{d,n}$, is the equivalent of the topic assignment, related on the per-document topic distribution θ_d; and the term $w_{d.n}$ is dependent on all of the topics $\beta_{1:K}$ and the topic assignment $z_{d,n}$ [7].

Once we obtained the results of the LDA, we extracted the relative frequencies in order to create different cartograms of the specific topics, highlighting the spatial distributions using GIS Software (Q-GIS Development Team, Open Source Geospatial Foundation Project. http://qgis.osgeo.org.).

3. Results and discussion

The results fully answered the research questions, it was possible, with the adoption of this process, to obtain an increased amount of geo-spatial data, useful to better understand the public view on climate change in Italian areas. With this process, we obtained 37.430 pure tweets, by extracting the users' profile location and linking

to the corresponding tweet, otherwise, considering only the position attached in each single tweet, we would have obtained only 1827 pure tweets (6% of the total corpus).

From the LDA, in Table 1 we represented the document-topic probabilities highlighting 10 topics with 5 keywords obtained with two validation sets of a topic coherence metric, where $\alpha = 0.1$ and $\beta = 0.1$. Overall, in the 10 topics we can find three principal themes of discussion: a) climate change effects; b) perception and uncertainty; c) government policies. In the first theme, some topics (e.g., topic 1, topic 4, topic 8, and topic 9) focus on the climate change effects on the society, reporting problems related to the weather, environmental sustainability, and animal extinction. In the second theme, it is possible to analyse the fear, concern, and anxiety that climate change is creating in society (e.g., topic 2), while also a conspiracy "cluster" emerges in topic 7.

Table 1: *Topic modelling results ordered by the keywords weight*

Topic	***Keywords***
Topic1	inadeguatezza (0.049), vedere (0.025), estate (0.025), caldissime (0.025), boschi (0.025)
Topic2	preoccupata (0.037), clima (0.037), urgente (0.025), stop (0.025), ansia (0.015)
Topic3	sviluppo (0.018), popolare (0.018), futuro (0.018), crisi (0.014), rapido (0.010)
Topic4	mitigazione (0.039), glasgow (0.033), gradi (0.025), caldo (0.015), torrido (0.010)
Topic5	convincere (0.032), gradi (0.032), pensiamo (0.020), estinzione (0.011), negano (0.009)
Topic6	marina (0.030), animali, (0.030), emmanuelmacron (0.025), eu (0.017), commissione (0.010)
Topic7	tv (0.039), mente (0.020), terrorismo (0.17), allarme (0.015), record (0.010)
Topic8	sostenibilità (0.081), ristorazione (0.042), ambiente (0.032), collasso (0.015), svizzero (0.007)
Topic9	morte (0.049), specie (0.037), panda (0.020), koala (0.012), nonostante (0.002)
Topic10	incapacità (0.032), fare (0.0015), possibilità (0.010), governo (0.003), america (0.001)

In the third theme, the public view addresses the climate change crisis from a future perspective (topic 3), discussing action and policies in Europe and the United States (e.g., topics 6 and 10). Once we obtained the relevant topics and the keywords, we constructed a cartogram for each of them (as represented in figure 1.) reported below as relative frequencies. In topic 1, two Italian regions, Lombardia (0.39) and Lazio (0.37) particularly emerged, discussing the effects of global warming on planet earth. In topic 2, the anxiety and uncertainty about climate change effects are discussed more in Lazio (0.27) and Lombardia (0.19) region, followed by Emilia Romagna (0.12) and Molise (0.11). The third topic, Lazio (0.15), Emilia-Romagna (0.12), Veneto (0.12), and Toscana (0.11) regions, focused more on the crisis and the future development of climate change, citing the rapid growth of the world population. In topic 4, Abruzzo (0.15), Sicilia (0.15), Puglia (0.12), and Umbria (0.12) users,

talked more about the high temperature and the possible mitigation of it, also citing the United Nations Climate Change Conference (COP26) held in Glasgow in 2021. In topic 5, emerges issues relating to the human extinction due to the global warming problem in Liguria (0.015), Sicilia (0.13), Puglia (0.12), and Toscana (0.12) regions.

The sixth topic focuses on the government policies, Lazio (0.15), Trentino-Alto Adige (0.15), and Piemonte (0.12) users discussed it, citing, for example, the EU commission. In topic 7, a conspiracy topic emerged, discussed more by Lazio (0.27) and Puglia (0.14) region, highlighting the unjustified alarmism by the media about climate change. In topic 8, Emilia Romagna (0.17), Friuli Venezia Giulia (0.12), and Liguria (0.15) discussed the sustainability of cuisine in an environmentally sustainable food system.

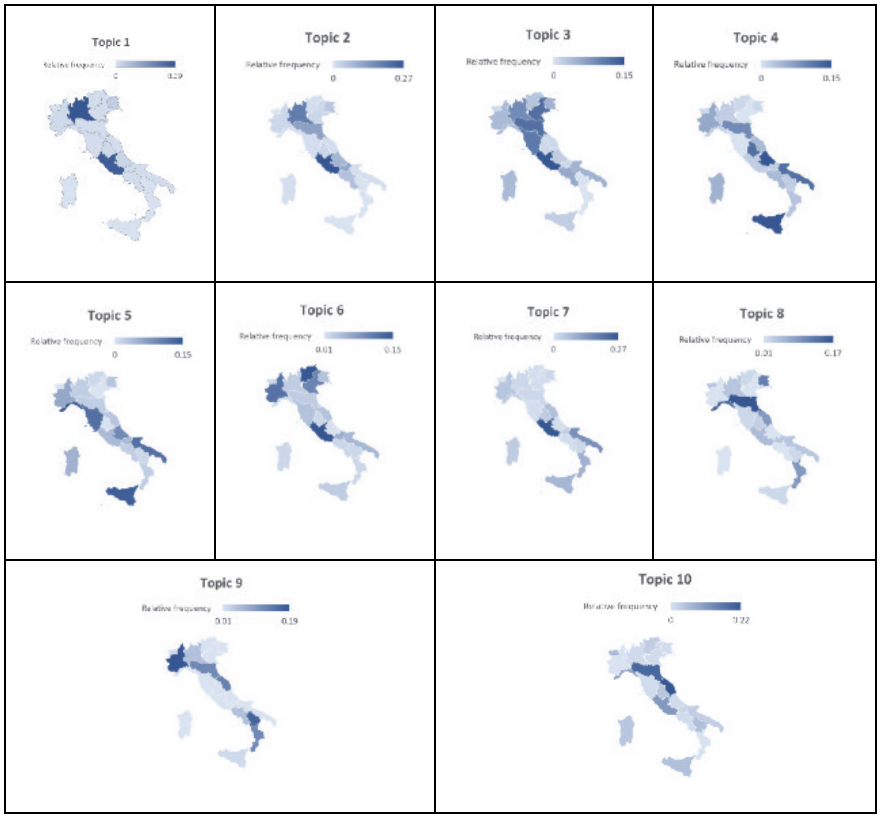


Figure 1: *Geo-spatial topic analysis outputs*

In topic 9, particular attention to the animal extinction (see the keyword "koala" and "panda") emerged, specifically in Piemonte (0.19), Basilicata (0.17), Marche (0.13), Emilia (0.12), and Calabria (0.12) regions. Finally, in the last topic (Topic 10),

users from Marche (0.22), Emilia-Romagna (0.19), and Lazio (0.11) discussed more the action and the policies of the American government.

4. Conclusions and future works

In conclusion, the approach provided a new implementation for improving geo-spatial topic modeling, increasing the availability of geodata compared to the standard approach. However, this study proposed a preliminary approach that in future works could be improved. The correlation between the Italian language and tweets that were written by an Italian user, does not allow us to say a priori that they are always related, in fact, sometimes users may enter a misleading (or old) location in their profile.

In these terms, this contribution is the first step and future works could be implemented using more keywords with larger datasets of Twitter data. Furthermore, it will be possible to implement different text-mining techniques to understand the sentiment of the discussions and analyse the dataset periodically (e.g., what were the most discussed topics from 2007 to 2011? Or, in which areas was a particular topic most discussed in 2020?).

To have a more complete vision of the object of study, different spatial analyses (e.g., spatial autocorrelation analysis) can be taken into consideration, we did not apply them to this study due to low contiguity of the regions.

Acknowledgment The work carried out in this paper was supported by the project SCORE which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003534.

References

1. PCC, 2014: Climate Change: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 (2014)
2. Panu P. Anxiety and the ecological crisis: An analysis of eco-anxiety and climate anxiety. Sustainability. **12**(19), 7836 (2019).
3. Coffey, Y., Bhullar, N., Durkin, J., Islam, M. S., & Usher, K. Understanding eco-anxiety: a systematic scoping review of current literature and identified knowledge gaps. The Journal of Climate Change and Health, **3**, 100047 (2021).
4. Bird, S., Klein, E., & Loper, E.. Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc." (2009).
5. Blei, D. M. Probabilistic topic models. Communications of the ACM, **55**(4), 77–84 (2012) doi: https://doi. org/10.1145/2133806.2133826.
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. Journal of machine Learning research, 993-1022 (2003).
7. Tong, Z., & Zhang, H. A text mining research based on LDA topic modelling. In International Conference on Computer Science, Engineering and Information Technology 201-210 (2016).

Comparing local structures of spatio-temporal point processes on linear networks

Confronto fra strutture locali di processi di punto spazio-temporali su reti lineari

Nicoletta D'Angelo, Giada Adelfio, and Jorge Mateu

Abstract We employ the Local Indicators of Spatio-Temporal Association (LISTA) functions on linear networks to build a statistical test for local second-order structure. This allows to identify differences in the spatio-temporal clustering behaviour of two point patterns, a point pattern of interest and a background one, both occurring on the same linear network. We illustrate the proposed methodology analysing a traffic-related problem.

Abstract *Utilizziamo le funzioni LISTA su reti lineari per costruire un test statistico per la struttura locale del secondo ordine. Ciò consente di identificare le differenze nel comportamento di clustering spazio-temporale di due pattern di punti, uno di interesse e uno di background, entrambi sulla stessa rete lineare. Illustriamo la metodologia proposta analizzando un problema relativo al traffico.*

Key words: Hypothesis testing, Linear networks, Local Indicators of Spatio-Temporal Association, Local properties, Second-order characteristics, Spatio-temporal point patterns

1 Introduction

Point processes on linear networks are recently considered to analyse events occurring on particular network structures, as it is the case, for example, with traffic accidents. They were firstly introduced in the spatial context (see Baddeley et al. (2020) for a review) and then extended to the spatio-temporal case (Moradi and Mateu (2020); Cronie et al. (2020), to cite a few).

Nicoletta D'Angelo and Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy e-mail: nicoletta.dangelo@unipa.it; giada.adelfio@unipa.it

Jorge Mateu
Department of Mathematics, University Jaume I, Castellon, Spain e-mail: mateu@uji.es

While most of the literature about point processes on networks is concerned with non-parametric estimation of the spatio-temporal intensity and second-order analysis of point patterns occurring on linear networks, none of these approaches has considered local properties of the introduced second-order statistics to measure local first- or second-order structure.

In the Euclidean point process theory context, while the use of global spatio-temporal second-order summary statistics is a well-established practice to describe global interaction structures between points in a point pattern, the use of local tools was firstly advocated by Siino et al. (2018).

As local second-order characteristics provide information on how events relate to nearby events, in this paper, we aim at employing the LISTA functions on linear networks to build a local test to identify differences between the local spatio-temporal second-order structure of two patterns occurring on the same linear network.

The above-mentioned test is proposed in D'Angelo et al. (2021), where the authors provide simulation studies to assess the performance of the testing procedure of local second-order structures, considering in particular homogeneous, inhomogeneous, clustered spatio-temporal point processes to study how the test behaves under these different settings, as well as more complex spatio-temporal point processes, such as self-exiting processes, where each point of the pattern can potentially produce offspring. They found, as expected, that the power of the test increases as the number of points and the clustered structure of the point pattern of interest increase too. The test for local spatio-temporal second-order structure is implemented in the package `LISTAnet`[1] available on `GitHub`.

In this paper, we provide an application of the proposed methodology to a traffic-related dataset that is of particular importance.

Section 2 presents the definition of LISTA functions on linear networks. In Section 3 we recall the permutation test for assessing local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network. Section 4 contains the application to a traffic dataset. Conclusions come in Section 5.

2 LISTA functions on linear networks

Definition 1. Local Indicators of Spatio-Temporal Association
The Local Indicators of Spatio-Temporal Association (LISTA) functions are a set of functions that are individually associated with each of the points of the point pattern and can provide information about the local behaviour of the pattern (Siino et al., 2018).

In the Euclidean context, several second-order characteristics of point processes are considered for building LISTA functions, such as the K-function, the pair correlation function, and the product density function.

[1] https://github.com/nicolettadangelo/LISTAnet; R package version 0.1.0

D'Angelo et al. (2021) provide the operational definition based on the local second-order spatio-temporal summary statistics on a linear network.

Definition 2. LISTA functions on linear networks
Any local second-order spatio-temporal summary statistic $\lambda^i_{2,L}((\mathbf{u},t),(\mathbf{v},s))$ computed on an observed point pattern $\mathbf{x}$ on a linear network L, that satisfies the following operational definition: $\hat{\lambda}_{2,L}((\mathbf{u},t),(\mathbf{v},s)) = \frac{1}{n-1}\sum_{(\mathbf{u}_i,t_i)\in\mathbf{x}} \hat{\lambda}^i_{2,L}((\mathbf{u}_i,t_i),(\mathbf{v},s))$, where $\hat{\lambda}_{2,L}$ is an estimator of the global second-order intensity function, can be called a LISTA function on a linear network.

In this paper, we consider the local spatio-temporal K-function on networks, defined for each point $(\mathbf{u}_i,t_i)$ as

$$\hat{K}^i_L(r,h) = \frac{1}{\hat{\lambda}^2|L||T|}\sum_{(\mathbf{u}_i,t_i)\neq(\mathbf{v},s)} \frac{I\{d_L(\mathbf{u}_i,\mathbf{v})<r, |t_i-s|<h\}}{M((\mathbf{u}_i,t),d_L(\mathbf{u}_i,\mathbf{v}),|t_i-s|)}, \qquad r\geq 0, h\geq 0$$

obtained weighting for a constant intensity $\hat{\lambda} = n/(|L||T|)$, since our aim is to find out differences in the local structure of two point processes by comparing their second-order characteristics, whichever is the first-order intensity. Here $|L|>0$ and $|T|>0$ are the total length of the network L and of the time interval T, respectively, $M((\mathbf{u}_i,t_i),r,h)$ is the number of points lying exactly at the shortest-path distance $r\geq 0$ and the time distance $h\geq 0$ away from the point $(\mathbf{u}_i,t_i)$.

In this particular context, ignoring the network would result in bias when computing the K-functions, as without the network correction $M((\mathbf{u},t),r,h)$ (introduced by Moradi and Mateu (2020) for the spatio-temporal context) we would overestimate the number of points in any spatio-temporal lag considered.

3 Testing for local second-order structure on a network

We test the null hypothesis of no differences in the spatio-temporal second-order structure of the two observed point patterns $\mathbf{x}$ and $\mathbf{z}$ with respect to the i^{th} point $\mathbf{x}_i = (\mathbf{u}_i,t_i)\in\mathbf{x}$, with $i=1,\ldots,n$. Formally, for each point $(\mathbf{u},t)\in\mathbf{x}$, the hypothesis test reads as

$$\begin{cases} \mathscr{H}_0: & \text{no difference in the second-order local structure of } (\mathbf{u},t) \quad \text{w.r.t} \quad \{\{\mathbf{x}\setminus(\mathbf{u},t)\}\cup\mathbf{z}\} \\ \mathscr{H}_1: & \text{significant difference in the second-order local structure of } (\mathbf{u},t) \quad \text{w.r.t} \quad \{\{\mathbf{x}\setminus(\mathbf{u},t)\}\cup\mathbf{z}\} \end{cases}.$$

Rejecting $\mathscr{H}_0$ for most of the points in the point pattern $\mathbf{x}$ means that the local second-order structures of $\mathbf{x}$ and $\mathbf{z}$ around those arbitrary points $(\mathbf{u},t)$ are different, and therefore the second-order global structure of $\mathbf{x}$ and $\mathbf{z}$ is different, meaning that they are most likely coming from different underlying point processes.
Nevertheless, in addition to this global conclusion, the most interesting results concern the rejection of $\mathscr{H}_0$ for specific points, allowing for the identification of points for which the spatio-temporal local behaviour varies with the spatio-temporal loca-

tion. This could also result in the identification of areas where the spatio-temporal local second-order structure of the points in the pattern of interest **x** differs the most and needs to be further analysed. If, for instance, the significant points tend to cluster in time and space, this could indicate multiple underlying generating processes, in contrast to a unique one. Furthermore, the identification of a significant point may already be an indication of outlying behaviour for the remaining points.

3.1 The permutation test

To take into account the geometry of the underlying network, we develop a testing procedure based on the homogeneous LISTA functions reviewed in Section 2. A permutation test approach is considered, where the null distribution of a given statistical test is estimated by randomly permuting the labels attached to the two point patterns. The steps of the testing procedure are as follows

1. Define the point pattern of interest **x** and the background pattern **z**, assumed to be realisations of the point processes X and Z, with $N(X) = n$ and $N(Z) = m$. Typically, **x** is the point pattern with the most clustered structure;
2. Set k as the number of permutations;
3. For each point $(\mathbf{u}_i, t_i) \in \mathbf{x}$
   a. Estimate the LISTA function $\hat{K}^i_L(r,h)$;
   b. Compute the local deviation test $T^i = \int_0^{t_0}\int_0^{r_0}\left(\hat{K}^i_L(r,h) - \hat{K}^{-i}_{L,H_0}(r,h)\right)^2 \mathrm{d}r\mathrm{d}h$ where $\hat{K}^{-i}_{L,H_0}(r,h)$ is the LISTA function for the i^{th} point, averaged over the $j = 1,\ldots,k$ permutations, and where r_0 and t_0 are the maximum spatial and temporal ranges considered in the K-functions;
   c. A p-value is computed for each point of the pattern **x** as $p^i = \sum_{j=1}^{k} \frac{\mathbf{I}(T^{i,j}_{H_0} \geq T^i)}{k}$, where $T^{i,j}_{H_0} = \int_0^{t_0}\int_0^{r_0}\left(\hat{K}^{i,j}_L(r,h) - \hat{K}^{-i,j}_{L,H_0}(r,h)\right)^2 \mathrm{d}r\mathrm{d}h$, with $\hat{K}^{i,j}_L(r,h)$ being the LISTA functions on the k patterns generated from the null hypothesis, and $\hat{K}^{-i,j}_{L,H_0}(r,h)$ being the LISTA function averaged over the points of the k^{th} permutation.

The testing procedure ends with providing a vector p^i of $N(X) = n$ p-values. The null hypothesis is rejected if $p^i \leq \alpha$, where α is the fixed nominal value of the type I error.

We note that using a Bonferroni correction, too conservative results under the null hypothesis where obtained. So we are in such a case that the non-corrected test provides good significance levels while the corrected one is too conservative (see existing literature Siino et al. (2018); Tamayo-Uria et al. (2014) where no such correction was considered).

4 Medellin traffic data

We analyse traffic accidents in the city of Medellin (Colombia), a dataset containing the locations of traffic accidents consisting of 1215 traffic accidents in 2019 in downtown of Medellin. The entire data were published in the OpenData portal of Medellin Town Hall at https://www.medellin.gov.co/geomedellin/index.hyg. The dataset actually consists of a multitype point pattern, categorised based on the damages caused by the accident. In particular, 675 "damages only" events are recorded, and 536 "injuries to people".

In our analysis, we do not consider the four events representing car accidents that lead to deaths, and we define the accidents leading to damages as the point pattern of interest $\mathbf{x}$, and the accidents causing injuries to people as the alternative (background) pattern $\mathbf{z}$.

We thus test the hypothesis of difference in the spatio-temporal local second-order structure between the two point patterns considered, setting the number of permutations to $k = 99$. The application of the test provides a p-value for each point in $\mathbf{x}$, and by comparing those to the confidence level $\alpha = 0.05$ we can identify *significant* events, i.e. those points for which $\mathscr{H}_0$ is rejected and therefore the local second-order structure is significantly different from the background process, and *non-significant* events. The significant events are 100 out of the 675 events, and they are shown in Figure 1.

Fig. 1: *Left panel*: Significant events. *Right panel*: Non-significant events.

The application results of the local test support the hypothesis that the two point patterns behave differently in terms of spatio-temporal local second-order structure.

Furthermore, displaying the significant events and the non-significant events allows identifying the areas containing those events whose spatio-temporal local second-order structure is different from the background pattern.

For instance, the results of the proposed application suggest that damages tend to cluster more near road intersections.

5 Conclusions

In this work, we have presented a test to assess local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network, based on the Local Indicators of Spatio-Temporal Association on linear networks, and we have applied it to a case study, analysing traffic data represented by a marked point process occurring on the streets of Medellin, where the marks represent the type of car accident. We have been able to explore how individual points are related to their neighbouring events and classify points with similar spatio-temporal local structures, identifying the areas that differ in terms of clustering behaviour. Furthermore, displaying the significant events has allowed to identify the areas containing those events whose spatio-temporal local second-order structure is different from the background pattern.

The analysed point patterns are basically identified by characteristics of the events, i.e. the marks, recorded together with their spatial and temporal location. Future analyses could regard the application of the local test to patterns whose splitting is due to a statistical procedure, such as the stochastic declustering of the self-exciting point processes. Indeed, the topic of fitting parametric point process model, accounting also for the underlying network structure, is still quite unexplored, and only a recent attempt has been carried out to fit self-exciting models to processes occurring on a network (D'Angelo et al., 2022).

References

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., and Davies, T. M. (2020). Analysing point patterns on networks—a review. *Spatial Statistics*, 42:100435. doi.org/10.1016/j.spasta.2020.100435.

Cronie, O., Moradi, M., and Mateu, J. (2020). Inhomogeneous higher-order summary statistics for point processes on linear networks. *Statistics and computing*, 30:1221–1239.

D'Angelo, N., Adelfio, G., and Mateu, J. (2021). Assessing local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network. *Spatial Statistics*, 45:100534.

D'Angelo, N., Payares, D., Adelfio, G., and Mateu, J. (2022). Self-exciting point process modelling of crimes on linear networks. *Statistical Modelling. Accepted*.

Moradi, M. M. and Mateu, J. (2020). First- and second-order characteristics of spatio-temporal point processes on linear networks. *Journal of Computational and Graphical Statistics*, 29(3):432–443.

Siino, M., Rodríguez-Cortés, F. J., Mateu, J., and Adelfio, G. (2018). Testing for local structure in spatiotemporal point pattern data. *Environmetrics*, 29(5-6):e2463.

Tamayo-Uria, I., Mateu, J., and Diggle, P. J. (2014). Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spatial Statistics*, 9:192–206.

DISTATIS-based spatio-temporal clustering approach: an application to business cycles' time series

Un approccio di clustering spazio-temporale basato sull'algoritmo DISTATIS: un'applicazione alle serie storiche relative ai cicli economici

Raffaele Mattera and Germana Scepi

Abstract In this paper we propose a clustering approach based on the DISTATIS algorithm for the study of time series showing spatial spillover effects. This clustering approach involves the use of a spatio-temporal distance, obtained from the combination of temporal and spatial distances. An application to real time series, related to Euro Area business cycles, enriches this paper.

Abstract *In questo articolo proponiamo un approccio di clustering basato sull'algoritmo DISTATIS per lo studio di serie temporali che mostrano effetti di spillover spaziale. Questo approccio di clustering prevede l'uso di una distanza spazio-temporale, ottenuta dalla combinazione di distanze temporali e spaziali. Un'applicazione alle serie storiche reali, relative ai cicli economici dell'Area Euro, arricchisce questo lavoro.*

Key words: time series analysis, spatial association, business cycle, cluster analysis, LISA, DISTATIS

1 Introduction

The presence of spatial spillover effects is well acknowledged when we deal with macroeconomic time series. An interesting example is given by the interaction among business cycles, since macroeconomic fluctuations are affected by spillover effects generated by neighbouring countries (e.g. see Hájek and Horváth, 2016).

Raffaele Mattera
Department of Economics and Statistics, University of Naples "Federico II"
Department of Social and Economic Sciences, Sapienza University of Rome e-mail: raffaele.mattera@unina.it

Germana Entropy
Department of Economics and Statistics, University of Naples "Federico II" e-mail: Entropy@unina.it

With the introduction of the Euro, different countries are now characterized by the same monetary policies. However, the effect of monetary policies can be asymmetric if there is not a business cycles' syncronization. Economists argue that, in theory, the monetary union enhances the business cycle' syncronization of the countries involved (Artis and Zhang, 2002). However, there are a lot of studies documenting the absence of a common business cycle in the Euro Area (e.g. Ahlborn and Wortmann, 2018; Artis and Zhang, 2002). Most of these studies use cluster analysis to group countries together on the basis of different business cycles. They find evidence of the so-called core–periphery division, where a "core" group (typically identified with the founding EU Member countries) is opposed to another group, called "peripherical" that typically contains all the other countries.
The definition of real clusters based on business cycles is a very relevant problem for public policy. Previous studies provide very different clusters' composition because they use different businsess cycles' estimating methods, clustering algorithms and dissimilarities measures (Artis and Zhang, 2002; Ahlborn and Wortmann, 2018; Aguiar-Conraria and Soares, 2011).
Surprisingly, the relevant spatial nature of the phenomenon has not been considered by previous studies yet. We fill this gap by introducing spatial dimension in the clustering procedure. In particular, we propose a clustering approach based on the use of a spatio-temporal distance, obtained from the combination of temporal and spatial distances. Optimal weights are computed by the DISTATIS algorithm (Abdi et al., 2005).
The rest of the paper is structured as follows. Section 2 discusses the proposed spatio-temporal clustering algorithm, while section 3 provides an application to Euro Area business cycles' time series. Some final remarks can be found at the end.

2 Methodology

Assuming we observe the values of a variable $C_{i,t}$ for a set of $N(i = 1,\ldots,N)$ spatial units over $T(t = 1,\ldots,T)$ time. In the business cycle case, $C_{i,t}$ represents the cyclical fluctuation of an i-th nation over the time T. When we deal with spatial units, it is common to study the presence of a spatial association that indicates if there is a similarity in the values among contiguous units. The existence of a spatial association among neighbouring locations can be assessed either globally or locally. While global measures define the overall presence of spatial correlation in a dataset, local measures can be used to identify the degree of similarity of each i-th unit to its neighbours. Local Indicators of Spatial Association (LISA) can be used to study the spillover effects (e.g. see Ying, 2000). Note that classical LISA are defined as static measures, so that they show a unique value for each time t.
Clustering of spatial data can be developed by considering both models based on spatial dissimilarity measures (e.g. based on LISA as in Scrucca, 2005) or spatially constrained models (Romary et al., 2015). When the variables are observed at the

same time on space and time, it is possible to use both non spatial models with spatio-temporal distance (Izakian et al., 2012) or spatially constrained models with temporal distance (D'Urso et al., 2021).
We follow a not spatial clustering approach (Parition Around Medoids, see Kaufman and Rousseeuw, 1990) with the following spatio-temporal distance between two units i and j:

$$D_{i,j}(s,t) = \lambda_s D_{i,j}(s) + \lambda_t D_{i,j}(t) \tag{1}$$

where $D_{i,j}(s,t)$ is the distance over space s and time t, $D_{i,j}(s)$ and $D_{i,j}(t)$ are, respectively, the spatial and temporal distances with λ_s and λ_t the spatial and temporal weights, such that $\lambda_s + \lambda_t = 1$. The spatial distance measures the similarity between two units in terms of spillover effects. It is an Euclidean distance computed on the standardized Getis and Ord (1992) index. $D_{i,j}(s)$ is an *instantaneous* spatial distance calculated on the last observation $t = T$. $D_{i,j}(t) = \sqrt{2(1-\rho_{i,j})}$, is a correlation-based distance, where $\rho_{i,j}$ is the correlation coefficient between two business cycles' time series $C_{i,t}$ and $C_{j,t}$.
A crucial step involves the computation of the weights λ_s and λ_t. For this purpose, we propose to adopt the DISTATIS algorithm of Abdi et al. (2005). The DISTATIS algorithm aims to analyze and synthetize a set of distance matrices. It is a generalization of the classical multidimensional scaling (MDS). The DISTATIS algorithm computes the weights as the eigenvalues of the following data matrix cross-product:

$$\mathbf{X}^T\mathbf{X} = \left[vec\left\{\mathbf{S}(s)\right\}, vec\left\{\mathbf{S}(t)\right\}\right]^T \left[vec\left\{\mathbf{S}(s)\right\}, vec\left\{\mathbf{S}(t)\right\}\right] \tag{2}$$

where $vec(\cdot)$ is the vectorization operator, $\mathbf{S}(s) = \Xi D_{i,j}(s)\Xi^T$ and $\mathbf{S}(t) = \Xi D_{i,j}(t)\Xi^T$, given $\Xi = I_N - 1_N m^T$ with I_N the Identity matrix, m a vector of length N containing equal elements $m_i = 1/N$ and 1_N a vector of length N containing values equal to 1.
We adopt a Partition Around Medoids (PAM) clustering algorithm. Therefore, the final clustering problem takes the following form:

$$\min: \quad \sum_{i=1}^{N}\sum_{k=1}^{K} D_{i,k}(s,t)^2 \tag{3}$$

where $D_{i,k}(s,t)$ is equal to (1) for the k-th centroid. Therefore, the PAM clustering algorithm finds the optimal partition of the dataset into C clusters with the aim of minimizing the distances of the units from their centroid.

3 Application to Euro Area business cycles' time series

We consider the quarterly GDP data of the Euro Area countries from the Eurostat website[1], observed from 1995 to 2021. We estimate the business cycle for each country by means of the Hodrick-Prescott (HP) filter (see Fig. 1).

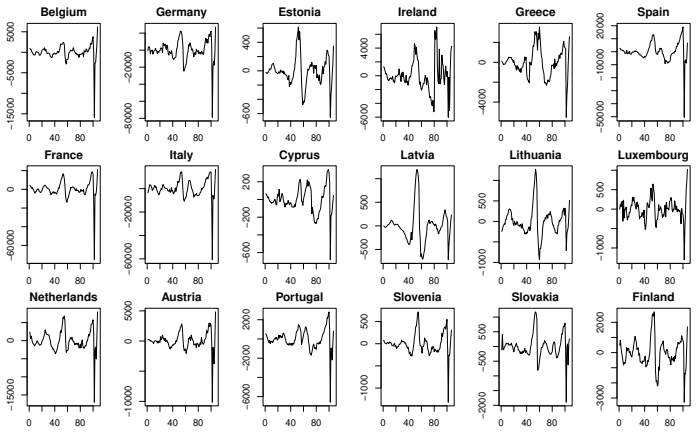


Fig. 1 Business cycles estimates obtained with the Hodrick-Prescott (HP) filter

The standardized Getis and Ord (1992) indices are reported in Tab. 1.

Table 1 Standardized Getis and Ord (1992) index by countries

Belgium	Germany	Estonia	Ireland	Greece	Spain	France	Italy	Cyprus
2.4741	3.283	-1.1964	0.4221	-1.2099	-0.5814	1.6503	0.8014	-0.0734
Latvia	Lithuania	Luxembourg	Netherlands	Austria	Portugal	Slovenia	Slovakia	Finland
-0.4225	-1.216	1.8447	1.4647	-0.4771	-1.0429	1.1013	1.3716	-0.0735

Positive values in Tab. 1 indicate that around the i-th location there is positive spillover, while negative values indicate negative spillover. From Tab. 1 we can notice that most of "economic virtous" countries are associated with positive values (e.g. Belgium, Germany, France), while most of PIIGS economies (e.g. Greece, Spain) show negative values. The clustering results, obtained with the proposed spatio-temporal approach, are shown in Fig. 2. The optimal number of clusters is selected by maximizing the Average Silhouette Width (ASW). Firstly, we note that, although the temporal dimension has a higher weight than the spatial one ($\lambda_t = 0.7329, \lambda_s = 0.2670$) in the definition of the distance, the spatial component is quite relevant in the clusters' composition: Portugal and Spain are clustered together, as well as Finalnd and Baltic countries (yellow cluster), while the Benelux

[1] Data can be retrieved at the following link https://ec.europa.eu/eurostat/web/main/data/database

countries (Netherlands, Belgium and Luxemburg) are clustered with their neighbours France, Germany and Italy. Overall, we find evidence favoring the "core-periphery" patterns in the business cycle synchronization, but the final clusters' composition differs from that showed by previous studies. For example, Ahlborn and Wortmann (2018) showed three gropus, whereas we find only two. Moreover, we find that Spain and Portugal belong to "periphery" cluster rather than the "core" one. Moreover, Ahlborn and Wortmann (2018) considers Ireland as a fuzzy unit, while our clustering group Ireland in the "periphery" cluster with a low degree of fuzziness. Such differences can be explained by the relevance of spatial spillovers.

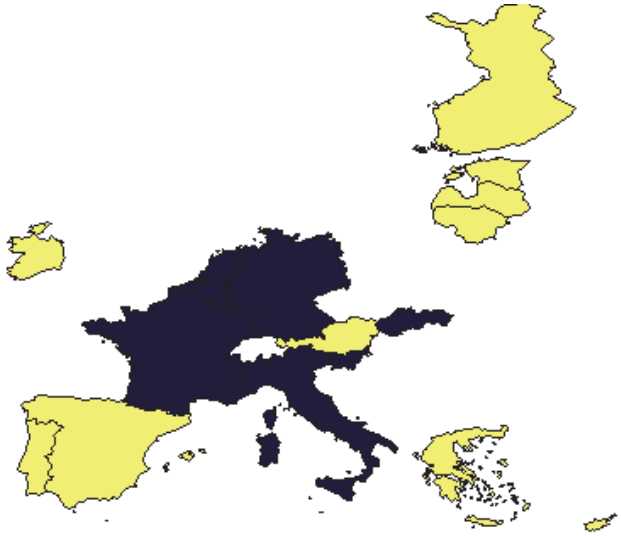


Fig. 2 Clustering results: cluster 1 is with blue color, cluster 2 with yellow one.

4 Final remarks

In this paper we propose a novel approach for clustering spatio-temporal data based on the use of a novel spatio-temporal distance. The spatial component of this distance is defined by means of Local Indicators of Spatial Association (LISA), while temporal one by the correlation among the time series. The two components of the distance are weighted by the DISTATIS algorithm's estimated weights. We notice that DISTATIS is applied, for the first time, in the definition of a spatio-temporal distance.
We apply the proposed clustering approach in the relevant problem of business cycles' syncronization of Euro Area countries.
A future study will be devoted to the comparison of the proposed clustering ap-

proach with other established techniques for spatio-temporal data. Moreover, we aim to evaluate the robustness of the results employing both alternative filters for business cycle estimation and different approaches for measuring spatial and temporal distances.

References

Abdi, H., O'Toole, A. J., Valentin, D., and Edelman, B. (2005). Distatis: The analysis of multiple distance matrices. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 42–42. IEEE.

Aguiar-Conraria, L. and Soares, M. J. (2011). Business cycle synchronization and the euro: A wavelet analysis. *Journal of Macroeconomics*, 33(3):477–489.

Ahlborn, M. and Wortmann, M. (2018). The core–periphery pattern of european business cycles: A fuzzy clustering approach. *Journal of Macroeconomics*, 55:12–27.

Artis, M. J. and Zhang, W. (2002). Membership of emu: A fuzzy clustering analysis of alternative criteria. *Journal of economic integration*, pages 54–79.

D'Urso, P., De Giovanni, L., and Vitale, V. (2021). Spatial robust fuzzy clustering of covid 19 time series based on b-splines. *Spatial Statistics*, page 100518.

Getis, A. and Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3).

Hájek, J. and Horváth, R. (2016). The spillover effect of euro area on central and southeastern european economies: a global var approach. *Open Economies Review*, 27(2):359–385.

Izakian, H., Pedrycz, W., and Jamal, I. (2012). Clustering spatiotemporal data: An augmented fuzzy c-means. *IEEE transactions on fuzzy systems*, 21(5):855–868.

Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125.

Romary, T., Ors, F., Rivoirard, J., and Deraisme, J. (2015). Unsupervised classification of multivariate geostatistical data: Two algorithms. *Computers & geosciences*, 85:96–103.

Scrucca, L. (2005). Clustering multivariate spatial data based on local measures of spatial autocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica*, 20(1):11.

Ying, L. G. (2000). Measuring the spillover effects: Some chinese evidence. *Papers in regional science*, 79(1):75–89.

Developments in composite indicators

Bayesian networks for monitoring the gender gap

Le reti bayesiane come proposta per monitorare il gender gap

Flaminia Musella and Lorenzo Giammei and Silvana Romio and Fulvia Mecatti and Paola Vicard

Abstract Composite indicators are a common choice for synthesizing complex phenomena. Over the years, they have grown in popularity and are now applied in many social and environmental sciences. Among others, a subject of increasing interest is gender equality analysis. Gender composite indicators, even if easy to read, may provide a limited picture of the problem. Here we discuss the potentiality to integrate the use of composite indicators for gender gaps with Bayesian networks, powerful tools for explaining the complex association structure in the dataset and developing scenarios to orient policy-making. An example is carried out on Italian province-level data.

Abstract *Gli indicatori compositi rappresentano una scelta comune per sintetizzare fenomeni complessi; la loro popolarità e la vastità di ambiti in cui sono applicati ha raggiunto anche l'analisi della parità di genere. Seppur gli indicatori compositi di genere siano di facile lettura, essi rischiano di fornire un quadro limitato al problema. Nel presente paper discutiamo la possibilità di integrare l'uso di indicatori compositi per il gender gap con un modello statistico-probabilistico basato sulle reti bayesiane e sviluppato a partire da un set di dati raccolto a livello provinciale.*

Key words: Composite indicator, Gender statistics, Multivariate dependencies

---

Flaminia Musella
Link Campus Unversity, Casale di San Pio V 44 Rome, e-mail: f.musella@unilink.it

Lorenzo Giammei
Sapienza University, Piazzale Aldo Moro 5 Rome, e-mail: lorenzo.giammei@uniroma1.it

Silvana Romio
University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1 Milan, e-mail: silvana.romio@unimib.it

Fulvia Mecatti
University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1 Milan, e-mail: fulvia.mecatti@unimib.it

Paola Vicard
Roma Tre University, Via Silvio d'Amico 77 Rome, e-mail: paola.vicard@uniroma3.it

1 Introduction

As revealed by United Nations Statistics Division – UNSD Global Gender Statistics Programme, gender statistics stands upon data collection disaggregated by sex with the aim of detecting and untangling gender bias in the analysis of real phenomena. A critical examination of the potential effect of unbalanced opportunities and denied rights can greatly benefit from being done according to statistical reasoning to point out the tangled relations among predictors of the gender gaps. Furthermore, gender statistic is motivated by the growing demand for gender-sensitive statistical information, coming from society and made even more urgent by the Covid-19 planetary crisis and its consequences that go well beyond public health. Finally, the increasing interest in gender equality is justified by the Goal 5 of UN Agenda 2030, which aims to achieve a sustainable gender balance for the benefit of society as a whole, both men and women.

The most popular strategy for gender data analysis is based on building composite indicators. They are easy to implement and interpret but unable to fully catch the multivariate and inter-sectional nature of gender inequalities. In this paper the potentiality of complementing composite indicators with an approach based on Bayesian networks (BNs) is discussed. BNs are multivariate statistical models for managing the intricate net of dependencies among variables. Applied to gender statistic, BNs can help describing the dependence of men-women distance from a set of observed variables, among which those used for the composite indicators. Hence BNs become a powerful tool for developing scenarios of gender balance and, together with composite indicators, could support decision-makers in orienting policies activated in different fields (*i.e.* economic, social and health) and simultaneously impacting the gender gap.

In this paper we firstly introduce composite indicators (Section 2) by focusing on the European Gender Index (GEI henceforth); in Section 3 we both review Bayesian Networks and present an illustrative example based on real Italian data collected at province-level of disaggregation; conclusions are addressed in Section 4.

2 Composite Indicators and Gender Equality Index

Composite indicators have grown in popularity due to their ease of reading and effectiveness in comparative analyses [6]. Their implementation is common in many disciplines, ranging from social to environmental and governance applications [1, 2]. Global institutions, among which the OECD, World Bank and EU, are also increasingly adopting synthetic indices for supporting policy makers around the globe [7]. Nevertheless, scientific debate about composite indexes is divided between supporters that highlight their effectiveness in describing an overall complex phenomenon and opponents that instead claim that results may be statistically limited and sometimes misleading [8]. Here we focus on composite indicators for monitoring the gender gap. These have been developed since 1995, trying to catch

different aspects of the gender gap. Some of them aim at measuring gender equality (positive gap) and some others gender inequalities (negative gap). Table 1 summarizes the evolution of the main gender indexes as they have appeared over time.

Table 1 Gender indexes

INDEX	Name	Institute	B-Year
GDI	Gender Development Index	UN Development Programme	1995
GEM	Gender Empowerment Index	UN Development Programme	1995
GEI	Gender Equality Index	EIGE	2000
GGGI	Global Gender Gap Index	World Economic Forum	2006
SIGI	Social Institutions and Gender Index	OECD	2009
GII	Gender Inequality Index	UN Development Programme	2010
WEOI	Women's Economic Opportunity Index	Economist Intelligence Unit	2010
MGII	Multidimensional Gender Inequality Index	University	2014
BIGI	Basic Indicator of Gender Inequaltiy	University	2019

Composite gender indexes are built accounting for several aspects, called domains, related to different areas of life and sometimes sliced into sub-domains. Within these areas, each index catches a specific set of variables that are meant to indirectly measure gap/equality between genders. These variables stand for ingredients of the obtained synthetic indicator. There is a large debate on how to combine the contribution of each variable to the domain since the weighting strategies may be seen simultaneously as opportunities and limitations [5]. Consistently with the most popular choice in Europe, we refer to GEI as a gold standard composite index to be taken in account within our proposal. One of the main objectives of the European Institute for Gender Equality (EIGE) is promoting equality between women and men in European Union. To this aim, since 2010, the institution annually constructs and publishes the GEI index to monitor and compare countries with respect to their level of gender equality/inequality and efforts to reduce gaps. The composite indicator ranges between 0, standing for total inequality, and 100, meaning perfect equality; it is based on six core domains, in some sub-domains, and 27 (univariate marginal) indicators. The index construction consists of the following steps: (i) firstly, indicators are measured by gender and the differences between women and men are computed, (ii) variables are aggregated in sub-domains by an unweighted arithmetic mean; (iii) sub-domains are combined in domains by a geometric mean, still unweighted; (iv) finally, domains are aggregated by means of a weighted geometric mean with weights established by experts. GEI domains are related to the following areas: Work, Money, Knowledge, Time, Power and Health. These domains remain the driver dimensions in our proposal based on Bayesian networks as well.

3 The Bayesian Network version of GEI

In this section we introduce Bayesian Networks [9] and show their use when investigating the gender gap. We begin by briefly describing the basic terminology of graphs and the main elements of a BN. Then we propose an implementation of the method on real province-level gender data.

A graph $G(V,E)$ consists of a set of nodes (or vertices) V and a set of edges E. Two nodes connected by an edge are said to be adjacent. Edges can be directed (arrows) if they are oriented from a node X_i to another node X_j, or undirected (lines) if such orientation is not made explicit with arrow direction. If a graph only contains directed edges, it is called a directed graph. When a directed edge goes from a node X_i to a node X_j, X_i is called a parent of X_j. We will denote with pa_i the set of the parents of node X_i. A path is a sequence of edges traced along pairs of adjacent nodes, such that each edge starts with the vertex ending the preceding edge. A path from X_i to X_j, where all the edges are oriented along the direction of the path, is a directed path. If a directed path begins and ends with the same node, it is called a cycle. A directed acyclic graph (DAG) is a directed graph where cycles are forbidden.

Bayesian Networks are composed by a DAG and a joint probability distribution over its nodes. The vertices of the DAG represent random variables and the edges describe the relations between them. Given a DAG G with node set $\mathbf{V} = \{X_1, ..., X_n\}$, the joint probability distribution $P(V)$ over its nodes can be factorized according to the structure of the graph as follows

$$P(x_1, ..., x_n) = \prod_i P(X_i | pa_i). \tag{1}$$

A BN is here proposed to show how this methodology could contribute to the study of the gender gap and support the analysis conducted with composite indicators. A province-level dataset is used to estimate the graph through a process called structural learning [3]. When the relationships within variables, i.e. associations and dependencies are unknown, structural learning allows to discover the relations from data and encode them into a DAG. The BN is learnt through the PC algorithm [4], a procedure that performs multiple conditional independence tests on data and then translates the obtained independence statements into a DAG. The variables contained in the dataset are listed in Table 2 and cover GEI ingredients, as well as additional structural and socioeconomic features. The logical groups in Table 2 also serve as a guide to learn the BN. It is assumed that structural and socioeconomic features are not affected by GEI ingredients and that structural features are not affected by socioeconomic features. These constraints are translated into graphical form by the algorithm by forcing the corresponding edge directions to be absent.

The obtained DAG is shown in Figure 1. Province-level GEI has been inserted in the network as an additional node, with incoming arcs from each ingredient variable. The presence of a GEI node allows ranking provinces in terms of gender equality. The colour of the vertices denotes the variables group (blue= structural features, orange= socioeconomic features, green= GEI ingredients, white= GEI at province-

Table 2 Variables logical groups

Structural features	Socioeconomic features	GEI Ingredients
Internet access	Province added value	Employment rate gender gap
Firm size	Total employment rate	Female work segregation
Population density	Education rate	Income gender gap
Geographical area		Female education segregation
		Female education rate
		Life expectancy gender gap
		Female entrepreneurs ratio
		Average n. of children

level). The graph provides a clear picture of how all the variables interact. The structure shows those variables affecting the GEI ingredients and those on which a policy maker could intervene to increase gender parity. The relations encoded in the graph structure are consistent with subject matter knowledge emerging from the literature.

4 Discussion and Research Perspectives

Gender parity has become a central objective in policy-making processes and allocation of resources. However, reducing inbalances between genders requires a deep understanding of the the risk factors and the interactions therein, which cannot be achieved without mainstreaming of gender research to improve quality of data and

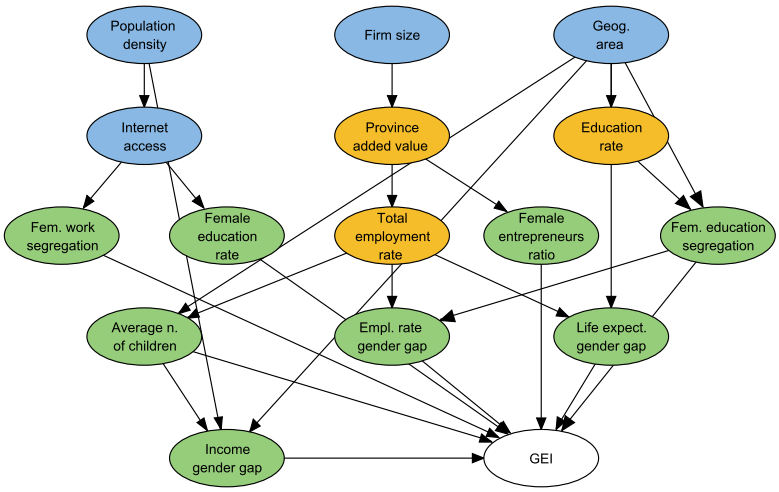


Fig. 1 Bayesian network model including GEI

statistical tools. Constructing composite indicators is the most common statistical approach to investigate inequalities in opportunities and rights based on gender. This methodology is usually straightforward to implement and provides a concise measure of the gender gap, but also presents some limits. Here we proposed a statistical approach based on Bayesian Networks which could be used together with composite indices. This advanced statistical model, based on structural learning, translates the information contained in the data into an effective easy-to-read graphical form, thus revealing how variables interact in a fully multivariate environment. This allows direct identification of the determinants of the gender gap, thus providing a solid data-driven base to support and motivate policies. Implementing BNs together with composite indicators would thus mean complementing the compact descriptive information provided by the latter with a more prediction-oriented multivariate model.

The proposed implementation of BNs on province-level data constitutes an illustrative application of the methodology. The obtained DAG reveals meaningful relations between variables that are consistent with the subject matter knowledge described in the related literature. The network could be thus interrogated to simulate predictive scenarios and explore how the DAG can be used as a decision support system. Further analysis will be carried out to investigate the full potential of this approach.

References

1. Bandura, R. Measuring country performance and state behavior: A survey of composite indices. Technical report, Office of Development Studies, United Nations Development Programme (UNDP), New York (2005).
2. Bandura, R. (2011). Composite indicators and rankings: Inventory 2011. Technical report, Office of Development Studies, United Nations Development Programme (UNDP), New York (2011).
3. Daly, R., Shen, Q., Aitken, S. Learning Bayesian Networks: Approaches and Issues. The knowledge engineering review, 26 (2), 99–157 (2011).
4. Glymour, C.; Spirtes, P.; Scheines, R. Causal Inference. Erkenntnis, 35 (1–3), 151–189 (1991).
5. Greco, S., Ishizaka, A., Tasiou, M. et al. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Soc Indic Res 141, 61–94 (2019). https://doi.org/10.1007/s11205-017-1832-9
6. Joint Research Centre-European Commission. *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing (2008).
7. Saltelli, A. Composite indicators between analysis and advocacy. Social Indicators Research, 81(1), 65–77 (2007).
8. Sharpe, A. Literature review of frameworks for macro-indicators. Ottawa: Centre for the Study of Living Standards (2004).
9. Pearl, Judea. Models, reasoning and inference (2000).

An Alternative Aggregation Function for the UNDP Human Development Index

Un metodo alternativo per aggregare le dimensioni dell'Indice di Sviluppo Umano

Manuela Scioni[1] and Paola Annoni[2]

Abstract Several composite indices have been proposed in literature to measure well-being or quality of life. A fundamental point in the construction of composite indices in this context is the degree of compensability among the different components of the index. We propose an alternative aggregation function in these cases and apply it to the UNDP Human Development Index (HDI). The method we propose adopts the generalized mean of order 0.5 for the HDI dimensions and adds a penalty to countries with an unbalanced dimension profile, to mitigate the level of compensability. The proposed method allows for adjusting the level of compensability between composite index components, providing a useful instrument when constructing an aggregate metric in general, with particular relevance in the social field.

Abstract *Numerosi indici compositi sono stati proposti in letteratura per misurare il benessere o la qualità della vita. In questo contesto un aspetto fondamentale è il grado di compensabilità ammesso tra le diverse componenti dell'indice. Obiettivo di questo lavoro è proporre un metodo di aggregazione da utilizzare in queste situazioni e analizzarne l'applicazione all'Indice di Sviluppo Umano dell'UNDP (HDI). Il metodo proposto utilizza la media generalizzata di ordine 0.5 a cui viene aggiunta una penalità per i paesi le cui dimensioni presentano un profilo non bilanciato, in modo da mitigare l'effetto compensabilità. Tale metodo permette di controllare il livello di compensabilità tra le componenti, aspetto utile nel costruire una metrica aggregata in generale, e in modo particolare in ambito sociale.*

Key words: Human Development, Composite Index, Aggregation Method, Compensability, Unbalance penalisation

[1] Manuela Scioni, Department of Statistical Sciences, University of Padua. Tel: 0039 0498274112. manuela.scioni@unipd.it. ORCID: 0000-0003-3192-4030

[2] Paola Annoni, European Commission, Directorate-General for Communication Networks, Content & Technology; Digital Economy, Recovery Plan and Skills Unit, Brussels, Belgium. Paola.Annoni@ec.europa.eu ORCID: 0000-0001-7628-6656

1. Introduction

Based on the need to go beyond conventional economic measures, and in particular beyond the Gross Domestic Product (GDP), several composite indices have been proposed in the last decades to measure well-being or quality of life. The issue of aggregation in the search for a scalar measure of quality of life is "often perceived as the single most important challenge faced by quality-of-life research" (Stiglitz, Sen and Fitoussi, 2009, pg 56). A fundamental point in the construction of composite indices in this context is the degree of compensability allowed among individual indicators because the deficiency on one aspect shall not be offset by the surplus on another. For this reason, for summarizing the indicators into a single number "partially compensatory" or "non-compensatory" approaches shall be preferred. A well-known composite index that serves as a representative example of this "dilemma" is the Human Development Index (HDI) by the United Nations Development Programme (UNDP, 2010; 2019). Having received quite few criticisms (Desai, 1991; Sagar and Najam, 1998), the HDI developers moved away from the arithmetic mean in 2010. Since then the index is obtained as geometric mean of three dimensions: D1) Life expectancy at birth, D2) Schooling (arithmetic mean of expected years of schooling and mean years of schooling) and D3) Gross National Income (GNI) per capita (natural logarithm). This paper employs the "generalized mean-min" method (Annoni and Scioni, 2022) as an alternative method for the HDI dimensions and investigates the effects on country scores. The "generalized mean-min" method introduces an additive penalisation to country's unbalanced profiles, thus mitigating the compensability among the HDI dimensions. The paper unfolds as follows: Section 2 provides an overview of some aggregation procedure that concerns this application. Section 3 describes results. Concluding remarks are in Section 4.

2. Controlling the level of compensability

The family of the generalised means of order β allows for controlling the level of compensability between the indicators/dimensions in a composite index (Annoni and Weziak-Bialowolska, 2016; Casadio Tarabusi and Guarini 2013; Decancq and Lugo 2013). Let x_{ji} denote the score of indicator (or dimensions) j $(j = 1, \dots, q)$for unit i $(i = 1, \dots, n)$, the generalized mean of order β can be expressed as follows:

$$I_i^{(\beta)} = \left(\frac{1}{q} \sum_{j=1}^{q} w_j x_{ji}^{\beta} \right)^{1/\beta} \qquad \beta \neq 0 \tag{1}$$

where $\{w_1, \dots, w_q\}$ are non-negative weights normalised to have unitary sum and β is

a constant, possibly negative. The indicators (dimensions) are all assumed to be in $\Re+$ and positively oriented with the phenomenon to be measured, for example in the case of the HDI, the higher the dimension's value, the higher the level of human development. Functions $I_i^{(\beta)}$ depend on the value of β. Special cases include the arithmetic mean for $\beta = 1$ and the minimum function (MIN) for the limit value $\beta \to -\infty$ (Decancq and Lugo, 2013; Casadio Tarabusi and Guarini, 2013). The former is a fully compensatory aggregation function whilst the latter is a fully non-compensatory aggregation function. If $\beta = 0$ the generalised mean is defined as the geometric mean, a partially compensatory function. While linear aggregation assumes constant trade-offs among the indicators, geometric aggregation offers inferior compensability for indices with lower values (diminishing returns) (Greco et al., 2019; Munda and Nardo, 2005). In terms of compensability, the family of generalised means of order $\beta \in (0,1)$ occupy an intermediate position, between the arithmetic and the geometric mean.

We apply here the "generalized mean – min" *GMM* function as an alternative aggregation method of the HDI dimensions that introduces a penalty proportional to the unbalance of each country score across the index dimensions (Annoni and Scioni, 2022). The *GMM* function, that is an adaptation of the "mean - min" method originally proposed by Casadio Tarabusi and Guarini (2013), is a two parameters function and uses the generalized mean of order $\beta = 0.5$ plus a penalty applied for the countries with very different values across the three HDI dimensions. For each country i, the penalised HDI score - HDI_{GMM}- is computed as follows:

$$HDI_{GMM}(x_1, x_2, x_3) = \bar{x}_{GM} - \alpha\left(\sqrt{(\bar{x}_{GM} - \min_x)^2 + \gamma^2} - \gamma\right) \quad (2)$$

where (x_1, x_2, x_3) are the HDI dimension scores for country i, $\bar{x}_{GM}$is the generalised mean of order $\beta = 0.5$, $\min_x$= min (x_1, x_2, x_3), the parameter α is the intensity of penalisation for unbalance $(0 \leq \alpha \leq 1)$ and $\gamma \geq 0$ is the level of substitutability between the dimensions. Parameters α and γ are key elements for the choice of the penalisation. Casadio Tarabusi and Guarini (2013) propose $\alpha = \gamma = 1$ as "reasonable values" corresponding "to an intermediate easy case of adjustment with incomplete and progressive compensability" (pg. 31), but guidelines for the selection of these parameter values is not provided in their original proposal. Annoni and Scioni (2022) proposed an empirical, data-driven method to choose the value of the parameters.

Between the two parameters, α is the one with the most limited variability and with the mildest impact on the compensability effects. We set $\alpha = 0.75$. The values $\alpha = 0.5$ and $\alpha = 1$ were also tested but discarded because they led to iso-curves either too linear ($\alpha = 0.5$) or too sharp-cornered ($\alpha = 1$)[1].

The range of variation of γ depends on the order of magnitude of the unbalance term $(\bar{x}_{GM} - \min_x)$. The maximum value of $\Delta = \sqrt{(\bar{x}_{GM} - \min_x)^2 + \gamma^2} - \gamma$ is equal to $(\bar{x}_{GM} - \min_x)$ if $\gamma = 0$ and monotonically decreases as γ increases. If $\gamma \gg (\bar{x}_{GM} - \min_x)$, then γ prevails over the unbalance making it irrelevant in the penalisation term. To select the value of γ we examine two types of graphs:

[1] Results are not shown for sake of brevity.

1) The iso-curves of the penalisation component $\Delta = \sqrt{(\bar{x}_{GM} - \min_x)^2 + \gamma^2} - \gamma$, expressed as a function of γ and of the unbalance $(\bar{x}_{GM} - \min_x)$, moving γ between 0 and the maximum value of the unbalance;

2) The histograms showing the distribution of the unbalance

Figure 1 shows these two graphs for the 2019 HDI (https://hdr.undp.org/en/composite/HDI). The histograms, that depend on the HDI data and not on parameters α and γ, help us assessing the entity and distribution of the unbalance in the case study. Penalisation iso-curves inform us on the choice of γ. For a given a value of γ (here possible values of γ are chosen in between 0 and 0.22), the iso-curves allow us to assess how the penalisation component Δ changes for different values of the unbalance. The two graphs shall be jointly read. For example, for a positively skewed distribution of the unbalance, we would choose a γ value close to 0, in order to differentiate as much as possible the penalization value. In the case of the HDI, the unbalance distribution is quite symmetrical, so we select the value of γ so as to reach a compromise between two extreme situations: the case with minimum or null penalty, Δ close to 0, and the case with maximum penalty. The former case corresponds to the darker area in Figure 1 (left-had side), whilst the latter to the lighter area. This analysis led us to choose $\gamma = 0.1$.

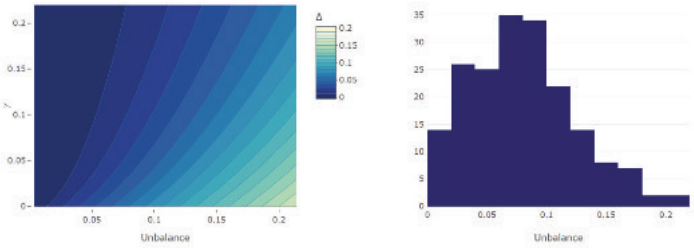


Figure 1 Iso-curves of the penalization components and histograms of the unbalance distribution across the countries

To validate this empirical way of proceeding, a second step is needed. If the aim is to get a level of compensability that is in between maximum penalisation and no penalisation, the shape of the observed iso-curves shall be in between the L-pipe shape and the straight line. To this purpose we aggregate pairwise each HDI dimensions using the *generalised mean-min* aggregation function (2) with $\alpha = 0.75$ and $\gamma = 0.1$ and analyse their corresponding iso-curves. The value of γ is confirmed since the iso-curves have a shape in between the straight line (fully compensatory) and the L-pipe (fully non compensatory)[1].

4. Results

The map in Figure 2 shows the percentage difference between the benchmark (HDI) and the penalised scores (HDI_{GMM}) across the 189 countries assessed by the UNDP.

[1] Figure not shown

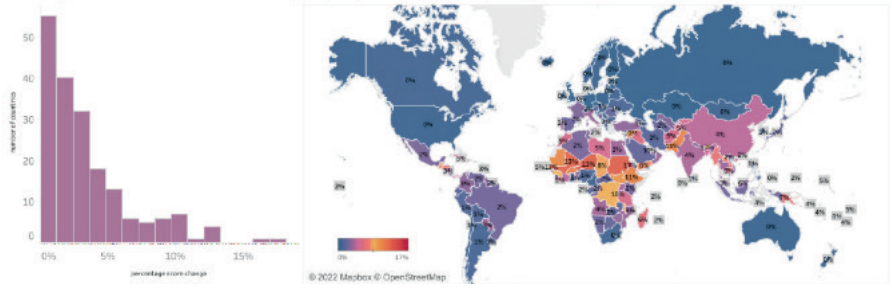


Figure 2 Percentage difference between the HDI and the penalised HDI_{GMM} country scores. Positive changes correspond to a deterioration of the score (the change is defined as: $(HDI - HDI_{GMM})/HDI$.

As expected, the proposed method has the effect of an overall penalisation: all the countries get a score lower than or equal to the HDI one, here considered as the benchmark. The 2019 HDI average is 0.72, while the average HDI_{GMM} is 0.70 (on a 0-1 scale). The level of penalisation goes from a minimum of 0 to a maximum of 0.10, with an interquartile range between 0.008 and 0.035. The percentage difference between the 2019 HDI and the HDI_{GMM} ranges in between 0 and 17.4%.

The penalisation method does not completely disrupt the HDI picture, though, as the correlation between the HDI and HDI_{GMM} scores remains high (0.995 with a p-value<0.001).

The penalisation is less pronounced for countries characterised by a balanced profiles. Amongst the most stable ones, we find Australia and Cameroon. They score very differently on the 2019 HDI, 0.94 for Australia and 0.56 for Cameroon, but their scores do not change with the *GMM* method because their performance is balanced across the three dimensions (Figure 3, top). Among the countries that deteriorate the most, we find Qatar (HDI=0.85, percentage difference=10%) and Senegal (HDI=0.51, percentage difference=17%) (Figure 3, bottom), both penalised by a relatively poorer performance on the Education dimension.

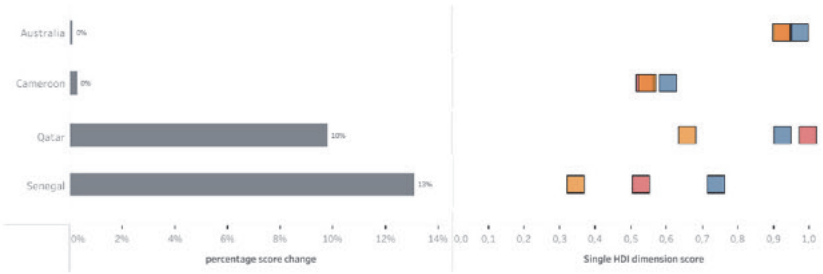


Figure 3 Percentage difference between the HDI and HDI_{GMM} for some countries and corresponding dimensions' scores.

4. Concluding Remarks

Building metrics of well-being or quality of life poses various methodological challenges. The paper addresses the issue of the type of aggregation and uses the Human Development Index (HDI) developed by the UNDP as case study. We apply an alternative aggregation function, the "generalized mean-min" method, to the 2019 HDI that first aggregate the three dimensions through the generalised mean of order 0.5 and apply an additive penalty term (Annoni and Scioni, 2022). The aim is to control the level of compensability between the HDI dimensions. The penalty element is proportional to the unbalance of the country scores across the dimensions and depends on two parameters, α and γ, representing respectively the intensity of the penalisation and the level of compensability (Casadio Tarabusi and Guarini, 2013). The values of the two parameters are chosen following a data-driven approach. In the 2019 HDI case, the values $\alpha = 0.75$ and $\gamma = 0.1$ are found to be the ones ensuring the desired level of compensability, intermediate between full and null compensability. The comparison between the 2019 HDI scores and the penalised ones shows that all countries see their performance either staying the same or deteriorating. As expected, the penalisation effect is particularly evident for countries with an unbalanced profile across the index dimension scores. The proposed aggregation function further reduces the level of compensability among the HDI dimensions with respect to the geometric mean, currently adopted for the index. This is exactly the effect we want for a metric of this type where all the aspects are equally important and a good aspect shall not compensate for a poor one in another.

References

1. Annoni, P., Scioni, M.: The Unbalance Penalisation Method for Metrics of Social Progress. Soc. Indic. Res. Doi: 10.1007/s11205-021-02876-4. (2022)
2. Annoni, P., Weziak-Bialowolska, D.: A measure to target anti-poverty policies in the European Union regions. Appl. Res. Qual. Life, 11, 181-207 (2016)
3. Casadio Tarabusi, E., Guarini, G.: An Unbalance Adjustment Method for Development Indicators. Soc. Indic. Res. 112, 9-45 (2013)
4. Decancq, K., Lugo, M. A.: Weights in multidimensional indices of wellbeing: An overview. Econom. Rev. 32, 7-34. (2013).
5. Desai, M.: Human development: Concepts and measurement. Eur. Econ. Rev. 35, 2-3, 350-357 (1991)
6. Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G.: On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Soc. Indic. Res. 141, 61–94 (2019)
7. Munda, G., Nardo, M.: Noncompensatory/nonlinear composite indicators for ranking countries: A defensible setting. Appl. Econ. 41(12), 1513–1523 (2009)
8. Sagar, A. D., Najam, A.: The human development index: a critical review. Ecol. Econ. 25(3), 249–64. (1998)
9. Stiglitz, J. E., Sen, A., Fitoussi, J. P. Report by the Commission on the Measurement of economic performance and social progress. (2009)
10. UNDP. *Human Development Report 2010: The Real Wealth of Nations - Pathways to Human Development*. New York. UNDP. http://hdr.undp.org/en/content/human-development-report-2010. (2010)
11. UNDP. Human Development Report 2019 - Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. New York: UNDP. (2019)

An ultrametric model for building a composite indicator system to study climate change in European countries

Un modello ultrametrico per la costruzione di un sistema di indicatori compositi per lo studio del cambiamento climatico nei paesi europei

Giorgia Zaccaria and Pasquale Sarnacchiaro

Abstract Nowadays climate change is one of the most urgent topics in public debate, also included into the 2030 Agenda for Sustainable Development adopted by the United Nations, because of its risks for human life. It represents a multidimensional phenomenon defined by different dimensions that pertain to greenhouse gas emissions, human causes of climate change, impacts on humans and natural systems, and efforts of human to avoid and adapt to the consequences, each of which is described by a set of variables directly observed. In this paper, we introduce a new model-based approach to study this phenomenon and its different characterization in European countries. The proposal aims at building a system of composite indicators in order to understand the main determinants of this phenomenon, and to provide guidelines for policy decisions to combat climate change in the European Union framework.

Abstract *Il cambiamento climatico è uno dei temi più urgenti nel dibattito pubblico, incluso nell'Agenda 2030 per lo sviluppo sostenibile adottata dalle Nazioni Unite, a causa dei rischi che può comportare per la vita umana. Esso rappresenta un fenomeno multidimensionale definito da diverse dimensioni, ognuna delle quali misurata da un insieme di variabili direttamente osservabili, che riguardano le emissioni di gas serra, le cause del cambiamento climatico, le sue conseguenze per gli umani e la natura e gli sforzi messi in campo dal genere umano per evitare e adattarsi a tali conseguenze. In questo lavoro, introduciamo un nuovo approccio basato su modello per studiare questo fenomeno e le sue diverse caratterizzazioni nei paesi europei. La proposta metodologica ha l'obiettivo di costruire un sistema di indicatori compositi per identificare i principali fattori che determinano il cam-*

Giorgia Zaccaria
Department of Law and Economics, University of Rome Unitelma Sapienza, Rome, Italy
e-mail: giorgia.zaccaria@unitelmasapienza.it

Pasquale Sarnacchiaro
University of Naples Federico II, Naples, Italy
e-mail: sarnacch@unina.it

biamento climatico e fornire di conseguenza indicazioni per attuare politiche nel contesto europeo volte a combattere il cambiamento climatico.

Key words: Ultrametric models, hierarchical structures, model-based clustering, composite indicators, climate change, sustainable development goals

1 Introduction

Climate change is a compelling topic in public debate because of its risks for human life. One of the main causes of this phenomenon pertain human activities, e.g., industrial processes, energy consumption by fossil fuels, agricultural exploitation of lands. An urgent action to combat these changes is needed. For these reasons, the United Nations included "climate action" within the goals of the 2030 Agenda for Sustainable Development, adopted in 2015 [6], that are monitored at European level by Eurostat throughout indicators and statistics.

In order to study and evaluate multidimensional phenomena as climate change, composite indicators are often used since they are able to summarize a big amount of information via a synthetic measure. Their construction is characterized by different choices that are usually normative, i.e., based upon experts' opinions or a theory known a priori [8]. For these reasons, methods for composite indicator construction are often criticized and considered to be not statistically rigorous [7]. In this paper, we discuss a new model-based approach to build a system of composite indicators for evaluating climate change at European level. In order to study this phenomenon and its different characterization in Europe, we consider the model introduced by Cavicchia, Vichi and Zaccaria [3]. The latter is an ultrametric Gaussian Mixture Model (GMM), where each component of the mixture has an ultrametric structure which is one-to-one associated with a hierarchy of latent dimensions. The idea herein is to extend the ultrametric GMM proposed by [3] to build a system of composite indicators, i.e., including the quantification of latent concepts for each level of the hierarchy, in order to study on which dimensions and in which countries policies to address climate change should be focused.

The paper is organized as follows. In Section 2 the essential background to introduce the methodology used in the paper is provided. Section 3 illustrates the ultrametric Gaussian Mixture Model, which is then applied to study climate change in European countries (Section 4). A final discussion completes the paper in Section 5.

2 Background

In order to define the parsimonious parameterization of the covariance matrix in Section 3, let us introduce the definition of a strict extended ultrametric covariance matrix [3].

Definition 1 (Strict Extended Ultrametric Covariance Matrix). A matrix $\boldsymbol{\Sigma}$ is said to be a *strict extended ultrametric covariance matrix* if all its elements $\sigma_{jl} \in \mathbb{R}$, for $j,l = 1,\ldots,p$, and the following conditions hold:

(i) *symmetry*: $\sigma_{jl} = \sigma_{lj}$ for $j,l = 1,\ldots,p$;
(ii) *positivity of the diagonal*: $\sigma_{jj} > 0$ for all $j = 1,\ldots,p$;
(iii) *ultrametric inequality*: $\sigma_{jl} \geq \min\{\sigma_{jh},\sigma_{lh}\}$, for $j,l,h = 1,\ldots,p$;
(iv) *strictly diagonal dominance*: $\sigma_{jj} > \sum_{\substack{l=1 \\ l\neq j}}^{p} |\sigma_{jl}|$ for $j = 1,\ldots,p$.

Condition (iv), together with conditions (i) and (ii), is sufficient for the positive definiteness of a matrix as shown by Brouwer and Haemers [2]. Definition 1 extends the definition of an ultrametric matrix introduced by Dellacherie et al. [4, pp. 60-61], which is limited to nonnegative matrices.

In the next section, we model the component covariance structure of a Gaussian Mixture Model via a strict extended ultrametric covariance matrix.

3 Methodology

Let $\boldsymbol{x} = (\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n)$ be a random sample - where $\boldsymbol{x}_i$ is a p-dimensional random vector - which is drawn from a population composed of G subpopulations. Suppose that, conditional on the membership to the subpopulation, the density of $\boldsymbol{x}_i$ is the following

$$f(\boldsymbol{x}_i|\boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \, \phi\big(\boldsymbol{x}|\boldsymbol{\mu}_g, \mathbf{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}g} + \boldsymbol{\Sigma}_{\mathrm{B}g})\mathbf{V}_g' - \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{W}g}\mathbf{V}_g') + \mathrm{diag}(\mathbf{V}_g\boldsymbol{\Sigma}_{\mathrm{V}g}\mathbf{V}_g')\big), \tag{1}$$

where $\pi_g, g = 1,\ldots,G$, are the mixing proportions (prior probabilities), and the component covariance structure is parameterized according to a strict extended ultrametric covariance matrix subject to the following constraints

$$\mathbf{V} = [v_{jq} \in \{0,1\} : j = 1,\ldots,p, q = 1,\ldots,Q]; \tag{2}$$

$$\mathbf{V}\mathbf{1}_Q = \mathbf{1}_p \quad \text{i.e.} \quad \sum_{q=1}^{Q} v_{jq} = 1 \quad j = 1,\ldots,p; \tag{3}$$

$$\boldsymbol{\Sigma}_{\mathrm{B}} = \boldsymbol{\Sigma}_{\mathrm{B}}', \mathrm{diag}(\boldsymbol{\Sigma}_{\mathrm{B}}) = \mathbf{0}, {}_B\sigma_{qh} \geq \min\{{}_B\sigma_{qs}, {}_B\sigma_{hs}\} \; q,h,s = 1,\ldots,Q, \; s \neq h \neq q; \tag{4}$$

$$\min\{{}_W\sigma_{qq} : q = 1,\ldots,Q\} \geq \max\{{}_B\sigma_{qh} : q,h = 1,\ldots,Q, h \neq q\}; \tag{5}$$

$${}_V\sigma_{qq} > |{}_W\sigma_{qq}|\Big(\sum_{l=1}^{p} v_{lq} - 1\Big) + \sum_{\substack{h=1 \\ h\neq q}}^{Q} |{}_B\sigma_{qh}| \sum_{l=1}^{p} v_{lh} \quad q = 1,\ldots,Q. \tag{6}$$

$\boldsymbol{\Psi} = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\mathrm{V}g}, \boldsymbol{\Sigma}_{\mathrm{W}g}, \boldsymbol{\Sigma}_{\mathrm{B}g}, \mathbf{V}_g : g = 1, \ldots, G\}$ is the overall parameter vector, where $\mathbf{V}_g$ is a membership matrix which defines a partition of the variable space into a reduced number of groups Q, each one associated with a latent dimension, $\boldsymbol{\Sigma}_{\mathrm{V}g}$ is a diagonal matrix of order Q whose diagonal entries represent the variance of the variable groups, $\boldsymbol{\Sigma}_{\mathrm{W}g}$ is a diagonal matrix of order Q whose diagonal values identify the covariance within the variable groups, and $\boldsymbol{\Sigma}_{\mathrm{B}g}$ is the matrix of order Q whose off-diagonal elements represent the covariance between the groups. The ultrametric paramaterization of the component covariance structure leads to a hierarchy of latent concepts which can assume different characterizations in the components (subpopulations) of the mixture.

The proposal is estimated via a grouped coordinate ascent algorithm [9, 1] by maximizing the objective function defined by Hathaway [5], which is equivalent to maximize the log-likelihood function of (1).

4 A composite indicator system for evaluating climate change in European countries

Every year, Eurostat collects data for monitoring progress towards the Sustainable Development Goals (SDGs) at European level. In order to study climate change, impacts on humans on it and actions to combat it, indicators related to different goals and measured for the 27 European countries and the United Kingdom can be considered. Indeed, other than the indicators defined to monitor Goal 13 - *Climate Action*, e.g., greenhouse gas emissions, we consider variables for measuring the level of pollution in groundwater and rivers (Goal 6 - *Clean Water and Sanitation*), energy consumption and renewable energy production (Goal 7 - *Affordable and Clean Energy*), exposure to air pollution (Goal 11 - *Sustainable Cities and Communities*) and generation of waste (Goal 12 - *Responsible Consumption and Production*). In detail, we include in our analysis indicators from different goals of the SDGs measured on the 27 European countries and the United Kingdom, and we use their population size (i.e., number of inhabitants) to normalize the variables in order to perform comparison among countries.

The motivation of this study lies on the assumption that it is crucial to combine information from different goals in order to define a general composite indicator for measuring climate change, and a system of specific composite indicators that define it. Indeed, the researcher could be interested in identifying dimensions of the phenomenon under study to which policies should be addressed in order to progress towards climatic goal.

5 Conclusions

In this paper, we provide a new hierarchical model-based approach to build a system of composite indicators, and inspect the nested relationships among specific indicators that define the general latent concept under study. The proposal is based upon the ultrametricity notion, which is related to hierarchical (tree) structures, and its implementation in Gaussian mixture models [3]. In particular, we focus on the analysis of climate change by identifying its dimensions and its different characterizations in European countries. The proposal aims at providing guidelines for policy decisions to combat climate change in the European Union framework.

References

1. Bezdek, J., Hathaway, R., Howard, R., Wilson, C., Windham, M.: Local convergence analysis of a grouped variable version of coordinate descent. J. Optim. Theory. Appl. **54**(3), 471–477 (1987)
2. Brouwer, A., Haemers, W.: Spectra of graphs. Springer, New York (2012)
3. Cavicchia, C., Vichi, M., Zaccaria, G.: Gaussian mixture model with an extended ultrametric covariance structure. Adv Data Anal Classif, doi: 10.1007/s11634-021-00488-x (2022)
4. Dellacherie, C., Martinez, S., Martin, J.S.: Inverse M-matrices and ultrametric matrices. Lecture Notes in Mathematics. Springer International Publishing (2014)
5. Hathaway, R.: Another interpretation of the EM algorithm for mixture distributions. Stat. Probab. Lett. **4**(2), 53–56 (1986)
6. Heads of State and Government and High Representatives: Transforming our world: the 2030 agenda for sustainable development, a/res/70/1. Tech. rep., United Nations (2015). URL `https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf`
7. Mazziotta, M., Pareto, A.: Methods for constructing composite indices: One for all or all for one? Rivista Italiana di Economia Demografia e Statistica **67**(2), 67–80 (2013)
8. OECD-JRC: Handbook on Constructing Composite Indicators. Methodology and User Guide. OECD Publisher, Paris (2008)
9. Zangwill, W.: Nonlinear programming: a unified approach. Prentice-Hall, Englewood Cliffs (1969)

Functional Weighted Malmquist Productive Index: a proposal for a dynamic composite indicator

L'indice di produttività di Malmquist funzionale pesato: una proposta per un indicatore composito dinamico

Annalina Sarra and Eugenia Nissi and Tonio Di Battista

Abstract Over time, composite indicators have gained extraordinary popularity in a wide variety of areas of research. Undeniably, the many stages in the construction process of a composite index involve compromises and subjective choices. In this work, restricting our attention on the weighting of indicators, we rely on a data-oriented weighted method. A combination of the Beta profile measure and Malmquist Index, is the proposed approach for meeting the challenge of a constructing a dynamic composite indicator. The originality of the study lies in tracking the temporal evolution of the phenomenon of interest by means of functional tools. The appropriateness of our procedure under specific situations, such as presence of outliers, variation of sample size and collinearity between indicators, is validated through a simulation experiment.

Abstract *Nel tempo, gli indicatori compositi hanno guadagnato una popolarità straordinaria in un'ampia varietà di aree di ricerca. Innegabilmente, le molte fasi del processo di costruzione di un indice composito comportano compromessi e scelte soggettive. In questo lavoro, limitando la nostra attenzione alla ponderazione degli indicatori, ci affidiamo a un metodo ponderato orientato ai dati. Una combinazione della misura Beta profile e dell'indice Malmquist è l'approccio proposto per affrontare la sfida di costruire un indicatore composito dinamico. L'originalità dello studio risiede nel tracciare l'evoluzione temporale del fenomeno di interesse attraverso strumenti funzionali. L'adeguatezza della nostra procedura in situazioni specifiche, come la presenza di valori anomali, la variazione della dimensione del campione e la collinearità tra indicatori, viene convalidata attraverso un esperimento di simulazione.*

Annalina Sarra
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: annalina.sarra@unich.it

Eugenia Nissi
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: eugenia.nissi@unich.it

Tonio Di Battista
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: tonio.dibattista@unich.it

Key words: composite indicators, DEA, β profile, Functional Weighted Malmquist Productivity Index, simulation study

1 Introduction

In the last years, there is a large agreement in the scientific community regarding the multifaced nature of complex social and natural phenomena, like, for instance, well-being, environment, poverty, human development, innovation. As a result, traditional measurements based on a single variable are now often replace by multidimensional methods (see, among others, [14], [11]). A large number of studies grounded in a multidimensional approach rely on the use of composite indicators (CIs). According to the Organisation for Economic Co-operation and Development (OECD) Glossary of Statistical Terms, "a composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model of the multi-dimensional concept that is being measured" [10]. Thanks to their ability to convey large amount of information in a unique measure, CIs are currently employed in various areas of research, for policy analysis, benchmarking comparison, performance monitoring, public communication and decisions. In such way, they enhance the public debate on important issues. Building a composite indicator is far from trivial and there is not a single universal method for achieving taxonomic measures ([1], [9], [12]). Despite their widespread use from governments and international organizations, CIs are not exempt from criticism that surrounds the different steps involved in their construction. Following the recommendations of the document published by the OECD [12], it is well established that the construction decisions range from the selection of a set of sub-indicators, to the standardization method, the weighting and the choice of the aggregation function. In each step, the analyst is tasked with the responsibility of multiple decisions, often compromises and subjective choices, that can have a large impact on the final outcome (e.g.ranking). Our paper contributes to the research on the creation of comparable CIs by presenting a proposal for the construction of a dynamic index, that enables comparative trend analysis. In the proposal, we focus the attention on a fundamental component of index aggregation, namely the weighting of indicators. Specifically, we rely on a data-oriented weighting method, known as Data Envelopment Analysis (DEA). Different works document the application of the DEA in the field of CIs, providing alternative weighting systems (see, among others, [5]). This works takes one step further and addresses the issue of change of performance of units over time, proposing a Functional Weighted Malmquist Productivity Index (hereafter, FWMPI). The rest of the paper is organized as follows. In Section 2, we present the proposed methodological approach and outline the underlying techniques. In Section 3, we give the details of a simulation experiment, while Section 4 concludes the paper.

2 Methodology: the framework

Data Envelopment Analysis (DEA), first introduced by Charnes et al. [3], is a well-known non-parametric method based on linear programming, used for measuring the relative efficiencies of homogenous entities, called Decision Making Units (DMUs), characterised by multiple outputs and inputs. Mathematically, DEA model generates an efficiency score, ranges in an interval from 0 to 1, with efficient DMUs achieving a score of 1 and inefficient DMUs a lower score. The application of DEA to the field of the composite indicators helps to overcome some key limitations. Specifically, the quality of DEA of looking for endogenous weights, for the aggregation of the domains considered into a single composite index, explains a major part of the appeal of DEA-based CIs in real policy-related settings [4]. Within the DEA methodology, the correct way to track the temporal evolution of the phenomenon of interest is by computing the Malmquist Productivity Index (MPI), introduced by Caves et al. [2]. The overall tendency in productivity changes of DMUs over time periods is traditionally obtained through the average of productivity indices of sequential times, implicitly assuming that all sectional indices equally affect the level of productivity. In this work, to eliminate the equal weight effect, we propose a Functional Weighted Malmquist Productive Index (FWMPI). Let us assume to have a set of n DMUs, each having m unitary input denoted by a vector $\mathbf{x}_j$ and s output denoted by a vector $\mathbf{y}_j$, for $(j = 1,\dots n)$, over the periods t and $t+1$. The Malmquist Productivity Index (MPI) allows to calculate the relative performance at different periods of time and is a cross measure, which considers the production of a DMU as the efficient frontier built in the next or previous instant:

$$MPI_0 = \left[\frac{\theta^t x_0^t y_0^t}{\theta^t x_0^{t+1} y_0^{t+1}} \frac{\theta^{t+1} x_0^t y_0^t}{\theta^{t+1} x_0^{t+1} y_0^{t+1}} \right]^{\frac{1}{2}} \tag{1}$$

where $\theta^t x_0^t y_0^t$ and $\theta^t x_0^{t+1} y_0^{t+1}$, $\theta^{t+1} x_0^t y_0^t$ and $\theta^{t+1} x_0^{t+1} y_0^{t+1}$ are respectively the input oriented efficiency measures of DMU_0 at period t and t+1. MPI_0 measures the productivity change between periods t and t+1 [8].

The Malmquist indices computed at two sequential times t and t+1 are denoted by $MPI_{jt}, \quad j = 1\dots n, \quad t = 1,\dots k$. We obtain more objective weights in aggregating the Malmquist indices by averaging the entries of MPI matrix, displayed in Table 1, where the rows represent the MPI measures over times, and the columns represent the DMUs.

Table 1 MPI matrix

	MPI1	MPI2	....	MPIk
DMU_1	MPI_{11}	MPI_{12}	....	MPI_{1k}
DMU_2	MPI_{21}	MPI_{22}	....	MPI_{2k}
DMU_3	MPI_{31}	MPI_{32}	....	MPI_{3k}
....	....	....	....	....
....	....	....	....	....
DMU_n	MPI_{n1}	MPI_{n2}	....	MPI_{nk}

In this work, following the idea of Fallahnejad [7], based on entropy measure and Malmquist method, we propose a functional approach to the problem at hand, adopting the β diversity profile [13]. The overall procedure allows to obtain a functional weighted MPI that facilitates the comparative analysis of the weighted MPI curves. In this regard, it is important to stress that the ranking of DMUs are achieved by means of functional tools, such as analysis of derivatives, radius of curvature and length of a curve, as suggested in [6].
Our proposal can be illustrated via some established steps, as detailed below.
In **Step 1**, the matrix in Table 1 is normalized dividing each element of the column by the sum of column:

$$p_{jt} = \frac{MPI_{jt}}{\sum_{j=1}^{n} MPI_{jt}} \tag{2}$$

The above normalization allows to eliminate anomalies due to different measurement units and scales.
Step 2: find the β diversity profile for all normalized MPI, computed as:

$$\Delta_\beta = \sum_{i=1}^{n} \frac{(1-p_{jt})^\beta}{\beta} p_{jt} \quad \beta \geq -1 \tag{3}$$

In Eq.3, $\sum_{i=1}^{n} \frac{(1-p_{jt})^\beta}{\beta}$ can be interpreted as a measure of relevance of each DMU for the change in productivity over time expressed by each MPI and p_{jt} is defined as above.

Step 3 involves the computation of the degree of diversification, defined as:

$$d_t = 1 - \Delta_\beta \tag{4}$$

In **Step 4** the degree of importance of FMPI at time t is obtained by setting

$$w_t(\beta) = \frac{d_t(\beta)}{\sum_{s=1}^{k} d_s(\beta)} \quad \forall \beta \tag{5}$$

In **step 5**, the functional weighted MPI is calculated as:

$$FWMPI_j(\beta) = \sum_{t=1}^{k} w_t(\beta) MPI_{jt} \quad \forall \beta. \tag{6}$$

3 Simulation design

In this section, we set up a simulation experiment to compare the performance of the Fallahnejad's algorithm and our approach under specific situations, such as presence of outliers, variation of sample size and collinearity between indicators. Specifically, the aim is to document which of the two compared methods leads to a ranking that is substantially closer to outlier-free data. In the simulation setup, we consider

the case of extreme values generated from the Uniform distribution as detailed below, and compare indicators from different sample sizes, over three time periods. The average rank shift (ARS) is used to measure the rank difference between the scalar (Fallahnejad's algorithm) and functional (our proposal) versions of weighted Malmquist index for the DMUs. ARS is defined as:

$$ARS = \frac{1}{n}\sum_{i=1}^{n} |RANK(CIs) - RANK(\tilde{CIs})| \tag{7}$$

In Eq.7, CIs and $\tilde{CIs}$ represent the composite indicator values of a DMU calculated from data contaminated and non-contaminated with extreme values (or Fallahnejad's algorithm *versus* FWMPI). In our simulation experiment, for each period of time, we generate four indicators (I_1,I_2, I_3, I_4) from a Uniform distribution within the interval [5,15] and assume there is no collinearity between them. We successfully vary the collinearity between indicators from no to high correlation. In our setting, we contaminate indicators with extreme values generated from Uniform distribution:

$\tilde{I_1c} \sim \mathrm{U}(k_1 max(I_1); k_2 max(I_1))$,
$\tilde{I_2c} \sim \mathrm{U}(k_1 max(I_2); k_2 max(I_2))$,
$\tilde{I_3c} \sim \mathrm{U}(l_1 min(I_3); l_2 min(I_3))$,
$\tilde{I_4c} \sim \mathrm{U}(l_1 min(I_4); l_2 min(I_4))$,
respectively.

Note that the parameters k_1, k_2 contribute to outlyingness of extreme values on the right tail, while the parameters l_1, l_2 are responsible for the left tail. We set $k_1 = 2$, $k_2 = 3$, $l_1 = 0.2$ and $l_2 = 0.3$. We also enlarge the outlyingness of extreme values over time by increasing k_1, k_2 to 4 and 6 for time period 2 and to 6 and 10 for time period 3, respectively. To randomly select the observations to be contaminated, we define a contamination level ε. Simulation results are obtained under three different scenarios, retained constant over time: no contamination level of contamination ($\varepsilon = 0$) , $\varepsilon = 0.025$ and $\varepsilon = 0.05$. Results of analysis based on simulated data indicate that the FWMPI has the lowest rank shift. Therefore, it follows that our approach enables to weaken the impact of outliers on the ranking of treated DMUs.

4 Conclusions

In recent years CIs have substantially gained in popularity, becoming prevalent as a useful tool for policy evaluations and public communication for understanding multidimensional phenomena. However, it is well acknowledged that if composite indicators are poorly constructed they can affect the overall ranking and send misleading policy messages. It follows that a thorough assessment of assumptions at all stages of construction is required. In this paper, we have contributed, with a dynamic approach, to the research on the creation of comparable composite indicators. Specifically, we have focused on the important issue when constructing composite

indices that comes from the choice of weighting schema of indicators. The existing literature offers a rich menu of alternatives for determining the weights of the components of synthetic measures. As a starting point in our proposal, we have opted for DEA as a data-oriented weighted method. Once we have obtained a measure of the efficiency change over time by means of Malmquist indexes, we proceeded on constructing a functional weighted MPI by integrating the Malmquist DEA scores with the Beta profile. Our methodological approach takes advantages of tools borrowed from Functional Data Analysis, for successfully meeting the challenge of comparative trend examinations. The appropriateness and versatility of the proposal have been validated through a simulation experiment where specific situations, such as presence of outliers, variation of sample size and collinearity between indicators, are taken into account. The sensitivity analysis conducted in this paper, allows us to conclude the robustness of using a functional weighted MPI improves the reliability of CIs ranking over time for policy making.

References

1. Booysen, F.: An Overview and Evaluation of Composite Indices of Development. Soc. Indic. Res. **59**, 115-151 (2002)
2. Caves, D. W., Christensen, L. R., Diewert, W. E.,: The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica (2008) **50**, 1393-1414 (1982)
3. Charnes, A., Cooper, W., Rhodes, E.: Measuring the efficiency of decision making units. Eur. J. Oper. Res. **2**, 429-444 (1978)
4. Cherchye, L., Moesen, W., Rogge, N., Puyenbroeck, T.V.: Benefit of the doubt composite indicators. KEI-project, European Commission's Sixth Framework Programme (2008)
5. Despotis, D.K.: A reassessment of the human development index via data envelopment analysis. J. Oper. Res. Soc. **56**(8), 969–980 (2005)
6. Di Battista, T., Fortuna, F., Maturo, F.: Environmental monitoring through functional biodiversity tools. Ecol. Indic. **60**, 237-247 (2016)
7. Fallahnejad, R.: Entropy based Malmquist Productivity Index in Data Envelopment Analysis. Int. J. Data Envel. Anal. **5**(4), 1425–1434 (2017)
8. Fare R., Grosskopf S., Norris M., Zhang Z.: Productivity growth, technical progress, and efficiency change in industrialized countries. Am. Econ. Rev. **84**(1), 66-83 (1994)
9. Mazziotta, M., Pareto, A.: A non-compensatory composite index for measuring well-being over time. Cogito, J. Multidiscip. Res. **5**(4), 93–104 (2013)
10. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. : Handbook on constructing composite indicators. Paris: OECD Publishing (2005)
11. Nicholas, A., Ray, R., Sinha, K.: Differentiating between dimensionality and duration in multidimensional measures of poverty: Methodology with an application to china. Rev. Income Wealth **60**, 48-74 (2017)
12. OECD: Handbook on Constructing Composite Indicators. Methodology and user guide. Paris: OECD Publications. (2008)
13. Patil, G., Taillie, C.: Diversity as a concept and its measurement. J. Am. Stat. Assoc. **77**, 548–567 (1982)
14. Sen, A. K.: Development as freedom. Oxford University Press, Oxford. (1999)

CFA & PLS-PM for UX-AI product infused
CFA & PLS-PM per l'utilizzo di smart speakers basati su AI

Emma Zavarrone and Rosanna Cataldo[1]

Abstract Since their appearance, artificial intelligence (AI) products have increasingly integrated into our daily life and today they are becoming an ever more pervasive reality. However, our expectations often exceed the current potential performance of AI products, causing disappointment and frustration. Starting from this consideration and taking smart speakers as a reference, the work studies new scale validation oriented to evaluate the user Experience (UX) with AI-infused products. The work compares the results of three-order model under the perspective both in the Partial Least Squares-Path Modeling and in the confirmatory approach.

Abstract *Oggi gli smart speakers basati sull'Intelligenza Artificiale stanno diventando una realtà sempre più pervasiva. Purtroppo, non sempre le aspettative degli utenti sono realmente soddisfatte dalle performance dei prodotti. Il lavoro propone e valida una nuova scala volta a valutare le dimensioni più adatte per valutare le esperienze degli utenti (UX). L'analisi confronta i risultati, considerando l'analisi confermativa e i modelli ad equazioni strutturali.*

Key words: AI-infused products, Confirmatory Factor Analysis, Partial Least Squares- Path Modeling

[1] Zavarrone Emma, IULM University, email: emma.zavarrone@iulm.it
Cataldo Rosanna, University of Naples "Federico II", email: rosanna.cataldo2@unina.it

1 Introduction

Proposing measurement scales to frame the User Experience (UX) of a new generation of products infused with Artificial Intelligence (AI) has been the source of many researches. The literature review (Spallazzo and Sciannammè, 2021) based on 129 papers has highlighted the overlapping of measurement scales able to capture the UX generalist dimensions without covering the aspects linked to Human Computer Interaction (HCI) and AI. These results are consistent with previous studies which had already tackled the issue (Vermeeren *et al.*, 2010; Bargas-Avila and Hornbæk, 2011; Rivero and Conte, 2017; Pettersson *et al.*, 2018). They spotted the main trends and paved the way to new UX research, highlighting, for instance, that it should cope with ever-evolving technologies and non-human intelligence (Pettersson *et al.*, 2018). In general, all these studies focused only on usability, pleasantness of use, emotion/affect, and enjoyment. Furthermore, the literature reveals no agreement on the use of terminology, and sometimes the same words are used without careful attention to the semantic nuances or interchangeably as dimensions and descriptors and items. Our research is placed in this scenario. The purpose of this work is the construction and validation of a scale, defined UX-AI, that takes into consideration, in addition to the aspects already known in the literature, other different aspects. Section 2 is devoted to introducing the model hypothesis characterized by detecting the common core shared between HCI and UX dimensions. Section 3 presents data and compares the results under the perspective both in Partial Least Squares- Path Modeling (PLS-PM) and in the confirmatory approach.

2 Model

The research of a theoretical measurement framework in which the most crucial latent dimensions characterized the UX are clustered, has moved us to choose a reflexive paradigm. Under this perspective we follow the development of UX-AI scale adopting the classical paradigm proposed by Gerbing and Anderson (1988) approach. The paradigm is characterized by four steps: literature review; focus group and item generations, item selection and model validation. The first two steps have been realized conjointly with the Meet AI, a research group of Politecnico of Milan. The hypothetical latent dimensions and their links have been identified by Spallazzo and

Sciannammè (2021), composed by *Pragmatic, Hedonic, Aesthetic,* and *Affective* ones can be considered as second order constructs. Each second order construct can be organized in terms of the first-order constructs as described in Table 1.

Table 1: Latent dimensions and their descriptors

Second Order Constructs	First-Order Constructs
Pragmatic	Helpfulness, Efficiency, and Functionality, easiness, simplicity, clearness, navigation, learnability, reliability, and convenience
Hedonic	Enjoyability, excitement, creativity, inventiveness, innovativity
Aesthetic	Appearance, attractiveness, enjoyability, excitement, creativity, inventiveness, and innovativity
Affective	Valence and arousal

The *Pragmatic* dimension concerns the use of an artifact in terms of *helpfulness, efficiency,* and *functionality*, *easiness, simplicity, clearness, navigation, learnability, reliability, and convenience*. The *Aesthetics* presents multiple facets, can mainly be interpreted in terms of the first order constructs: *appearance* (where *clarity* and *sophistication* are two recurring themes), or *attractiveness* (*good* and *pleasant* qualities being frequently assessed). The *Hedonic* dimension often overlaps with the affective one, as their borders are very subtle. *Enjoyability* and *excitement*, *creativity*, *inventiveness,* and *innovativity* appear to have their weight. Finally, the *Affective* component is largely influenced by basic emotions: *pleasure, fear, sadness, happiness, disgust, anger*, and *surprise* disseminated through referral and appraisal.

The innovative dimensions not explored in the literature but identified following the focus group step are indicated: ***Trustworthiness:*** concerns *accuracy, data management, data protection, reliability,* and *transparency*. ***Conversational dimension:*** is mostly relatable to *accuracy, context awareness, understanding, feedback quality*, *fluidity* and *naturalness*. ***Intelligence:*** presents some traits reminding human capabilities (e.g. *learning, understanding needs, companionship*), and others strictly linked to the machine dimension (e.g. *data elaboration, connectivity*). ***Meaningfulness:*** mostly appeal to the human-computer/product relationship.

The proposed model suggests configuring an endogenous UX third-order dimension linked through the second and first order constructs according to Figure 1 and Table 2.

Table 2: Hierarchical model

Third Order Construct	Second Order Constructs	First-Order Constructs
UX	Intelligence (INT)	Accuracy
		Adaptability
		Context awareness
		Understanding
	Trustworthiness (TRU)	Accuracy
		Data management
		Privacy
		Reliability
		Transparency
		Pragmatic (PRA)
		Hedonic (HED)
		Conversational dimension (CON)
		Meaningfulness (MEAN)

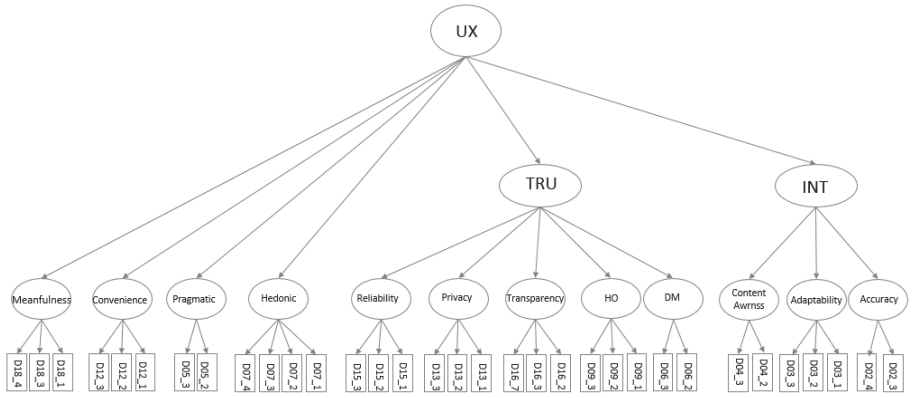


Figure 1: Reflective Hierarchical Model

The item generation step has defined 139 items with an ordinal four-point scale. The choice of even points has been setted for avoiding the uncertainty by forcing the respondent to express an opinion. Recent studies (Rhemtulla et al., 2012) have highlighted that a score with an ordinal nature doesn't represent a limit for the development of a scale in a reflective perspective. The generated pool of items has been evaluated by experts in order to test both the content and face validity and reliability in a pre-test analysis. In this step a considerable reduction of items has been realized obtaining 65 items, on average seven first-order constructs. In a second step a further reduction led to obtaining 33 items, on average three first-order constructs. This structure was validated under both Confirmatory Factor Analysis (CFA) and

PLS-PM approaches. Last step has been devoted to the comparison of the composite reliability

3 Data and Results

A further items reduction has been operated obtaining 36 items through EFA, applied on a polychoric correlation matrix composed of 75 items (collected on 671 Americans and British users) estimated with the WLS method. The validation stage has been characterized by the second survey composed of 735 units. Demo social variables have been measured conjointly with the device type for both the samples. Table 3 shows the structure of the second sample that has been used for applying the Confirmatory Factor Analysis (CFA) using a Diagonally weighted least squares (WLSMV) estimator and PLS-PM Analysis. The CFA excludes three items so the item final number becomes 33. The CFA and PLS-PM goodness of fit index and Cronbach coefficient for each dimension and for overall have been reported in Table 4.

Table 3: Socio-demographic analysis

Gender	58% Male; 42% Female
Country	50% UK; 50% USA
Age	27% 18-30; 23% 31-40; 18% 41-50; 14% 51-60; 18% over 60
Device	64% Amazon Echo; 17% Google Nest; 13% Apple; 6% Other
Use_Device	48% Frequently; 36% Sometimes; 12% Seldom; 4% Never

Table 4: CFA and PLS-PM Goodness of fit and Cronbach coefficients

		N Items	CFA	PLS-PM
Cronbach coefficients	**UX**	33	0.967	0.968
	INT	7	0.873	0.879
	PRA	2	0.800	0.807
	HED	4	0.914	0.914
	TRU	14	0.941	0.944
	CONV	3	0.866	0.872
	MEAN	3	0.873	0.877
Goodness measures			RMSEA= 0.042	GoF= 0.698
			SRMR= 0.033	
			CFI= 0.986	
Composite Reliability			0.99	0.97
AVE			0.89	0.73

About the reliability of the questionnaire, all dimensions present high Cronbach coefficients for both approaches. It is generally held that a Cronbach coefficient of over 0.7 is acceptable (Devellis, 2012; Cataldo et al. 2017).
About goodness of fit measures, for CFA the Root Mean Squared Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) are absolute measures of fit and that are used to assess how well the a priori models fits the sample data (a value lower than 0.05 indicates a good model fit), while the Comparative Fit Index (CFI) is a relative measure and assesses the relative improvement in fit of model compared with the baseline model (the conventional threshold for a good fitting model is greater than 0.90). For PLS-PM the Goodness of Fit can be used as a measure of goodness of the model.
As can be seen from Table 4, in our analysis the model presents a good fit as all measures exceed their respective thresholds. Under the two approaches composite reliability (CR), used for measuring internal consistency of items and average validity extracted (AVE), used for measuring on average, how much variations of items scale can be explained by the constructs have been computed. Overall CR and AVE have values greater than 0.75 and 0.5 respectively confirming the validity and reliability of the UX-AI scale.

References

1. Bargas-Avila, J. A., & Hornbæk, K.: Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 2689-2698). (2011)
2. Cataldo, R., Grassia, M.G., Lauro, N.C., & Marino, M.: Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. Quality & Quantity. Springer, 51(2), 657–674. (2017)
3. Devellis, R. F.: Scale development: Theory and application, Thousand Oaks, CA: Sage. (2012).
4. Pettersson, I., Lachner, F., Frison, A. K., Riener, A., & Butz, A.: A bermuda triangle? A review of method application and triangulation in user experience evaluation. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-16). (2018)
5. Gerbing, D. W., & Anderson, J. C.: An updated paradigm for scale development incorporating unidimensionality and its assessment. Journal of marketing research, 25(2), 186-192, (1988)
6. Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods, 17(3), 354–373.
7. Rivero, L., & Conte, T.: A systematic mapping study on research contributions on UX evaluation technologies. In Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems (pp. 1-10). (2017)
8. Spallazzo, D. and Sciannamè, M.: UX Descriptors for AI-infused Products. figshare. doi:10.6084/M9.FIGSHARE.14345498.V1. (2021)
9. Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K.: User experience evaluation methods: current state and development needs. In Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries (pp. 521-530). (2010)

Fertility, adulthood, and economic uncertainty

Uncertainty and fertility intentions: a comparison between the Great Recession and the Covid-19 crisis

Incertezza e intenzioni di fecondità: la Grande Recessione e la crisi pandemica a confronto

Chiara Ludovica Comolli

Abstract Theories of the fertility response to business cycle fluctuations consider crises exclusively as economic experiences. Yet, economic downturns are also social phenomena, affecting communities and morality. When societal changes are of a sufficient magnitude, they tend to break down the social fabric and represent additional sources of uncertainty which may produce effects on reproductive decisions beyond those of economic unpredictability alone. These aspects have been overlooked by the literature on the determinants of low fertility in western societies in the context of crises. Using the Swiss Household Panel 1999-2020, this study evaluates the effects of forms of uncertainty on childbearing intentions during the two most recent crisis: the Great Recession of 2008 and the Covid-19 pandemics.

Abstract *Le teorie sulla risposta della fecondità alle fluttuazioni del ciclo economico considerano le crisi esclusivamente come esperienze economiche. Tuttavia, le crisi economiche sono anche fenomeni sociali che influenzano le comunità e la moralità. Quando i cambiamenti sono di portata sufficiente, tendono a rompere il tessuto sociale e rappresentano fonti di incertezza che possono produrre effetti sulle decisioni riproduttive oltre a quelli della sola imprevedibilità economica. Questi aspetti sono stati trascurati dalla letteratura sui determinanti della bassa fecondità nelle società occidentali nel contesto delle crisi. Utilizzando i dati del Swiss Household Panel 1999-2020, questo studio valuta gli effetti di queste forme di incertezza sulle intenzioni di fecondità durante le due crisi più recenti: la Grande Recessione del 2008 e le pandemie del Covid-19.*

Key words: fertility, intentions, uncertainty, Covid-19, Great Recession

[1] Chiara Ludovica Comolli, Università di Bologna; email: chiara.comolli@unibo.it

1 Introduction

The 2008 Great Recession triggered negative fertility trends in Europe, but the declines persisted long after the economic recovery [30]. Part of this unaccounted fertility decline can be explained by the rise of economic *uncertainty* – a condition in which the future cannot be deduced by present information [11] – that persisted long after the economic recovery. Economic uncertainty alone, however, does not explain why fertility is still declining more than ten years after the Great Recession, including the United States, or in countries like the Nordic European states which were marginally affected by the crisis and economic insecurity [10].

Against this backdrop, Covid-19 and the measures adopted to contain its diffusion have led to another crisis. The social, financial and labour market losses have been considerable, and the post-pandemic prospects remain highly uncertain. This will possibly produce additional negative effects on childbearing. The Covid-19 crisis is first and foremost a health emergency, but in the longer run individuals may adjust their behaviour to the considerable economic and social costs of the pandemic. As a result, fertility is likely to be affected too [2]. Early evidence hints to a negative effect but the magnitude, contextual heterogeneities, and the mechanisms explaining childbearing delays are unknown [24].

While theories of the fertility response to business cycle fluctuations consider major crises exclusively as economic experiences, their effects are rarely confined to one domain. Economic downturns are also social phenomena. They have an impact on psychological outcomes, communities and morality, and on social interactions and individual's sense of belonging to certain groups. When these societal changes are of a sufficient magnitude, they tend to break down the social fabric and they represent additional sources of uncertainty which may produce effects on reproductive decisions beyond those of economic unpredictability alone. All these features of social change directed towards the fragmentation of societies, greater complexity, and the increase of insecurity about future societies can be subsumed under the notion of *social uncertainty*. These aspects have been largely overlooked by the literature on the determinants of low fertility in contemporary societies and especially in the context of crises [for an exception, see 8]. Using long-term longitudinal Swiss households' data stretching from 1999 to 2020, this study focuses on the last two decades and in particular it compares the effects on childbearing intentions of the Great Recession and the early effects of the Covid-19 pandemics. The study aims at uncovering the link between social and economic uncertainties, and determining the elicited fertility responses.

2 Background

The New Home Economics (NHE) theory sees childbearing as a rational choice based on the costs and benefits of children and posits that a decline in households' income is linked to the postponement of childbearing [5]. When incomes drop and unemployment rates rise, long-term commitments - such as housing purchases or having children - tend to be postponed. The initial declines in Total Fertility Rates (TFR) in the aftermath of the Great Recession, were thus not surprising as births were postponed hoping for a future with better economic prospects. Despite the economic recovery in the second half of the 2010s, however, the US' and European countries' TFRs continued their descent, with countries as diverse as Greece, England, Wales, Finland and Norway reporting all-time-low TFRs in 2019 [13]. Part of this decline has been linked to persistent *economic uncertainty* causing a re-evaluation of future prospects and the postponement of commitments like having children [9].

In fact, objective economic conditions are not the unique rationale of childbearing. First, the subjective perception of those conditions leads to different evaluations of current and future prospects and childbearing opportunities [32]. Second, preferences, attitudes and aspirations together with social norms and values guide individuals' attachment to paid work, reproductive decisions and the link between the two [22]. Third, institutions and structural constraints influence childbearing [13]. Finally, individuals do not live and take decisions in isolation but are embedded in a social context [12].

Individuals use community actions as sources of information to navigate complex situations [28] and to choose their own course of action [26]. Social interaction and the ensuing exchange of information, services and goods between individuals is a powerful determinant of fertility [4]. Yet, social interactions do not only produce *information* but also produce *resources* or *social capital* [27]. Social capital permeates spatial as well as social settings [33]. Social psychologists and political scientists show that social capital tends to be affected by long-lasting periods of lower opportunities. Enduring inequalities elicit lower social cohesion and mounting intra- and inter-generational tensions. Inequality lowers civic spirit, trust, and civic engagement, and correlates with political polarization and the emergence of populist parties [31]. Sociologists add that inequality is socially dysfunctional. It causes marginalization, increasing anxiety over the preservation of social status and reducing social trust. Economic crises threaten individuals' sense of control. To compensate, group-based control increases, fostering in-group limited trust at the expense of generalized trust, hence reducing the *radius of trust* [16].

Generalized trust represents a crucial human strategy to reduce the complexity arising from the burden of the multiplicity of possible events [23], and to mitigate uncertainty [7]. On the contrary, the shredding of good practices induces a sense of insecurity that pervades lives, especially during times of social changes. Lower trust levels have been linked to lower fertility [1] via a lower likelihood to marry [6], a lower quality of institutions (low childcare public provisions) and a lower likelihood of couples to outsource child care [3]. However, while these studies see social capital and morality as invariable traits of groups or societies, the abovementioned literature shows that major crises and the ensuing "miniaturization" of the social fabric [16] do alter these traits, generating further uncertainties related to the future of social support, values and the quality of institutions, which might induce a further childbearing postponement. The hypothesis that economic crises generate not only economic but also social uncertainties and that the latter too drive contemporary fertility declines has never been explored.

3 The current study

The study will expand this literature, first, by integrating theories of *economic* uncertainty with those of *social* uncertainty, as the latter have been overlooked in the literature of low fertility. Second, the study will contextualize post-pandemic fertility trends into a long-term perspective that includes the post-Great Recession developments.

While Covid-19 and confinement measures have possibly led to larger economic losses than the Great Recession, the nature of this latest recession is more exogenous. This could suggest a quicker (V-shaped) recovery leading to smaller effects on fertility than those observed during the 2010s decade. However, the effects of the pandemic might cumulate with those of the Great Recession, exacerbating the existing negative fertility tendency.

3.1 Research questions

On the one hand, the pandemic and lockdowns are redefining the meaning of paid and unpaid work and their boundaries, which represents a discontinuity with the pre-pandemic period and might produce unexpected consequences on childbearing behaviour. While women normally work in less volatile sectors that suffer smaller job losses during recessions, in 2020 jobs' features such as sociality or

teleworkability defined new employment hazards, with women overrepresented in the most hit sectors [29]. The reversal of the gendered employment hazard might weaken the selection of women into certain occupations and change the relationship between female work and childbearing.

Moreover, physical distancing involves limiting access to external care and outsourcing. The relative increase in job insecurity among women coupled with cuts in outsourcing and shrinking boundaries between paid and unpaid work, will likely affect childbearing decisions. *The first research question of the study asks thus whether the relationship between women's employment and childbearing decisions has changed between the Great Recession period and the Covid-19 pandemic years.*

On the other hand, societies and social relationship are likely to be affected by long-lasting periods of lower opportunities. Studies in social psychology and sociology demonstrate that inequality lowers social trust [18]. The spread of Covid-denial, anti-mask, anti-vaccine movements illustrate how divisive for society the experience of the pandemic can be. Isolation may increase stress and conflict between partners, lowering relationship quality and wellbeing, and might induce couples to postpone marriage or anticipate union dissolution [17]. Lockdowns may break kinship and network ties normally sustaining families. Distancing limits socialization, moving physical encounters even more to the digital world. All in all, lower wellbeing, increasing virtual interactions at the expenses of interpersonal contact, and declining social cohesion and trust, might produce childbearing delays through delayed union formation, lower network support, greater uncertainty and lower likelihood of outsourcing [1,3]. *The second research question of the study is thus whether social uncertainties, beyond objective and subjective economic conditions, influence childrearing decisions and, third, whether this relationship has changed from the Great Recession period to the early years of the pandemic crisis.*

3.2 Data and method

To answer the three research questions presented above I use the 22 waves available of the Swiss Household Panel (SHP, 1999-2020), a panel survey of households in which all adult members are interviewed annually.

Two refreshment samples, in 2004 and 2013, were added to the initial sample of 1999 to compensate for attrition. These data are particularly useful for the current study because they cover a very long period of time including the pre and post Great Recession years up to the most recent data collected during the Covid-19 pandemics in 2020, in the early spring during the lockdown in a special Covid-19 module, plus the usual summer-fall wave of the SHP. Longitudinal data are needed to investigate

how changes over time in the explanatory variables of interest affect childbearing decisions.

The analytical sample will consist of women in reproductive age 15-49. Because it is too early now to observe changes in childbearing behavior induced by the pandemics, here I investigate short-time childbearing intentions, considered a good proxy of actual reproductive behavior. The dependent variable is thus the intention to have a first or another child in the next few years. In the SHP the question posed is whether a *child is wanted in the next 24 months* (yes/no).

The first set of independent variables pertaining to the economic realm are women's objective economic conditions (changes in employment status, working hours, sector of occupation or income) and perceived economic uncertainties (job insecurity). A second set of independent variables related to social uncertainty include generalized trust, measured through the question of whether *most people can be trusted or you can't be too careful in dealing with people* (answers in the scale 0-10: from Can't be too careful to Most people can be trusted), and the engagement on *volunteering activities* as a measure of *civic engagement*. In the SHP membership to voluntary associations is measured in term of *active/passive or no membership*[1]. Volunteering is additionally measured with respondents *having honorary or voluntary activities within an association, an organization or an institution*. Another explanatory variable of interest relates to the *sense of control* over life measured in the SHP through the question: "How often have you felt that you were unable to control the important things in your life?" (answers in the scale 1-5: from Never to Very often). Finally, possible mechanisms such as childcare usage and other domestic tasks outsourcing, division of unpaid work in the couple, subjective wellbeing (life and domain specific satisfaction), and relationship quality are also included in the SHP survey. Control variables such as women's age and nationality, marital status, educational level and number of children already had.

In a first set of analyses, I will run several (step-wise) Linear probability Models of the probability of expressing a positive intention to have an (additional) child in the next two years depending on women's objective and subjective economic conditions and, net of those, on the residual effect of social uncertainties on fertility intentions. In the second part of the analysis the main focus will be on the interaction term of economic and social uncertainties with a period categorical variable (1999-2008; 2009-2013; 2014-2019; 2020) to compare their association to fertility intentions in aftermath of the Great Recession with the year of the Covid-19 pandemic.

[1] For instance, local or parents' association, sports or leisure association, organization involved in cultural activities, music, or education, syndicate, employee's association, political party, organization concerned with protection of the environment, charitable organization, women's association, and tenants' rights association.

References

1. Aassve, A., Billari, F. C., and Pessin, L.: Trust and fertility dynamics. Social Forces, 1-30. (2016)
2. Aassve, A., Cavalli, N., Mencarini, L., Plach, S., and Sanders, S.: Early assessment of the relationship between the COVID-19 pandemic and births in high-income countries. Proceedings of the National Academy of Sciences, 118(36) (2021).
3. Aassve, A., Mencarini, L., Chiocchio, F., Gandolfi, F., Gatta, A., and Mattioli, F.: Trustlab Italy: a new dataset for the study of trust, family demography and personality (No. 115) (2018).
4. Balbo, N., Billari, F. C., and Mills, M.: Fertility in advanced societies: A review of research. European Journal of Population/Revue européenne de Démographie, 29(1), 1-38 (2013).
5. Becker G.S.: A Treatise on the Family. Enlarged edition. Cambridge, MA: First Harvard University Press (1993).
6. Cherlin, A. J., Ribar, D. C., and Yasutake, S.: Nonmarital first births, marriage, and income inequality. American sociological review, 81(4), 749-770 (2016).
7. Colquitt, J. A., LePine, J. A., Piccolo, R. F., Zapata, C. P., and Rich, B. L.: Explaining the justice–performance relationship: Trust as exchange deepener or trust as uncertainty reducer? Journal of applied psychology, 97(1), 1 (2012).
8. Comolli, C.L. and Andersson, G. (2021). *Partisan fertility response in the wake of the Great Recession in Sweden*. SRRD Working paper series 2021:25.
9. Comolli, C.L. and Vignoli, D.: Spreading uncertainty, shrinking birth rates. European Sociological Review, 37(4), 555-570 (2021).
10. Comolli, C.L., Neyer, G., Andersson, G., Dommermuth L., Fallesen, P., Jalovaara, M., Jónsson A., Kolk, M., Lappegård, T.: Beyond the economic gaze: Childbearing during and after recessions in the Nordic countries. European Journal of Population, 1-48 (2021).
11. Dequech, D.: Fundamental uncertainty and ambiguity. Eastern Economic Journal, 26(1), 41-60 (2000).
12. Elder, G. H.: Children of the great depression. Routledge (1974).
13. Esping-Andersen, G., and Billari, F. C.: Re-theorizing family demographics. Population and development review, 41(1), 1-31 (2015).
14. Eurostat:https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_frateandlang=en. Fertility rates by age. Last accessed 29 January 2021.
15. Fritsche, I., Moya, M., Bukowski, M., Jugert, P., de Lemus, S., Decker, O. and Navarro-Carrillo, G.: The great recession and group-based control: Converting personal helplessness into social class in-group trust and collective action. Journal of Social Issues, 73(1), 117-137 (2017).
16. Fukuyama, F.: The Great Disruption. Human Nature and the Reconstitution of Social Order. London: Profile Books (1999) .
17. Guetto, R., Vignoli, D., and Bazzani, G.: Marriage and cohabitation under uncertainty: the role of narratives of the future during the COVID-19 pandemic. European Societies, 1-15 (2020).
18. Haushofer, J., and Fehr, E.: On the psychology of poverty. Science, 344(6186), 862-867 (2014).
19. Kearney, M. S., and Levine, P. B.: Income inequality, social mobility, and the decision to drop out of high school (No. w20195). National Bureau of Economic Research (2014).
20. Kohler, H. P.: Social interactions and fluctuations in birth rates. Population Studies, 54(2), 223-237 (2000).
21. Kohler, H. P., Billari, F. C., and Ortega, J. A. The emergence of lowest-low fertility in Europe during the 1990s. Population and development review, 28(4), 641-680 (2002).
22. Lesthaeghe, R., and Van de Kaa, D. J. Two demographic transitions. Population: Growth and decline, 1, 9-24 (1986).
23. Luhmann, N.: Trust and power. New York, NY: Wiley (1979).

24. Luppi, F., Arpino, B., and Rosina, A.: The impact of COVID-19 on fertility plans in Italy, Germany, France, Spain, and the United Kingdom. Demographic Research, 43, 1399-1412 (2020).
25. Matsudaira, J. D.: Economic conditions and the living arrangements of young adults: 1960 to 2011. Journal of Population Economics, 29(1), 167-195 (2016).
26. Montgomery, M. R., and Casterline, J. B.: Social learning, social influence, and new models of fertility. Population and development review, 22, 151-175 (1996).
27. Putnam, R. D.: Making democracy work. Civic traditions in modern Italy. Princeton: Princeton University Press (1993).
28. Rossier, C., and Bernardi, L.: Social interaction effects on fertility: Intentions and behaviors. European Journal of Population, 25(4), 467-485 (2009).
29. Shibata, I.: The distributional impact of recessions: The global financial crisis and the COVID-19 pandemic recession. Journal of Economics and Business, 105971 (2020).
30. Sobotka, T., Skirbekk, V., and Philipov, D.: Economic recession and fertility in the developed world. Population and development review, 37(2), 267-306 (2011).
31. Uslaner, E. M., and Brown, M.: Inequality, trust, and civic engagement. American Politics Research, 33(6), 868-894 (2005).
32. Vignoli, D., Bazzani, G., Guetto, R., Minello, A., and Pirani, E.: Uncertainty and Narratives of the Future: A Theoretical Framework for Contemporary Fertility. In Analyzing contemporary fertility (pp. 25-47). Springer, Cham (2020).
33. Vitali, A., and Billari, F. C.: Changing determinants of low fertility and diffusion: A spatial analysis for Italy. Population, Space and Place, 23(2), e1998 (2017).

Interpreting the relationship between life course trajectories and explanatory factors. An example on the transition to adulthood

Due approcci per lo studio delle relazioni tra le sequenze di vita e fattori esplicativi. Un esempio sulla transizione allo stato adulto

Danilo Bolano, Matthias Studer and Reto Buergin

Abstract Using sequence analysis as analytic framework, we discuss two alternative methods to investigate the link between a trajectory (sequence of states) and explanatory factors of interest: implicitative statistics and discrepancy analysis. Data visualisation tools to show the results are discussed as well. An illustrative example on differences in transition to adulthood in Switzerland is shown.
Abstract In questo lavoro si discute due possibili approcci nelle analisi delle sequenze per studiare la relazione tra la sequenza osservata e alcuni fattori esplicativi. Nel primo approccio, basato sulla statistica implicativa, il focus risiede sulle differenze in un dato momento. Il secondo approccio, basato su una analisi di discrepanza, invece prende in considerazione l'intera traiettoria osservata. I due approcci sono illustrati via un esempio sulle differenze di genere e status socio-economico nella transizione allo stato adulto.

Key words: Sequence analysis, Life course research, Sequential data, Discrepancy analysis, Data visualisation tools

1 Introduction

The transition to adulthood is a complex and multifaceted demographic phenomenon that cannot be analysed simply as a "transition" in/out a given condition. It is rather a process that might involve multiple changes in roles and states (e.g changes in living arrangements, moving out parental home, starting a union, becoming parent)

Danilo Bolano
Bocconi University (IT) - Dondena Center. e-mail: danilo.bolano@unibocconi.it

Matthias Studer
University of Geneva (CH) - Center LIVES. e-mail: matthias.studer@unige.ch

Reto Buergin
ZHAW School of Engineering (CH). e-mail: reto.buergin@zhaw.ch

with the time being a crucial determinant. We focus here on sequence analysis. An holistic approach particularly effective in studying periods of life, such as young adulthood, full of events and transitions [2]. In sequence analysis, the variable of interest is not the timing of an event, or the probability of experiencing a given transition, but rather the entire trajectory. The life course of an individual is then seen as an ordered sequences of states.

In a social science perspective, it is crucial not only to describe the different aspects of the sequence, but also to study the association between the sequence and some explanatory variables in order to test specific hypotheses. The most common approach relies on (deterministic) cluster membership. In short, based on a dissimilarity measure, the sequences are clustered in homogeneous groups. Then a regression model is used to study the association between some covariates of interest and cluster membership. The cluster-based strategy, while easy to perform, is based on the relatively strong assumption that the variability of the sequences in each cluster is potentially non-informative and therefore it might be ignored. If this assumption do not hold, as often in applications in social sciences, the deterministic cluster-based procedure may lead to incorrect conclusions.

The present study discusses two alternative approaches to better investigate the effects of explanatory factors on the heterogeneity of the observed sequences: the implicative approach and a discrepancy based one. Moreover, we show data visualisation tools to easily display the results of the two proposed approaches. We illustrate the two approaches investigating the association between family life sequences in early adulthood (age 16 to 30) and potential covariates such as: gender, level of education, cohort and origin. Based on retrospective information (data from the Swiss Household Panel), we constructed yearly trajectories (sequences of states) on more than 40,000 individuals distinguishing between eight types of living arrangement: living alone, with parents, with partner, with others (e.g., flat-sharing), with partner and at least one child, with child(ren) but without partner, with parent and partner, with children and other, other living arrangements. We used such a detailed list of living arrangements to grasp the complexity of potential trajectories in early adulthood.

2 Implicative statistics

Implicative statistic [9] provides an estimation of the relevance of a rule such as "A implies B". This is achieving by measuring the gap between the expected and observed number of counter-examples of the rule. If the observed number of counter-examples is significantly lower than expected in the independence case, then the rule is statistically relevant. Technically, this gap and its statistical significance are computed using the adjusted residuals of a contingency table with a continuity correction [7]. In our context, this is done at each time point.

The implicative statistics can be easily and effectively represented graphically. The plot shown in Figure 1 represents at each time point t and for each state A

considered, the relevance of the rule (i.e. the gap). A higher value means that the rule is more relevant. The dashed horizontal lines represent the confidence thresholds of the rules. We consider that a rule is significant at the 5% level, if it is above the dashed horizontal line 95%. The rule is for instance in top panel of Figure 1: being a man (woman) implies being in state A at time t.

The plot shows that family formation, i.e., any living arrangement that consists of living with the partner and eventually having children, in early adulthood are particularly relevant among women (technically speaking for the rule "being a woman"). On the other hand the significance of the rule "being a man" is particularly strong for the states "living with parents" and only after age of 22 "living alone". The results than confirms the literature with women experiencing a transition our of parental home at earlier age than men. Plots like this allows to have a clear snapshot of differences across levels of the covariates but loses the longitudinal information since we look at sequences of transversal characteristics (e.g., at each time point). Thus, we cannot conclude that patterns like "Living alone - Partner - Partner with children" at early age is typical of women. Nevertheless, these plots are easy to read and may help to uncover important sequences characteristics of a given profile.

The next approach leverages the longitudinal perspective of the sequences using the discrepancy analysis.

3 Discrepancy based approach

The discrepancy analysis allows a direct analysis of the the strength of the sequence-covariate link (i.e., without any prior clustering) [8]. Defining discrepancy of the sequences in terms of their pairwise dissimilarities, the approach is a generalisation of the the principles of analysis of variance—either with a single (ANOVA) or a multifactor framework (MANOVA)—to measure and test the part of the discrepancy among sequences that can be explained by covariates. In practice the response variable is summarized by a dissimilarity matrix. For instance, one can assess what fraction of the differences between early family lives trajectories can be explained by gender and social stratification. It is a proper longitudinal approach as long as the underlying distance measure does (in our case we use the Hamming distance).

Discrepancy-based methods proposed in the literature so far allows only to identify if the associations are statistically significant or not but do not fully provide a way to give a meaning of such relationships. We show here how to use the residuals of discrepancy analysis to leverage the longitudinal nature of the data in interpreting the differences in early adulthood trajectories according to some factors of interest.

Discrepancy analysis with multiple explanatory factors [8] mainly relies on computing the total sum of squares $SS_T = tr\big(\mathbf{G}\big)$ where $\mathbf{G}$ is the Gower matrix [5, 6]. As conventional, we can re-write the total sum of squares as the sum of between sum of squares ($SS_B = tr\big(\mathbf{HG}\big)$) and the residual within sum of squares ($SS_W = tr\big[(\mathbf{I}-\mathbf{H})\mathbf{G}\big]$), where $\mathbf{H}$ is the hat matrix used in traditional linear model.

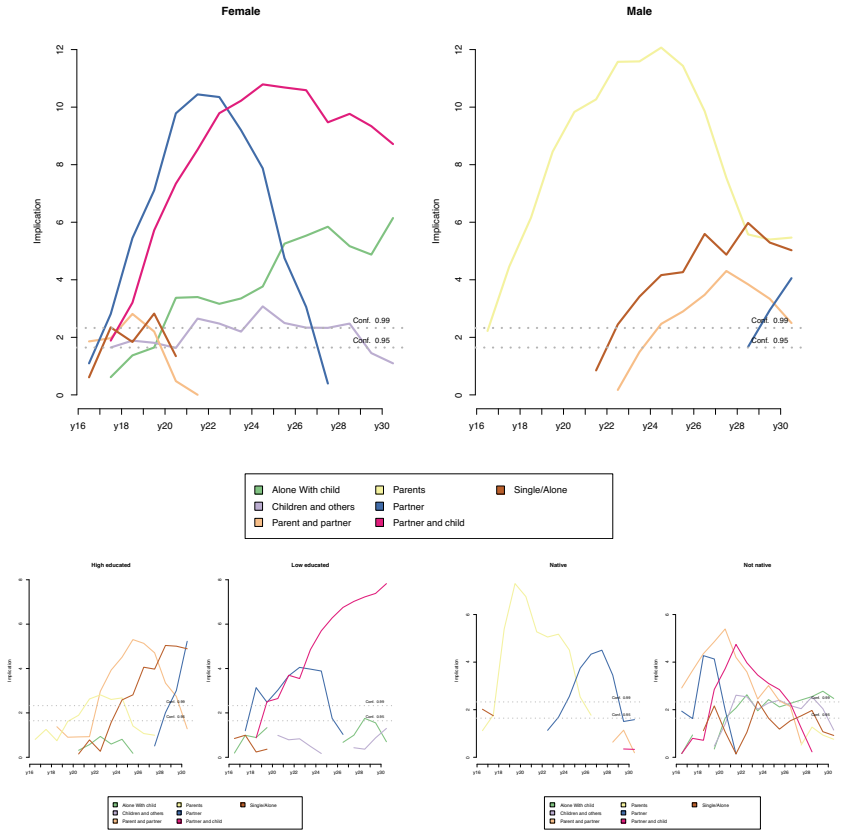


Fig. 1 Evolution of life trajectories in early adulthood according to explanatory factors.

For brevity, we cannot here discuss in details the mathematical formulations used in the general discrepancy analysis but interested readers can refer to [8].

The diagonal elements of the Gower matrix can be interpreted as the contribution of each sequence to the *overall* discrepancy, and also as the distance between a sequence and the gravity center of a set of sequences [8]. We will interpret those values as residuals: a high value means that a sequence is far from its gravity center, and a low value means that the sequence is close to it. Similar to the residual of a null model in a regression setting, these are residuals without accounting for any covariate. In a similar manner, the diagonal elements of the matrix $(\mathbf{I}-\mathbf{H})\mathbf{G}$ can be interpreted as the contribution of each sequence to the *residual* discrepancy. In a linear model, those values could be interpreted as the square of the residual of the fitted model (i.e, including explanatory factors).

By combining these two types of residuals (i.e. contribution to the overall and residual discrepancy), we can identify the sequences that are best represented when

we include a covariate in the model. These are the diagonal elements of the matrix $\mathbf{G} - \left[(\mathbf{I} - \mathbf{H})\mathbf{G}\right] = \mathbf{HG}$. We will call the diagonal elements of this matrix *gain*. A positive value means that the corresponding sequence is better represented if the explanatory factor is taken into account whereas a negative value means that the sequence is less well represented. In practice, we can compute a value for each sequence representing the gain (or loss) of information if we take a covariate into account. Both the "effect" of a single or of multiple factors can be analysed.

Figure 2 presents two index-plots of the sequences separated according to level of education and ordered according to the gain when taking education and the others explanatory factors (gender, cohort, being native) into account. On the right of each index plot, we have plotted the gain associated with each sequence. Sequences at the bottom have the highest gain, whereas sequences at the top have negative gain. The sum of the gains values equals to the explained sum of squares.

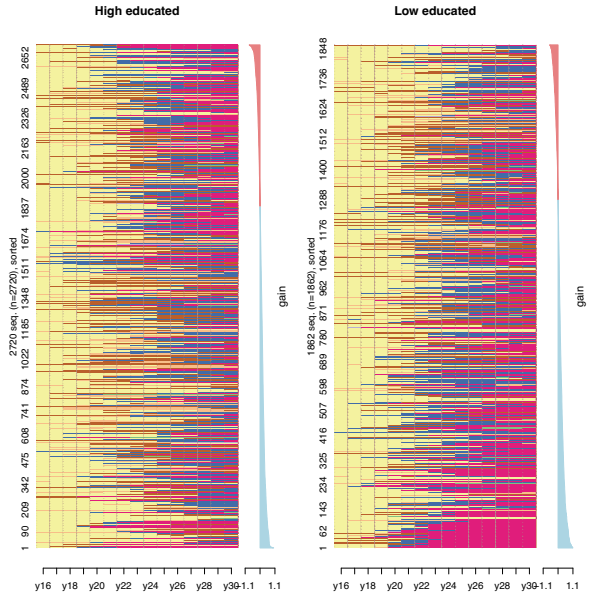


Fig. 2 Sequences ordered according to gain by low and highly educated individuals when taking into account also gender, cohort and being born in the country.

Looking at the sequences at the bottom of the graph, we can notice that after taking into account the effect of gender, cohort and origin, the gravity center of the sequences of those with high level of education has moved in the direction of a predominance of "red" state (having a partner of children) at an early age. This means that we observe a *tendency* of low educated of leaving the parent home and

start their own family at an earlier age than those highly educated. Similarly, it is possible to plot the effect for each single explanatory factor considered.

4 Conclusions

The study discussed two ways to study the relationship between sequence of states and some explanatory factors of interest in a sequence analysis framework. Going beyond the traditional cluster based approach, we have shown how different tools can be used to focus either cross-sectionally on the typical states of a subpopulation (for instance women, as opposed to men) measuring the gap between the expected and observed numbers of counter examples (implicative statistics) or more importantly leveraging the longitudinal nature of trajectories using a discrepancy-based approach. This work focuses on sequences of states taking an holistic approach but it worth mentioning that the study of trajectories can be done using step-by-step probabilistic approaches such as Markov based models [1, 3, 4] where it is straightforward to include both time-varying and time invariant explanatory factors in the model strategy.

Acknowledgements Danilo Bolano acknowledges financial support from EU H2020 project DisCont Discontinuities in Household and Family Formation (PI: Francesco C. Billari)

References

1. Bartolucci, F., Farcomeni, A., and Pennoni, F. Latent Markov Models for Longitudinal Data. Boca Raton, FL: Chapman & Hall/CRC Press. (2013)
2. Billari, F.C. and Piccarreta, R. Analyzing demographic life courses through sequence analysis. Mathematical population studies, **12**(2), pp.81-106 (2005)
3. Bolano, D. and Berchtold, A. (2016). General framework and model building in the class of Hidden Mixture Transition Distribution models, Computational Statistics & Data Analysis, **93**, pp. 131-145. (2016)
4. Bolano, D. Handling Covariates in Markovian Models with a Mixture Transition Distribution Based Approach. Symmetry. (2020)
5. Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53** (3/4), 325338. (1996)
6. Gower, J. C. Euclidean distance geometry. Mathematical Scientist **7**, 114. (1982)
7. Ritschard, G. De lusage de la statistique implicative dans les arbres de classification. In R. Gras, F. Spagnolo, and J. David (Eds.), Actes des Troisimes Rencontres Internationale ASI Analyse Statistique Implicative, Volume Secondo supplemento al N.15 of Quaderni di Ricerca in Didattica, pp. 305314. Palermo: Universit degli Studi di Palermo. (2205)
8. Studer, M., G. Ritschard, A. Gabadinho, and N. S. Muller. Discrepancy analysis of state sequences. Sociological Methods and Research **40**(3), 471510. (2011)
9. Suzuki, E. and Kodratoff Y. Discovery of surprising exception rules based on intensity of implication. In J. M. Zytkow and M. Quafafou (Eds.), Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD 98, Nantes, France, September 23-26, Proceedings, pp. 1018. Berlin: Springer. (1998)

The relationship between economic news and fertility: the case of Germany

La relazione tra notizie economiche e fecondità: il caso della Germania

Maria Francesca Morabito, Raffaele Guetto, Matthias Vollbracht, and Daniele Vignoli

Abstract This work contributes to the increasing literature on the relationship between the media coverage of the economy and fertility dynamics. We address this relationship for Germany, a nation with a prosperous economy, by combining individual-level data from the German Socio-Economic Panel (G-SOEP) and Media Tenor data on economic news reported by some German weekly magazines of great circulation. Our analysis scrutinizes the association between economic news items and fertility behaviour. Findings reveal that an increase in negative news on the state of the economy discourages fertility, while positive news items only exert a negligible effect.

Abstract *Il presente studio contribuisce alla crescente letteratura riguardante la relazione tra la copertura mediatica dell'economia e la fecondità. Ci focalizziamo sulla Germania, un paese caratterizzato da un'economia florida, combinando dati individuali derivanti dal German Socio-Economic Panel (G-SOEP) con dati Media Tenor relativi alle notizie economiche riportate da alcuni settimanali tedeschi di grande tiratura. La nostra analisi esplora la correlazione tra notizie economiche e fecondità individuale. I risultati mostrano che un aumento di notizie negative scoraggia la fecondità, mentre le notizie positive esercitato un effetto trascurabile.*

[1] Maria Francesca Morabito, Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; email: mariafrancesca.morabito@unifi.it

Raffaele Guetto, Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; email: raffaele.guetto@unifi.it

Matthias Vollbracht, Media Tenor International, Talacker Strasse, 41, 8001 Zurich, Switzerland; email: m.vollbracht@mediatenor.com

Daniele Vignoli, Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; email: daniele.vignoli@unifi.it

Key words: Fertility; Media; Germany; Economy; Economic News

1 Introduction

A common, albeit heterogenous, decline in fertility levels is investing several countries of Western Europe since 2010. It is involving even countries of North Europe, as well as some countries of the Continental area (e.g., Belgium and France), which did not experience the negative consequences of the Great Recession as harshly as the Southern European countries. In this scenario, the relative growth of the German Total Fertility Rate (TFR) was of 15% during 2010-2019, and represents an exception compared to declines of 8%, 14%, 21%, and 28% in France, Belgium, Norway, and Finland, respectively. As a result of these opposite dynamics, Germany's TFR in 2019 (1.53) reached, or surpassed, that of Norway (1.53) and Finland (1.35), traditionally "high-fertility" countries which are characterized by cultural, institutional, and economic conditions relatively more favourable for childbearing compared to other European areas.

The current puzzled relationships between institutional arrangements, gender ideologies, and economic situations, on one side, and fertility levels, on the other, lead to research into what other (new) forces are involved in shaping fertility choices. A recent line of research suggests the idea that the public discourse around the state of the economy, driven by economic narratives conveyed by the media and fuelled by interactions with parents and peers, is crucial in influencing fertility behaviour in conditions of uncertainty, over and above the (not-negligible) effect of individual and contextual objective economic circumstances [17, 18]. This view is consistent with a new generation of studies emphasizing the importance of subjective perceptions of the economic context as crucial fertility drivers [10, 19].

Media-conveyed narratives consist of news, debates, and all information reported by print news, television, and radio, as well as that available online. Previous studies proved the causal impact of media-conveyed narratives about economic uncertainty and the COVID-19 pandemic on marriage and fertility intentions via laboratory experiments [7, 8, 20]. And lastly, Guetto et al. [6] have documented a robust association between the incidence and tone of economic stories reported by the evening newscast (*TG1*) of Italy's most-viewed TV channel (*Rai 1*) and the probability of conception.

In the present study, we scrutinize the effects of the media coverage of the economy on individuals' fertility behaviour in Germany, net of micro- and macro-level indicators of the objective economic context. This research extends the empirical evidence to the German case, where fertility and economic conditions have had diametrically opposed trends compared to Italy, the only country for which, to the best of our knowledge, micro-level evidence on the relation between the media coverage of the economy and fertility is available [6].

2 Background

According to the agenda-setting theory [13], the media may exert their influence on the individual perception of reality through three main processes: the selection of the contents to report, the salience reserved to them, and, last but not least, the tone and the perspectives through which information is framed [5].

Many studies have addressed the relationship between the media coverage of the economy and economic perceptions of German citizens. Brettschneider [1] found a strong correlation between the proportion of citizens that consider the unemployment as the Germany's most important problem, and the share of news on unemployment. Lamla and colleagues showed that increases in the news coverage about current and future inflation levels improve the accuracy of German consumers' expectations, whereas it worsens when the media overemphasize negative facts [11, 12]. Furthermore, Garz found a correlation between the share of news on labour market policies and perceived job insecurity [3], and between the number of words related to unemployment reported in the news items and households' economic perceptions [4].

Nevertheless, we did not find any study addressing the relationship between the media coverage of the economy and fertility in Germany. At the area-level, Schneider [14] found a significant association between an indicator of the press coverage of the Great Recession and general fertility rates in the US, net of objective economic indicators. At the individual level, the causal effect of media-conveyed narratives of economic uncertainty on fertility intentions has been recently estimated for Italy and Norway [20] through controlled experiments in which participants were exposed to a (mock) news bulletin on the conditions of the national economy in the next three years. More related to this research, Guetto et al. [6] documented that the number of negative economic items reported by the main Italian newscast is negatively associated with fertility, whereas the number of positive news items has an opposite correlation. Furthermore, an increase in the percentage of economic news out of all news has been found to reduce fertility, even if this effect decreases as the tone of media narratives improves.

These findings are limited to a context of low fertility levels and strong economic hardship as that of Italy, but the outlined relationships may potentially vary among countries. Incidentally, the effect of new information on personal perceptions also depends on the standard to which individuals are accustomed [15]. We hypothesize that in countries characterized by prosperous economies, negative economic narratives may have a stronger impact than positive ones, since they are far from the imaginary of citizens, and vice versa. Furthermore, negative news items have proved to be more influential than positive ones on economic perceptions and expectations in wealthy countries, such as Germany [2] and Sweden [11]. Vignoli and colleagues found a stronger impact of negative economic narratives on fertility intentions in Norway, while positive economic narratives have been proved to be more influential in Italy [20]. Consistently, the (positive) effect of an increase in the number of positive news about the economy on fertility behaviour has been found to be stronger than that of an increase in negative news in Italy [6]. Thus, we may

expect that Germans, used to receive reassuring news on the economy, may be more influenced by a worsening of the media coverage of the state of the economy than to improvements.

3 Data and methodology

Media coverage data are provided by *Media Tenor International* and include the monthly number of economic news reported in German weekly magazines of great circulation (*Spiegel*, *Focus*, *Bild am Sonntag*). To each news item is assigned a tone (positive, negative, or no clear) according to explicit language or implicit evaluations reported in the news text. We merged these data with individual information from G-SOEP [9, 16]. We kept observations for female respondents aged 16-40. The monthly panel dataset covers the period Jan 2002 – Apr 2018 and includes 10,788 women (also belonging to the same household). Out of them, we observe 2,932 conceptions.

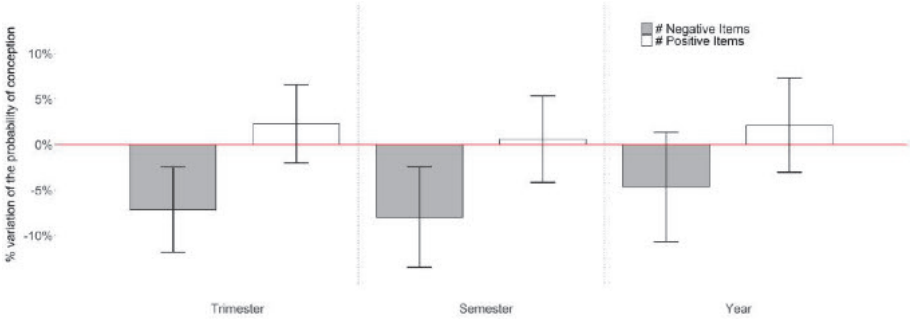
We applied Linear Probability Models with Fixed Effects (at both individual and household levels) to test if and to what extent changes in the media coverage of economic news are correlated to (within-individual) monthly variations of fertility. The dependent variable is a dummy that assumes value 1 in the month of conception of each child, and 0 otherwise. The main independent variables are measures of the coverage of the economy in German weekly magazines. Finally, models control for well-established micro-level fertility predictors (age, level of education, individual labour earnings, participation in education, and type of employment), a linear time trend, and macroeconomic indicators typically employed in fertility studies (such as the growth rate of real GDP, and the female unemployment rate).

4 Findings and discussion

We report the variation in the monthly probability of conception, compared to the mean risk observed in the sample (0.42%), associated to one standard deviation increase in the number of negative and positive news stories (Figure 1). The average numbers of economic news reported over the previous trimester, semester, and year, are included separately in three models, including the same control variables

Results reveal that the average amount of negative news is negatively correlated with the monthly probability of conception. On the other hand, the association between the moving average of the number of positive news and fertility is almost null (and not significant). The impact of negative items is more intense than that of positive ones, and statistically significant (except for the measure calculated on the coverage of the previous year). As shown in Figure 1, these effects are substantially relevant, controlling for micro- and macro-level indicators of objective economic conditions.

Figure 1: Percentage variation of the monthly probability of conception, compared to the mean risk observed in the sample, associated to one standard deviation increase in the media coverage of the economy in the previous trimester, semester and year.



Source: our elaboration on G-SOEP and Media Tenor data.

This is the first study documenting the association between economic news coverage and individual fertility in Germany, and the second addressing this relationship in a European country. This contribution is certainly not exhaustive on the subject, and further studies are needed. Indeed, in our models we do not control for changes in the fruition of magazines over time (nor for variations in the sensitivity of the consumers). However, we consider leading news magazines which also make use of online platforms to convey their contents. Furthermore, the potential influence of economic news coverage in weeklies magazines may result beyond their actual circulation. Items reported by media may also reach those who are not directly exposed to them, by fuelling the wider public discourse about the state of the economy.

References

1. Brettschneider, F.: Reality bytes: Wie die medienberichterstattung die wahrnehmung der wirtschaftslage beeinflußt. In: Falter, J., Gabriel, O.W., Rattinger, H. (eds.) Wirklich ein volk?, pp. 539-569, Leske+Budrich, Opladen (2000)
2. Dräger, L.: Inflation perceptions and expectations in Sweden – are media reports the "missing link"? Oxf. Bull. of Econ. and Stat. **77**, 681--700 (2015)
3. Garz, M.: Job Insecurity Perceptions and Media Coverage of Labor Market Policy. J. of Labor Res. **33**, 528--544 (2012)
4. Garz, M.: Effects of unemployment news on economic perceptions – Evidence from German Federal States. Reg. Sci. and Urban Econ. **68**, 172--190 (2018)
5. Goffman, E.: Frame analysis: An essay on the organization of experience. Harvard University Press, Cambridge (1974)
6. Guetto, R., Morabito, M.F., Vignoli, D., Vollbracht, M.: Media coverage of the economy and fertility (Working Paper No. 12). University of Florence, Department of Statistics, Computer Science, Applications "G. Parenti", Florence (2021a)
7. Guetto, R., Vignoli, D., Bazzani, G.: Marriage and cohabitation under uncertainty: the role of narratives of the future during the COVID-19 pandemic. Eur. Soc. **23**, S674--S688 (2021b)
8. Guetto, R., Bazzani, G., Vignoli, D.: Narratives of the future and fertility decision-making in uncertain times. An application to the COVID-19 pandemic. Vienna Yearb. of Popul. Res. **20**, 1--38 (2022)

9. IAB-SOEP Migration Samples, data of the years 2013-2019 (2019) doi: 10.5684/soep.iab-soep-mig.2019
10. Kreyenfeld, M.: Uncertainties in female employment careers and the postponement of parenthood in Germany. Eur. Sociol. Rev. **26**, 351--366 (2010)
11. Lamla, M.J., Lein, S.M.: The role of media for consumers' inflation expectation formation (Working Paper No. 201). Swiss Economic Institute, Zurich (2008)
12. Lamla, M.J., Sarferaz, S.: Updating inflation expectations (Working Paper No. 301). KOF Swiss Economic Institute, Zurich (2012)
13. McCombs, M.E., Shaw, D.L.: The agenda setting function of mass media. Public Opin. Q. **36**, 176--184 (1972)
14. Schneider, D.: The Great Recession, fertility, and uncertainty: Evidence from the United States. J. of Marriage and Fam. **77**, 1144--1156 (2015)
15. Schwarz, N., Bless, H.: Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In: Martin, L.L., Tesser, A. (eds.) The construction of social judgments, pp. 217-245. Erlbaum, Hillsdale (1992)
16. Socio-Economic Panel, data for years 1984-2019, SOEP-Core v36, EU Edition (2021) doi: 10.5684/soep.core.v36eu
17. Vignoli, D., Bazzani, G., Guetto, R., Minello, A., Pirani, E.: Uncertainty and Narratives of the Future. A Theoretical Framework for Contemporary Fertility. In: Schoen, R. (ed.) Analyzing Contemporary Fertility, pp. 25-47. Springer, Berlin (2020a)
18. Vignoli, D., Guetto, R., Bazzani, G., Pirani, E., Minello, A.: A Reflection on Economic Uncertainty and Fertility in Europe: The Narrative Framework. Genus **76**, 28 (2020b)
19. Vignoli, D., Mencarini, L., Alderotti, G.: Is the Effect of Job Uncertainty on Fertility Intentions Channeled by Subjective Well-Being? Advances in Life Course Res. **46**, 100343 (2020c)
20. Vignoli, D., Minello, A., Bazzani, G., Matera, C., Rapallini, C.: Narratives of the Future Affect Fertility: Evidence from a Laboratory Experiment. Eur. J. of Popul. **38**, 93--124 (2022)

Leaving home among Millennials in Italy: does economic uncertainty matter?

L'uscita dalla famiglia di origine fra i Millennials in Italia: il ruolo dell'incertezza economica

Silvia Meggiolaro and Fausta Ongaro

Abstract This paper analyses whether the process of leaving home is changing among recent cohorts of young Italians. We focus on the changes with reference to: a) timing and reasons of exit, and b) role played by (individual) economic uncertainty. We used retrospective data from the pooling of two rounds of a survey conducted by ISTAT in 2009 and 2016 and applied event history models. Results show that, in comparison with previous cohorts, both male and female of younger cohorts have a lower probability of leaving parental home to live with a partner, but, whereas this lower propensity persists for men, whatever the exit, women have a higher likelihood of leaving home for non-union related reasons. Younger cohorts also show a differentiated role of economic uncertainty.

Abstract *Il lavoro esamina le modalità di uscita dalla famiglia di origine dei giovani italiani nell'ipotesi che le coorti più recenti (Millennials) presentino comportamenti diversi da quelle precedenti per modalità di uscita e per ruolo svolto dall'incertezza economica individuale. A tale scopo si utilizzano dati retrospettivi delle due indagini condotte dall'ISTAT nel 2009 e 2016, analizzati con modelli event-history. I risultati mostrano che, tra i Millennials, gli uomini ritardano ulteriormente le uscite sia per unione che per altre ragioni, mentre le donne ritardano solo le uscite per unione e invece tendono ad anticipare quelle per altri motivi. Inoltre, la propensione a uscire dalla famiglia di origine sembra essere influenzata diversamente dall'incertezza economica individuale*

Key words: leaving home, economic uncertainty, event history models.

[1]Silvia Meggiolaro, Department of Statistical Sciences, University of Padova, meg@stat.unipd.it

Fausta Ongaro, Department of Statistical Sciences, University of Padova, ongaro@stat.unipd.it

1 Introduction

Leaving the parental home is a key event to be studied, since it is one of the markers in the process of transition to adulthood and the pathway through which it has been experienced may have important consequences for young adults' life course outcomes. Given the importance of the leaving home process of young adults, a quite rich literature has analysed the determinants of the decision to leave the parental home in Europe (e.g., Iacovou 2010; Schwanitz, Mulder and Toulemon, 2017), showing that – among the numerous (individual, family, contextual) factors influencing the decision of establishing an independent household - the economic ones may play an important role (Blaauboer and Mulder, 2010; Iacovou 2010; Bertolini et al., 2017; Schwanitz, Mulder and Toulemon, 2017).

This paper aims at studying whether and how the process of leaving home is changing among recent cohorts of young Italians (for which data are available, i.e. those born during the '80s and the first '90s), the so-called *Milllennials*, with reference to a) the timing and reasons of exit, and b) the role played by (individual) economic uncertainty on the timing and the reasons of exit.

But why is Italy worthwhile to be considered? Traditionally, Italian young adults tend to leave home at later ages compared to their counterparts in other European countries (Ongaro, 2001) and mainly for union-related reasons (marriage). However, there are not studies focusing on more recent cohorts. As previous cohorts of young Italians, indeed, also recent cohorts live in a context characterized by a persistent familistic welfare, but in comparison with older cohorts, Milllennials are also living in a context of relevant changes: from one hand the economic structural (labour market reforms of 1990) and conjunctural (the 2008-2013 Economic Crisis) changes; from the other hand, cultural changes, such as those determined by the diffusion of the Second Demographic Transition family behaviours, by the globalization, and by decreasing gender inequalities.

Does the leaving home of Italian Millennials differ from that of previous cohorts with respect to timing or reasons for leaving home? Does the individual economic uncertainty of recent cohorts play a role in the possible further postponement? Is there a different role of individual economic uncertainty for leaving home among Millennials in comparison with previous ones? Do young men behaviours differ from those of young women with respect to this points? To answer these questions, we used retrospective data coming from the pooling of two rounds of the survey 'Families and Social Subjects' conducted in Italy by the Italian Statistical Institute (ISTAT) in 2009 and 2016 and applied event history models (in some case in their competing risk version), distinguishing among the gender of the young people.

2 Data and strategy of analysis

The retrospective data on leaving home obtained from the survey 'Families and Social Subjects' allow to consider individuals who at the time of interview (in 2009

and in 2016, respectively, for the two rounds of the survey) were between 20 and 44 (and thus, born between 1965 and 1996). Using this sub-sample of population, we first studied the process of leaving home as a single destination, and then we considered two different pathways out of the parental home: leaving home to live without a partner versus leaving home to live with a partner.

Following previous studies (for example, Blaauboer and Mulder, 2010; Schwanitz, Mulder and Toulemon, 2017), process time starts at age 16 and ends at age 34. Cases were censored at the time of interview or at age 35 when the transition out of the parental home had not been made by that age. In this way, our sample totals 22,218 respondents (50.5% women and 49.5% men).

Since only the year of leaving parental home is recorded, discrete-time event history models were estimated: this required adopting a person-year scheme. Separate analyses were conducted for men and women. In a preliminary analysis, we considered the process of leaving home as a single destination (and thus a logistic regression of person-years is estimated); then, a competing risks approach was used (multinomial logistic regression), where leaving the parental home to live without a partner and leaving home to live with a partner are the outcomes of interest.

As regards the covariates, in this paper, we focused our attention on two core explanatory variables: the cohort of individuals and the "labour market situation", a (time-varying) variable which distinguishes four positions according to the employment status and the type of employment: not employed individuals (including inactive and unemployed individuals), temporary contract employees (those with fixed term or project contract), permanently employees (employed with a permanent contract), and self-employed individuals (including young entrepreneurs and freelancers). In addition, we considered the school enrolment, a time dependent covariate that has to be used in connection with the employment status to properly isolate the cases of unemployed and inactive individuals.

Other individual characteristics controlled for in our analyses were the age of individuals and the level of education. We considered also some family characteristics, namely, parental education, parental separation and maternal employment status during the respondent's adolescence (Schwanitz et al., 2017). Lastly, we considered the area of residence (at the interview) as a contextual characteristic to take into account the geographical heterogeneity and the impact of unobservable context variables (Sironi and Rosina, 2015).

The strategy of analysis started with models 1, which include only the cohort as a covariate (besides the age of individuals included in linear as well as in logistic form); then models 2 add the key covariates (the labour market situation and the student status) and all the control covariates; lastly, models 3 add the interaction term between the key covariates and the cohort.

3 Results

In a preliminary analysis (here not reported) we considered the process of leaving home as a single destination, in which we did not make any distinction of the route

taken out of the parental home. Results show that Millennials have a significant lower likelihood of leaving parental home with respect to previous cohorts for men, but not for women. As regards the role of economic uncertainty, if for young women only a situation of unemployment decreases the likelihood of leaving home, both among older and younger cohorts, for male Millennials some signals of new patterns seem to emerge: for them, unemployment continues to have a negative effect for the risk of departure from home, but with a less strong effect in comparison with previous cohorts. In addition, a condition of economic uncertainty such as defined by having a temporary contract employment increases the probability of leaving home in comparison with that of previous cohorts.

In fact, results of competing risks models describe a quite different situation (Table 1a and 1 b, for men and women, respectively). Results (mod.3), first, suggest that Millennials show a new pattern in the process of leaving parental home in comparison with previous cohorts: both male and female Millennials have a lower likelihood of leaving parental home to live with a partner (the odds of leaving for male Millennials are 22 percent ([1-exp(-0.25)]*100) lower than those of older cohorts; for female Millennials, 15% lower), but, whereas this lower propensity persists considering non-union related exits for men, at the opposite, female Millennials have a higher likelihood of leaving parental home for non-union related reasons in comparison with previous cohorts.

Second, younger cohorts also show a differentiated role of economic uncertainty on the decision about whether to move out parental home. As regards the likelihood of leaving parental home to live with a partner, economic uncertainty plays a similar role among younger and previous cohorts: unemployment lowers the likelihood of leaving parental home with a partner for young men, but not for women, in comparison with that of young adults with a permanent employment (thus confirming Italy as a conservative country with gender differentiated effect of unemployment, Gousia et al., 2021); moreover, experiencing precarious employment conditions - such as those defined by temporary contract - lowers the risk of departure from home with a partner both for men and women (being employed with a temporary contract decreases the odds of leaving by 19 percent = [1-exp(-0.21)]*100 in comparison with being permanently employed for men, and by 11 percent for women).

The situation is completely different considering the exit from parental home for non-union related reasons. In this case, the unemployment, but also an uncertain position such as defined by a self-employment, decreases the probability of leaving parental home both for men and for women, but, these effects are less strong for male Millennials in comparison with previous cohorts. On the opposite, with respect to a permanent contract, a situation of uncertainty as that defined by a temporary contract increases the likelihood of leaving home without a partner both for men and women, and – this effect seems to increase among the male Millennials (the interaction term being positive and significant for women).

Table 1a. Leaving the parental home to live with a partner or without a partner: discrete-time competing risks models. Men

	Model 1		Model 2**		Model 3**	
	With a partner	**Without a partner**	**With a partner**	**Without a partner**	**With a partner**	**Without a partner**
Age	-0.09***	-0.15***	-0.07***	2.18***	-0.08***	-0.16***
Age logged	2.75***	1.16***	-0.15***	1.17***	2.22***	1.06***
Cohort (ref: 1965-79)						
1980-96	-0.37***	0.06*	-0.25***	-0.02	-0.25***	-0.26***
Labour market situation TV* (ref: permanently employed)						
Temporary employed			-0.20***	0.29***	-0.21***	0.17**
Self employed			0.02	-0.28***	0.03	-0.46***
Not employed			-1.00***	-0.53***	-1.02***	-0.57***
Education (ref: out of education)						
Student			-0.60***	-0.06	-0.53***	-0.17**
Interactions						
1980-96 cohort * labour market situation						
Temporary employed					0.03	0.34***
Self employed					-0.05	0.57***
Not employed					0.10	0.16*
1980-96 cohort* student					-0.27	0.21**

TV* time-varying; ** controlled for respondent's education, family characteristics and area of residence

Table 1b. Women

	Model 1		Model 2**		Model 3**	
	With a partner	**Without a partner**	**With a partner**	**Without a partner**	**With a partner**	**Without a partner**
Age	-0.10***	-0.09***	-0.09***	-0.09***	-0.08***	-0.10***
Age logged	1.90***	0.94***	1.61***	0.93***	1.59***	0.94***
Cohort (ref: 1965-79)						
1980-96	-0.37***	0.42***	-0.19***	0.26***	-0.16***	0.27***
Labour market situation TV* (ref: permanently employed)						
Temporary employed			-0.14***	0.09	-0.12*	0.19*
Self employed			-0.02	-0.25**	0.00	-0.33**
Not employed			-0.02	-0.48***	-0.04	-0.46***
Education (ref: out of education)						
Student			-1.23***	0.28***	-1.04***	0.23***
Interactions						
1980-96 cohort * labour market situation						
Temporary employed					-0.04	-0.21
Self employed					-0.11	0.22
Not employed					0.09	-0.06
1980-96 cohort* student					-0.58***	0.09

TV* time-varying; ** controlled for respondent's education, family characteristics and area of residence

Thus, we can therefore argue that in a society where economic uncertainty is becoming a common condition, young adults are finding new strategies to

autonomy. This does not hold for a more traditional way of leaving home, such as that to live with a partner, for which, unemployment or economic uncertainty associated to a temporary job continues to have a detrimental effect: probably young people are less likely to experience a transition, such as in the case of entry into union, where responsibilities are greater than in the case of entry into independence.

4 Discussion and future research

This connection between economic uncertainty and responsibilities should be analysed more in depth with future studies considering more in details the different processes of leaving home, further disentangling the different reasons determining decision about whether to move out: the relatively low sample size of individuals in more recent cohorts does not allow more detailed analyses, but it would be worthwhile to distinguish leaving home to live with a partner in cohabitation from that in marriage, to verify whether the increase in the diffusion of non-marital unions bring also a new pattern of exit from the family of origin, implying less commitment than that implied by marriage and thus more compatible with uncertainty. Similarly, among individuals leaving parental home for non-union related reasons, those leaving to pursue educational opportunities should be distinguished from those leaving to pursue employment opportunities and from those leaving to establish independence, to verify the routes more favoured by economic uncertainty.

References

1. Bertolini, S., Bolzoni, M., Ghislieri, C., Goglio, V., Martino, S., Meo, A., Moiso, V., Musumeci, R., Ricucci, R., Torrioni, P.M. Labour market uncertainty and leaving parental home in Italy. In: Baranowska-Rataj, A., Bertolini, S., Goglio, V. (eds.) Country level analyses of mechanisms and interrelationships between labour market insecurity and autonomy. EXCEPT Working Paper No. 11, pp. 16-40, Tallin University, Tallin (2017).
2. Blaauboer, M., Mulder, C.H. Gender differences in the impact of family background on leaving the parental home. J. Hous. Built. Environ., 25(1): 53–71. doi:10.1007/s10901-009-9166-9 (2010).
3. Gousia, K., Baranowska-Rataj, A., Middleton, T., Nizalova, O. The impact of unemployment and non-standard forms of employment on the housing autonomy of young adults. Work Employ. Soc., 35(1), 157-177 (2021).
4. Iacovou, M. Leaving home: Independence, togetherness and income. Adv. Life Course Res., 5(4): 147–160. doi:10.1016/j.alcr.2010.10.004 (2010).
5. Ongaro, F. Transition to adulthood in Italy. In: Corjin, M., Klijzing, E. (eds.) Transition to Adulthood in Europe, pp. 173-208, Kluwer Academic Press, Bruxelles (2001).
6. Schwanitz, K., Mulder, C.H., Toulemon, L. Differences in leaving home by individual and parental education among young adults in Europe. Demographic Res. 37, 1975-2010 (2017).
7. Sironi E., Rosina A. Leaving the parental home in Italy during the economic crisis, Genus, 71, 199-216 (2015).

Adverse pregnancy outcomes in the United Kingdom following unexpected job loss

Esiti avversi della gravidanza nel Regno Unito a seguito della perdita inaspettata del lavoro

Alessandro Di Nallo and Selin Köksal

Abstract There is no evidence on the effect of a job loss on the extra risk of adverse pregnancy outcomes. We analyse conceptions, partnerships and work histories from the British survey "Understanding Society" to examine whether a woman's and/or her partner's involuntary job loss during a pregnancy increases the risk of a miscarriage and non-live birth. Our findings show that risk of non-live birth increases when the mother or the partner are exposed to an unexpected job loss. Moreover, we find higher risk of adverse outcomes following a job loss during the first trimester of pregnancy.

Abstract *Nella letteratura accademica, non esiste evidenza della relazione tra la perdita del lavoro e l'aumento del rischio di un esito avverso della gravidanza. Questo articolo analizza le gravidanze e le storie lavorative delle donne intervistate dall'inchiesta britannica "Understanding Society", ed esamina se la perdita involontaria del lavoro da parte di una donna e/o del suo partner durante una gravidanza aumenta il rischio di aborto spontaneo e morte alla nascita. I risultati mostrano che i rischi di esito avverso della gravidanza aumentano quando la donna e/o il suo partner sono esposti a una perdita imprevista del lavoro. Inoltre, troviamo un rischio maggiore di esiti avversi come risposta alla perdita di lavoro durante il primo trimestre di gravidanza.*

Key words: Pregnancy Loss, Miscarriage, Birth outcomes, Job Loss, Gender, UK

Alessandro Di Nallo
Dondena Centre for Research on Social Dynamics and Public Policy, and Department of Social and Political Sciences, Bocconi University, e-mail: alessandro.dinallo@unibocconi.it

Selin Köksal
PhD Candidate at the Department of Social and Political Sciences, e-mail: selin.koksal@unibocconi.it

1 Introduction

An estimated 11%-20% of clinically recognized pregnancies result in miscarriage. The risk of spontaneous termination of pregnancy is even higher for women who experience adverse economic events [3]. Besides somatic costs, spontaneous pregnancy losses (less than 20 weeks' gestation) are associated with an increased risk of long-term depression and anxiety among parents [1, 8]. Studies on stillbirths (above 28 weeks' gestation) signal that the incidence of spontaneous loss has remained relatively stable as compared with steady reductions in mothers' death at delivery, morbidity from childbirth and neonatal mortality [7]. This motivates international institutions' appeal to improve the understanding of preventable causes of pregnancy loss [27].

Many correlates of spontaneous pregnancy loss have been identified: maternal age, behavioural habits and genetic predisposition are among the most cited antecedents [28]. However, these and other known factors fail to explain most events of pregnancy loss [3]. Job market shocks and financial distress have been found to be negatively associated with a wide array of birth-related outcomes [9, 3, 16]. [2], [21] and [3] estimated the effect of economic downturns on perinatal outcomes at the ecological level. One ([18]) examined birth outcomes of couples who self-reported a husband's job loss. The evidence is mixed as some studies show adverse birth outcomes, such as live birth and birthweight [21, 24], and others display no or even better-than-expected results [5]. Some studies found an association between proxies of personal hardship and pregnancy loss in an ecological design [11, 23, 3].

However, it is not clear whether and how an *individual* level job loss plays any role in spontaneous pregnancy loss. To the best of our knowledge, no study has investigated whether a specific life shock at personal or couple-level, such as job loss, is a possible antecedent of miscarriage or non-live birth.

We contribute to existing literature in multiple ways. Firstly, we assess whether and to what extent a job loss of a woman or her partner's increases the risk of spontaneous pregnancy loss, represented by a miscarriage or a non-live birth. Secondly, we clarify whether the timing of a job loss during a pregnancy impacts on pregnancy outcomes, and if any disparities fall along socio-economic status. Thirdly, we use high-quality individual data on the spouses' socio-demographic characteristics and finely grained information on their and their partners' work histories as well as on their reproductive events. Unlike the extant literature, we are able to identify self-reported dates of conception and own or partner's non-voluntary job loss.

2 Data and Methods

We retrieve information on pregnancy outcomes from the UKHLS ('Understanding Society'), a population-representative survey of the United Kingdom spanning from 2009 to 2019. This dataset includes a wide array of information on conception date, duration of pregnancy, cause of pregnancy loss, delivery outcome. Consistent with

the economic literature, we retrieve data on individual employment, job characteristics, tenure and, if any, the voluntary or involuntary causes of a job loss.

We hypothesize that the number of spontaneous abortions would rise above the expected value among the pregnancies of women exposed to involuntary episodes of own or partner's job loss. We analyse associations between job loss and adverse pregnancy outcomes using two model specifications. We estimate a linear probability model of two outcomes: a miscarriage and a non-live birth. In the first model specification, the main explanatory variable is a dichotomous indicator of exposure to (a) woman's or (b) her partner's job loss during pregnancy. If conception and job loss are reported in the same month, we consider the couple to be exposed to the shock. In the second specification, the explanatory variable is distinguished in 6 categories to capture the *timing* of the job loss during over the gestational period. In chronological order, t_0 captures the job loss *preceding* conception by 4 to 1 month(s). The variables t_1, t_2, and t_3 identify, respectively, the job loss occurring during 1st, 2nd or 3rd trimester of gestation. t_4 identifies a job loss occurring between 10 and 24 months post-conception. Therefore, t_0 serves as a placebo test that signals if job displacement and a pregnancy loss are confounded because of pre-conception characteristics of the woman or her partner's.

The confounders include: woman's age (7 categories), ethnicity (8 categories), parents' SES at 16 (5 categories), whether the woman reported any miscarriage in the past, parity (first or second/higher order child), whether in a union (3 categories), time-varying educational attainment (3 categories), religious affiliation (7 categories), woman and her partner's job position (NSSeC-3 categories, lagged by 1 year). These controls are introduced in a step-wise approach. Of the 8,429 women who experience a pregnancy, 7,320 report a valid date of conception.

3 Results

Linear probability models clearly support the increased risk of a miscarriage (Table 1, columns 1-3) and non-live birth (Table 1, columns 4-6) following the exposure to women and partners' unexpected job loss. The statistical significance of the effect of women's job loss remains constant ($p < 0.01$) when accounting for an incremental set of controls. The *in utero* exposure to job loss increases the probability of miscarriage and non-live birth by 8.5 and 11.8 percentage points respectively. The impact of partner's job loss increases in magnitude ($p < 0.05$) when more controls are added. In this case, the probability of miscarriage and non-live birth increase by 8 and 8.3 percentage points respectively.

Linear models addressing the timing of job loss (Table 2) suggest that the influence is stronger in the first trimester both with respect to miscarriage (col. 1-3) and non-live births (col. 4-6). The *in utero* exposure to own and partner's job loss increases the probability of miscarriage and non-live birth by roughly 10 and 12 percentage points, respectively, for all specifications.

4 Discussion

The analysis suggests the existence of link between a couple's exposure to a non-volunary job loss and the risk of miscarriage and non-live birth. This link operates regardless of the partner hit by labour market shock and is highly sensitive to the timing with respect to conception. The first trimester of pregnancy is particularly susceptible to a job loss.

Table 1. Estimated effect of job loss on adverse pregnancy outcomes

Any job loss	Miscarriage (1)	Miscarriage (2)	Miscarriage (3)	Non-live birth (4)	Non-live birth (5)	Non-live birth (6)
Ref: No job loss						
Woman's job loss	0.087***	0.085***	0.085***	0.126***	0.118***	0.118***
	(0.018)	(0.018)	(0.018)	(0.034)	(0.034)	(0.034)
Partner's job loss	0.077**	0.080**	0.080**	0.068*	0.083**	0.083**
	(0.032)	(0.032)	(0.032)	(0.035)	(0.035)	(0.035)
Baseline controls	✓	✓	✓	✓	✓	✓
Own SES	✓	✓	✓	✓	✓	✓
Union characteristics		✓	✓		✓	✓
Partner's SES			✓			✓
Observations	7,230	7,230	7,230	7,230	7,230	7,230
R-squared	0.605	0.605	0.605	0.465	0.473	0.473

Baseline: ethnicity, parents' SES, age, religion, any prior miscarriage. Own SES: education, job status
Union characteristics: parity, partnership status or single. Partner's SES: education, job status
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2. Estimated effect of timing of job loss wrt conception on adverse pregnancy outcomes

Timing of job loss	Miscarriage (1)	Miscarriage (2)	Miscarriage (3)	Non-live birth (4)	Non-live birth (5)	Non-live birth (6)
Ref: -4/-1						
-12/-5	-0.022	-0.022	-0.022	-0.041	-0.041	-0.041
	(0.042)	(0.042)	(0.042)	(0.055)	(0.053)	(0.053)
0/3m	0.103***	0.101***	0.101***	0.134**	0.124**	0.123**
	(0.038)	(0.038)	(0.038)	(0.057)	(0.055)	(0.055)
4/6m	0.082	0.082	0.081	0.069	0.069	0.069
	(0.051)	(0.051)	(0.051)	(0.066)	(0.065)	(0.065)
7/9m	0.059	0.060	0.061	0.014	0.021	0.022
	(0.045)	(0.045)	(0.045)	(0.055)	(0.053)	(0.053)
11/24m	0.039	0.040	0.040	0.008	0.011	0.012
	(0.033)	(0.033)	(0.033)	(0.045)	(0.043)	(0.043)
No job loss/Other	0.003	0.002	0.001	-0.012	-0.019	-0.019
	(0.030)	(0.029)	(0.029)	(0.042)	(0.040)	(0.040)
Baseline controls	✓	✓	✓	✓	✓	✓
Own SES	✓	✓	✓	✓	✓	✓
Union characteristics		✓	✓		✓	✓
Partner's SES			✓			✓
Observations	7,230	7,230	7,230	7,230	7,230	7,230
R-squared	0.605	0.606	0.606	0.466	0.473	0.474

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Acknowledgements The authors gratefully acknowledge financial support from EU H2020 project "DisCont - Discontinuities in Household and Family Formation" (PI: Billari – ERC Grant no. 694262).

References

1. Blackmore, E. R., Côté-Arsenault, D., Tang, W., Glover, V., Evans, J., Golding, J. O'Connor, T. G. (2011), 'Previous prenatal loss as a predictor of perinatal depression and anxiety', The British Journal of Psychiatry 198(5), 373–378.
2. Bozzoli, C. Quintana-Domeque, C. (2014), 'The weight of the crisis: Evidence from newborns in argentina', Review of Economics and Statistics 96(3), 550–562.
3. Bruckner, T. A., Mortensen, L. H. Catalano, R. A. (2016), 'Spontaneous pregnancy loss in Denmark following economic downturns', American journal of epidemiology 183(8), 701–708.
4. Damaske, S. (2022), 'Gender, family, and healthcare during unemployment: Healthcare seeking, healthcare work, and self-sacrifice', Journal of Marriage and Family 84(1), 291–309.
5. Dehejia, R. Lleras-Muney, A. (2004), 'Booms, busts, and babies' health', The Quarterly journal of economics 119(3), 1091–1130.
6. Dooley, D. Prause, J. (2005), 'Birth weight and mothers' adverse employment change', Journal of Health and Social Behavior 46(2), 141–155.
7. Flenady, V., Koopmans, L., Middleton, P., Frøen, J. F., Smith, G. C., Gibbons, K., Coory, M., Gordon, A., Ellwood, D., McIntyre, H. D. et al. (2011), 'Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis', The lancet 377(9774), 1331–1340.
8. Frost, M. Condon, J. T. (1996), 'The psychological sequelae of miscarriage: a critical review of the literature', Australian and New Zealand Journal of Psychiatry 30(1), 54–62.
9. Gemmill, A., Catalano, R., Casey, J. A., Karasek, D., Alcalà, H. E., Elser, H. Torres, J. M.(2019), 'Association of preterm births among us latina women with the 2016 presidential election', JAMA network open 2(7), e197084–e197084.
10. Hatch, S. L. Dohrenwend, B. P. (2007), 'Distribution of traumatic and other stressful life events by race/ethnicity, gender, ses and age: A review of the research', American journal of community psychology 40(3-4), 313–332.
11. Hogue, C. J., Parker, C. B., Willinger, M., Temple, J. R., Bann, C. M., Silver, R. M., Dudley, D. J., Koch, M. A., Coustan, D. R., Stoll, B. J. et al. (2013), 'A population-based case-control study of stillbirth: the relationship of significant life events to the racial disparity for african americans', American journal of epidemiology 177(8), 755–767.
12. Howe, G. W., Levy, M. L. Caplan, R. D. (2004), 'Job loss and depressive symptoms in couples: common stressors, stress transmission, or relationship disruption?', Journal of Family Psychology 18(4), 639.
13. Inanc, H. (2018), 'Unemployment, temporary work, and subjective well-being: The gendered effect of spousal labor market insecurity', American Sociological Review 83(3), 536–566.
14. Jacobs, P. (1987), 'Chromosome abnormalities origin and etiology in abortions and livebirths', Human genetics.
15. Kearney, M. S. Levine, P. B. (2012), 'Why is the teen birth rate in the united states so high and why does it matter?', Journal of Economic Perspectives 26(2), 141–63.
16. Kramer, M. R., Hogue, C. J., Dunlop, A. L. Menon, R. (2011), 'Preconceptional stress and racial disparities in preterm birth: an overview', Acta obstetricia et gynecologica Scandinavica 90(12), 1307–1316.
17. Kramer, M. S., S eguin, L., Lydon, J. Goulet, L. (2000), 'Socio-economic disparities in pregnancy outcome: why do the poor fare so poorly?', Paediatric and perinatal epidemiology 14(3), 194–210.

18. Lindo, J. M. (2011), 'Parental job loss and infant health', Journal of health economics 30(5), 869–879.
19. Marcus, J. (2013), 'The effect of unemployment on the mental health of spouses–evidence from plant closures in Germany', Journal of health economics 32(3), 546–558.
20. Mare, R. D. (2016), 'Educational homogamy in two gilded ages: Evidence from intergenerational social mobility data', The ANNALS of the American Academy of Political and Social Science 663(1), 117–139.
21. Margerison, C. E., Luo, Z. Li, Y. (2019), 'Economic conditions during pregnancy and preterm birth: A maternal fixed-effects analysis', Paediatric and perinatal epidemiology 33(2), 154–161.
22. Margerison-Zilko, C. E., Li, Y. Luo, Z. (2017), 'Economic conditions during pregnancy and adverse birth outcomes among singleton live births in the united states, 1990–2013', American journal of epidemiology 186(10), 1131–1139.
23. Neugebauer, R., Kline, J., Stein, Z., Shrout, P., Warburton, D. Susser, M. (1996), 'Association of stressful life events with chromosomally normal spontaneous abortion', American Journal of Epidemiology 143(6), 588–596.
24. Noelke, C., Chen, Y.-H., Osypuk, T. L. Acevedo-Garcia, D. (2019), 'Economic downturns and inequities in birth outcomes: evidence from 149 million us births', American journal of epidemiology 188(6), 1092–1100
25. Scharber, H. (2014), 'Does "out of work" get into the womb? exploring the relationship between unemployment and adverse birth outcomes', Journal of health and social behavior 55(3), 266–282.
26. Upward, R. Wright, P. W. (2019), 'Don't look down: the consequences of job loss in a flexible labour market', Economica 86(341), 166–200. van den Berg, G. J., Paul, A. Reinhold, S. (2020), 'Economic conditions and the health of newborns: Evidence from comprehensive register data', Labour Economics 63, 101795.
27. WHO (2015), State of inequality: reproductive maternal newborn and child health: interactive visualization of health data, World Health Organization.
28. Wisborg, K., Kesmodel, U., Henriksen, T. B., Olsen, S. F. Secher, N. J. (2001), 'Exposure to tobacco smoke in utero and the risk of stillbirth and death in the first year of life', American journal of epidemiology 154(4), 322–327. 13

Bayesian modelling and inference 2

A Bayesian beta linear model to analyze fuzzy rating responses

Un modello bayesiano di regressione beta per l'analisi di risposte imprecise

Antonio Calcagnì, Massimiliano Pastore, Gianmarco Altoè, Livio Finos

Abstract In this short paper, we describe a Bayesian beta linear model to analyse imprecise rating responses. The non-random imprecision is extracted from crisp responses via the Item Response Theory tree (IRtree) method and it is represented by means of beta fuzzy numbers. The parameters of the beta linear model are estimated using the adaptive Metropolis-Hastings algorithm, with the fuzzy likelihood function being used as empirical evidence for the imprecise observations. A real case study is used to show the characteristics of the fuzzy beta regression model.

Abstract *Questo contributo descrive l'applicazione del modello di regressione beta nell'analisi di dati imprecisi. In questo contesto, l'imprecisione è riferita ad una fonte non casuale di incertezza ed è calcolata mediante il metodo fuzzy-IRTree. Il risultato di tale pre-trattamento è una collezione di insiemi fuzzy di tipo beta. I parametri del modello di regressione beta sono stimati mediante l'algoritmo Metropolis-Hastings di tipo adattivo mentre l'evidenza empirica dei dati è espressa mediante una funzione di verosmiglianza imprecisa. Il contributo si chiude con l'analisi di un caso studio.*

Key words: Fuzzy rating data, Beta linear model, Bayesian data analysis

---

Antonio Calcagnì,
DPSS, University of Padova, e-mail: antonio.calcagni@unipd.it
GNCS Research Group, National Institute of Advanced Mathematics (INdAM)

Massimiliano Pastore,
DPSS, University of Padova, e-mail: massimiliano.pastore@unipd.it

Gianmarco Altoè,
DPSS, University of Padova, e-mail: gianmarco.altoe@unipd.it

Livio Finos,
DPSS, University of Padova, e-mail: livio.finos@unipd.it

1 Introduction

Rating data are widespread across disciplines dealing with human-based measurements. In these cases, since the measurement process is based on cognitive actors, the collected data are often affected by non-random imprecision or fuzziness. This type of uncertainty has multiple origins, including the semantic aspects of the items being rated and the individual-level decision uncertainty underlying the response process [1]. To give an example, consider the situation where the item "I am satisfied with my life" is rated through a scale ranging from "strongly disagree" to "strongly agree". A stage-wise response process is usually involved in responding to these types of items. In particular, in a first step cognitive and affective information about the item being rated are retrieved and integrated (opinion formation stage) until the second decision stage is triggered, which includes the selection of the final rating response (e.g., "strongly disagree"). Because of the integration of conflicting cognitive and affective information about the item, fuzziness arises from the conflicting demands of the opinion formation stage [2]. Over the recent years, several fuzzy rating scales have been proposed to quantify fuzziness from rating data, including both direct/indirect fuzzy rating scales and fuzzy conversion scales (for an extensive review, see [1]). While direct fuzzy rating scales quantifies fuzziness by mapping response process to fuzzy numbers directly, fuzzy indirect scales aim at turning standard crisp ratings into fuzzy numbers by means of statistically-based procedures (e.g., see [3]). Unlike for the previous case, here the aim is to represent as much information as possible from the rating process in terms of a more complex number representation. Once fuzzy numbers have been obtained, they can be analysed either by means of standard statistical approaches or by adopting fuzzy statistical methods devoted to this purpose (e.g., see [4]).

In this contribution, we describe an application of a Bayesian beta linear model to the analysis of IRTree-based fuzzy data, a novel type of fuzzy responses which treat fuzziness in terms of decision uncertainty [1]. The remainder of this short paper is as follows. Section 2 describes the fuzzy beta data. Section 3 exposes the Beta linear model along with the parameter estimation procedure. Finally, Section 4 concludes this contribution by illustrating the application of the proposed method to a real dataset.

2 Data

IRTree-based fuzzy data represent a particular type of fuzzy numbers which are the output of a psychometric-based fuzzy conversion method (i.e., the Item Response Theory tree approach). In particular, they are computed in a way that the imprecision encapsulated into the matrix of crisp rating data $\mathbf{Y}_{n\times J}$ is mapped onto beta fuzzy numbers using all the rater's responses $\mathbf{y}_i$ to the J items being rated. The reader can refer to [1] for technical details about the conversion system. In this context, data consist of a collection of n (raters) $\times$ J (items) beta fuzzy numbers:

$$\widetilde{\mathbf{Y}} = (\xi_{\widetilde{y}_{11}}, \ldots, \xi_{\widetilde{y}_{1J}}, \ldots, \xi_{\widetilde{y}_{nj}}, \ldots, \xi_{\widetilde{y}_{nJ}})$$

where

$$\xi_{\widetilde{y}_{ij}}(y) = \frac{1}{C}\, y^{a-1}(1-y)^{b-1} \tag{1}$$
$$a = 1 + ms \quad \text{and} \quad b = 1 + s(1-m)$$

In this representation, $m \in (0,1)$ is the mode of the set, $s \in (0,\infty)$ is the precision of the set, whereas C is a constant ensuring that $\max_{y\in\mathcal{Y}} \xi_{\widetilde{y}}(y) = 1$. Note that $\xi_{\widetilde{y}_{ij}}$ is a normal and convex fuzzy set which lies in the interval $(0,1) \subset \mathbb{R}$. Because of the parametric representation involved by beta fuzzy numbers, the observed fuzzy data can be represented using two matrices, namely the matrix of modes $\mathbf{M}_{n\times J}$ and that of precisions $\mathbf{S}_{n\times J}$. Figure 1 shows an example of beta rating responses. It should be remarked that, in view of the fuzzy-IRTree representation adopted here, m represents the most plausible rating choice, s is the precision of m (i.e., smaller values indicate larger levels of hesitation in the rating choice), and $\xi_{\widetilde{y}_{ij}}$ codifies the decision uncertainty in terms of fuzziness (the larger the fuzziness, the highest the decision uncertainty). Ideally, in the case of no decision uncertainty, the fuzziness would vanish and the true rating realization would be precisely observed.

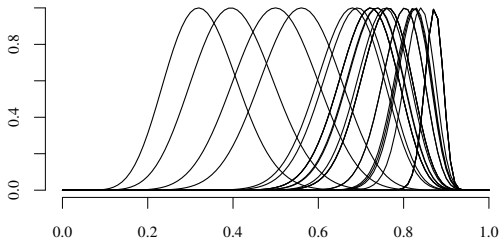


Fig. 1 An example of beta fuzzy responses.

3 Model and parameter estimation

To analyse fuzzy rating data we will adopt the Beta linear model proposed by [4], which is particularly well-suited for bounded rating responses. For a crisp collection of i.i.d. $(0,1)$-realizations $\mathbf{y} = (y_1, \ldots, y_n)$, the Beta density is as follows:

$$f_{\mathbf{Y}}(\mathbf{y};\mu,\phi) = \prod_{i=1}^{n} \frac{\Gamma(\phi_i)}{\Gamma(\phi_i\mu_i)\Gamma(\phi_i-\mu_i\phi_i)}\, y_i^{(\mu_i\phi_i-1)}(1-y_i)^{(\phi_i-\mu_i\phi_i-1)} \tag{2}$$

where

$$\mu = (1+\exp(\mathbf{X}\beta))^{-1} \quad \text{and} \quad \phi = \exp(\mathbf{Z}\gamma) \tag{3}$$

where $\mu \in (0,1)^n$ is the $n \times 1$ vector of location parameters and $\phi \in (0,\infty)^n$ the $n \times 1$ vector of precision parameters, which have been linearly expanded to account for the presence of covariates. To estimate model parameters $\theta = (\beta,\gamma)$, we use an adaptive Metropolis-Hastings algorithm where the transition distribution is approximated by means of a multivariate Normal distribution $q(\theta^{(t)}|\theta^{(t-1)}) = \mathcal{N}(;\theta^{(t-1)},\Sigma^{(t)})$, with the covariance matrix $\Sigma^{(t)}$ being adapted at each step by using a convenient subsample from the previous samples [5]. In this context, the acceptance ratio of the sampler is as follows:

$$\alpha^{(t)} = \frac{\mathcal{L}(\theta^{(t)};\mathbf{m},\mathbf{s})q(\theta^{(t-1)}|\theta^{(t)})f(\theta^{(t)})}{\mathcal{L}(\theta^{(t-1)};\mathbf{m},\mathbf{s})q(\theta^{(t)}|\theta^{(t-1)})f(\theta^{(t-1)})} \tag{4}$$

where $\mathcal{L}(\theta^{(t)};\mathbf{m},\mathbf{s})$ is the likelihood function for the fuzzy sample of data and $f(\theta^{(t)})$ is the prior density ascribed to the model parameters. In the case of i.i.d. and non-interactive fuzzy responses, the imprecise likelihood function is as follows [6, 7]:

$$\mathcal{L}(\theta^{(t)};\mathbf{m},\mathbf{s}) = \prod_{i=1}^{n} \int_0^1 \xi_{\tilde{y}_i}(y;m_i,s_i) \frac{\Gamma(\phi_i)y^{(\mu_i\phi_i-1)}(1-y)^{(\phi_i-\mu_i\phi_i-1)}}{\Gamma(\phi_i\mu_i)\Gamma(\phi_i-\mu_i\phi_i)}\, dy \tag{5}$$

4 Application

In the present application we aimed to investigate the predictors of sexual intimacy in a sample of $n = 450$ participants from Flanders (73% female, mean age 32.9 years, mean relationship length 7.68 years).[1] Because of its characteristics, assessing the determinants of sexual intimacy is a typical situation in which raters show some levels of decision uncertainty in providing their self-reported responses. The survey consisted of four questionnaires used to measure (i) the perceived sexual intimacy with the partner, (ii) the perceived partner responsiveness (i.e., the extent to which one experiences the partner as being responsive to emotional needs), (iii) the sexual desire, (iv) the avoidant attachment score (i.e., how ambivalent early developmental experiences affect the current relationship). The items have been measured on a 7-point Likert-type scale with response categories ranging from 1 ("definitely not") to 7 ("yes, definitely"). The items associated with sexual intimacy have been fuzzified using the fuzzy-IRTree method [1] and the ensuing fuzzy beta responses

[1] The dataset is publicly available at `http://osf.io/adgw2/`. For further details about the survey, see [8].

have been averaged to form the final intimacy indicator (see Figure 1 for a subsample of fuzzy responses). Following the findings by [8], three additive Beta linear models M1-M3 have been defined to predict sexual intimacy (see Table 1). The models have been varied in terms of covariates for the term μ, whereas no covariates have been used to model the precision term (i.e., $\phi = \exp(\mathbf{1}\gamma)$). For all the models, diffuse Normal densities have been used for the priors $f(\beta) = \mathcal{N}(\beta; \mathbf{0}, \mathbf{I}10)$ and $f(\gamma) = \mathcal{N}(\gamma; 0, 3)$ and four parallel MCMC have been run with 20000 samples (5000 samples for the burn-in phase) by means of the R package `MHadaptive`. The final model has been chosen according to the LOO information criterion as implemented by the R package `loo` [9]. According to the Gelman and Rubin's convergence diagnostics, all the chains reached the convergence (i.e., $\hat{R} = 1.00$). Table 2 reports the posterior quantiles along with the 95% HDIs for the six model parameters. Figure 2 shows the marginal posterior densities for the model parameters whereas Figure 3 plots the predicted curves against the observed fuzzy data as a function of both continuous and categorical predictors. The posterior results suggest that sexual intimacy is predicted by the perceived partner responsiveness, with a slight decrease of the outcome for the case in which the partner is male. The other predictors seem to play a marginal role in predicting sexual intimacy, with a negative relationship between avoidant attachment style and sexual intimacy as expected.

Model	Covariates	LOOIC
M1	`partner_respo`, `sex_desire`	873.80
M2	`partner_respo`, `sex_desire`, `attach_avoid`	863.50
M3	`partner_respo`, `sex_desire`, `attach_avoid`, `gender_partner`	**857.00**

Table 1 Application: Models for the sexual intimacy fuzzy rating data. Note that M3 is the best model according to the lowest LOO-IC criterion.

	β_0	$\beta_{\texttt{partner\_respo}}$	$\beta_{\texttt{sex\_desire}}$	$\beta_{\texttt{attach\_avoid}}$	$\beta_{\texttt{gender\_partner:Male}}$	γ
min	-1.01	0.11	-0.01	-0.05	-0.26	3.90
mean	-0.59	0.13	0.02	-0.03	-0.11	4.35
max	-0.12	0.15	0.04	-0.00	0.03	4.90
0.95 HDI_{lb}	-0.83	0.12	0.01	-0.04	-0.20	4.10
0.95 HDI_{ub}	-0.36	0.14	0.03	-0.01	-0.04	4.68

Table 2 Application: Posterior quantiles and 95% HDI for the model parameters. Note that β_0 is the intercept of μ and codifies the level `gender_partner = Female`.

References

1. Calcagnì, A., Cao, N., Rubaltelli, E., Lombardi, L.: A psychometric modeling approach to fuzzy rating data. Fuzzy Sets and Systems (2022)

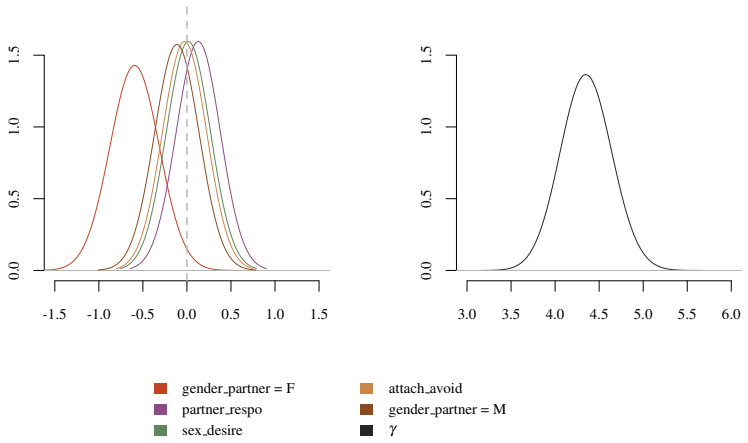


Fig. 2 Application: Marginal posterior densities for the model parameters.

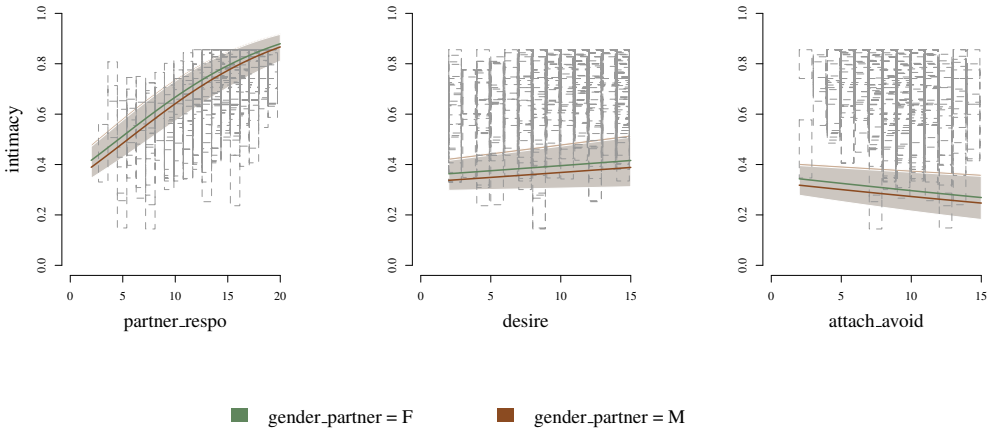


Fig. 3 Application: Observed fuzzy data for sexual intimacy as a function of the categorical predictor `gender_partner` (colors in the panels) and the three continuous predictors (panels). Fitted curves correspond to posterior means (see Table 2) whereas shadows represent the posterior 95% HDI of the predicted curves. Note that rectangles represent α-cuts of the observed fuzzy data with $\alpha = 0.5$, i.e. $\mathbf{y}_i^{\alpha} = \left[\min\left(\{y \in [0,1] : \xi_{\widetilde{y}_i}(y) > 0.5\}\right), \max\left(\{y \in [0,1] : \xi_{\widetilde{y}_i}(y) > 0.5\}\right)\right]$

2. Leary, M.R., Kowalski, R.M.: Impression management: A literature review and two-component model. Psychological bulletin **107**(1), 34 (1990)
3. Yu, S.-C., Wu, B.: Fuzzy item response model: a new approach to generate membership function to score psychological measurement. Quality and Quantity **43**(3), 381–390 (2009)
4. Coppi, R., Gil, M.A., Kiers, H.A.: The fuzzy approach to statistical analysis. Computational

statistics & data analysis **51**(1), 1–14 (2006)
5. Haario, H., Saksman, E., Tamminen, J.: An adaptive metropolis algorithm. Bernoulli, 223–242 (2001)
6. Gil, M., Corral, N., Gil, P.: The minimum inaccuracy estimates in $\chi 2$ tests for goodness of fit with fuzzy observations. Journal of Statistical Planning and Inference **19**(1), 95–115 (1988)
7. Calcagnì, A., Lombardi, L.: Modeling random and non-random decision uncertainty in ratings data: a fuzzy beta model. AStA Advances in Statistical Analysis, 1–29 (2021)
8. Van Lankveld, J.J., Dewitte, M., Verboon, P., van Hooren, S.A.: Associations of intimacy, partner responsiveness, and attachment-related emotional needs with sexual desire. Frontiers in Psychology **12** (2021)
9. Vehtari, A., Gelman, A., Gabry, J.: Practical bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and computing **27**(5), 1413–1432 (2017)

A Mixture Model for Multi-Source Cyber-Vulnerability Assessment

Un modello di mistura per la valutazione delle cyber-vulnerabilitá

Mario Angelelli, Serena Arima and Christian Catalano

Abstract Cyber-risk assessment is gaining increasing attention due to the potential impact of cyber-incidents on digital and physical systems, with consequent effects on individuals and organisations. This work introduces a Bayesian mixture model to relate individual cybervulnerability features and accessible information that could affect attackers' evaluation and cyber-incident occurrence. Individual cybervulnerability features are clustered according to a Dirichlet process. Future research directions aimed at prioritising cyber-vulnerabilities and access to information are discussed too.

Abstract *La valutazione del cyber-rischio è oggi un ambito di interesse crescente a causa del potenziale impatto che gli incidenti informatici possono avere su sistemi digitali e fisici, con conseguenti ricadute sulle persone e sulla collettività. In questo lavoro, si introduce un modello di regressione Bayesiana per collegare le caratteristiche individuali delle cyber-vulnerabilità con le informazioni accessibili che potrebbero influenzare la percezione di tali vulnerabilità e e il verificarsi di incidenti informatici. Il modello prevede un effetto casuale rappresentato da una mistura e volto a evidenziare raggruppamenti di cyber-vulnerabilità senza assunzioni stringenti a priori. Il lavoro si conclude con la discussione di future direzioni di ricerca volte alla prioritizzazione delle cyber-vulnerabilità e all'accesso alle informazioni.*

Key words: Mixture model, Bayesian regression, Cyber-risk

Mario Angelelli
University of Salento, e-mail: mario.angelelli@unisalento.it

Serena Arima
University of Salento, e-mail: serena.arima@unisalento.it

Christian Catalano
University of Salento, e-mail: christian.catalano@unisalento.it

1 Introduction

Cyber-incidents are of paramount relevance in current technological systems. Issues posed by compromised data Confidentiality, Integrity, and Availability ("CIA" paradigm) may affect the validity of experimental results in the scientific context, as well as the generated value in productive sectors. The plurality of incidents and attack categories (e.g. Denial-of-Service, Malware, but also social engineering) may lead to informative, economic, functional, or reputational impact. These issues are now further amplified by the high level of connectivity among devices and networks entailed by the Digital Transformation.

For these reasons, cyber-risk assessment is gaining increasing attention for an informed use of digital resources. Widely accepted standards for cyber-risk assessment comes from the qualitative evaluation of a set of features associated with each cyber-vulnerability. The National Institute of Standards and Technology (NIST) provides a severity ranking of emerging cyber-vulnerabilities based on their impact (along the "CIA" dimensions) and accessibility assessment [2]. However, the actual risk entailed by a cyber-vulnerability also depends on additional factors that may influence attackers' perception regarding a cyber-vulnerability's exploitability: exposure of vulnerable technologies, their diffusion in critical systems, exploit[1] availability (existence and value).

The present work introduces a statistical model to address the relation between cyber-vulnerability assessment and information on the extent of its exploitability. The multiplicity of information sources, impact dimensions, means of exploitation, and attackers' perception suggests adopting a model without *a priori* assumptions on the cyber-vulnerability classification. Formally, we consider a regression model involving: i. categorical regressors, i.e. cyber-vulnerability features, and ii. a set of response variables that express the priority of the cyber-vulnerability, which can be understood as severity, exposure, or availability of resources for exploitation. Multiple response variables reflect the multiplicity of information sources quantifying these interpretations. Then, we choose a Bayesian mixture model to take into account potential grouping effects.

2 Data description and methodology

2.0.1 Data description

The set of regressors that qualitatively assess the cyber-vulnerability features is based on the *attack vector* in the NIST classification. This is a 6-dimensional vector $\mathbf{X}$ of categorical variables.

[1] An exploit is defined as software that can be directly executed to realise a cyber-attack.

- Variables X_C, X_I, X_A reflect the assessed impact on data in terms of Confidentiality, Integrity, and Availability, respectively; each of these categories contains three levels ("none: 0", partial: 0.275", "complete: 0.660").
- Variables X_{AV}, X_{AC}, X_{Au} describe the accessibility prompted by the cyber-vulnerability: i. x_{AV} describes the Access Vector ("Requires local access: 0.395, "Local Network accessible: 0.646", "Network accessible: 1"); ii. x_{AC} assesses Access Complexity (ranked "high: 0.35, "medium: 0.61", "low: 0.71"); iii. x_{Au} assesses Authentication ("Requires no authentication: 0.704", "Requires single instance of authentication: 0.56", "Requires multiple instances of authentication: 0.45).

CVSS is widely accepted as a relevant measure for cyber-vulnerabilities along different dimensions[2]. *Known* vulnerabilities are ranked in terms of the CVSS, the score representing the severity of such vulnerabilities [2]. Conditioned to the attack vector, the CVSS severity score is deterministic: based on the quantitative assessment of regressors presented in Subsection 2.0.1, the impact I, the exploitability E, and the severity score CVSS, are

$$\begin{aligned} \mathrm{I} &= 10.41 \cdot (1-(1-X_C)\cdot(1-X_I)\cdot(1-X_A)), \\ \mathrm{E} &= 20 \cdot X_{AV} \cdot X_{AC} \cdot X_{Au}, \\ \mathrm{CSVV} &= (0.6\cdot \mathrm{I} + 0.4 \cdot \mathrm{E} - 1.5) \cdot 1.176 \cdot (1-\delta_{\mathrm{I}}) \end{aligned} \tag{1}$$

where δ denotes Kronecker's delta.

Different response variables y can be considered to address multiple information sources: we will focus on exposure of vulnerable technologies, software availability for exploitation, and their value. These views on cyber-vulnerabilities are associated with actual data from multiple sources (databases).

With these premises, the composition of the dataset is described as follows:

- statistical units are associated with a unique ID and are extracted from the Common Vulnerabilities and Exposures (CVE) database, which is the knowledge base regarding exploited vulnerabilities.
- Each CVE is associated with an evaluation of the regressors presented above, which is provided by the NIST.
- the Shodan database[3] reports exposed hosts for known vulnerabilities: this may represent a relevant driver for attackers' intervention. Shodan allows including CVE and Country as keys for queries.
- Reported exploits for CVEs can be extracted from ExploitDB[4].
- The information on exploits can be further refined from VulnDB[5], a database that collects information on the price range for exploits associated with a CVE. Fields extracted from VulnDB include the price range as 0-day (i.e. when the vulnerability was not disclosed and there were no available solutions to patch it),

[2] https://nvd.nist.gov/vuln-metrics/cvss

[3] https://exposure.shodan.io

[4] https://www.exploit-db.com/

[5] https://vuldb.com/

the price rance at the moment of querying, and the exploitability (specified by levels “No defined”, ”Unproven”, ”Proof-of-Concept”, and “Highly functional”).

For all these databases, we prepared scripts using Python in order to automatically extract the required data through APIs. We started from the selection of exposure data in Italy through Shodan to obtain a base set of CVEs. Subsequently, the scripts was adapted to extract Attack Vectors associated with these CVEs from NIST database, and to check the exploits availability from ExploitDB and VulnDB.

The final dataset consists of $n = 593$ records.

2.0.2 Model definition

The regression model is

$$y = \mathbf{X}^{\mathbf{T}} \cdot \beta + \rho + e \tag{2}$$

where

- β_p are the regression parameters, $p \in \mathscr{I} \subseteq \{1,\dots,6\}$. Being associated with categorical data, we adopt an ANOVA representation where each β_p correspond to $\beta_{p,\ell} \sim \mathscr{N}(0,100)$, $\ell \in \{0,\dots,L_p-1\}$, and L_p is the number of levels for the p-th variable.
- ρ is a random effect associated with groups of cyber-vulnerabilities, i.e.

$$\rho \sim \sum_{i=1}^{\infty} w_k \mathscr{N}(\mu_k, 100). \tag{3}$$

- $e \sim \mathscr{N}(0,1)$ is the residual.

In order to implement the Bayesian model, we consider ρ as arising from a Dirichlet process, which is realised using the stick-breaking representation [3]:

$$\begin{aligned}
\alpha &\sim \Gamma(1,1),\\
v_k &\sim \mathrm{B}(1,\alpha), \quad k \in \mathbb{N},\\
w_k &= v_k \cdot \prod_{i=1}^{k-1}(1-v_i), \quad k \in \mathbb{N},\\
\mu_k &\sim \mathscr{N}(0,100), \quad k \in \mathbb{N},\\
\rho|k &\sim \mathscr{N}(\mu_k,100) \quad \text{with } p(k|w) = w_k.
\end{aligned} \tag{4}$$

From this model, y_i, ρ_i, and e_i, $i \in \{1,\dots,n\}$ represent i.i.d. samples.

The evaluation of posterior distributions is based on this model and data X_i describing the attack vector for n cyber-vulnerabilities, and it is carried out through Markov Chain Monte Carlo (MCMC) sampling.

3 Preliminary results

MCMC algorithm allows us to obtain samples from the posterior distribution of all model parameters. The algorithm identified four main components starting from two regressors, namely the categorical variable X_{AV} and the $CVSS$ severity score, and a response variable expressing monitored exposure. The posterior distribution of regressors indicates that they are significant, as shown in Figure 1. The analysis

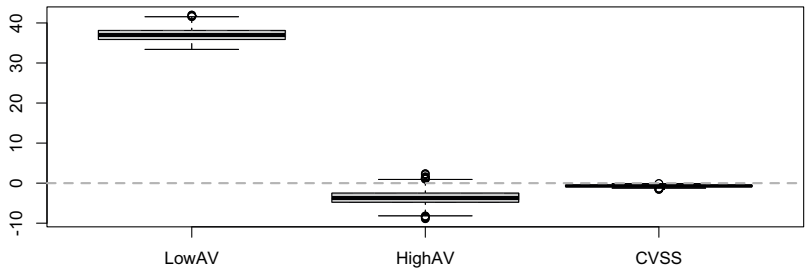


Fig. 1 Box plots for regression parameters associated with the categorical variable X_{AV} (corner: β_{LowAV}; differential: β_{HighAV}) and the severity score $CVSS$.

of the cluster membership for the n units shows that the attack vector does not fully determine the grouping of sampled vulnerabilities. In particular, CVEs associated with the same attack vector do not always lie in the same group k. This suggests that the extent to which vulnerable hosts can be identified, which is quantified by the Shodan response, does not only depend on technical features of the cybervulnerability (expressed by the attack vector), but also on contextual factors that depend on the diffusion and the use of a given technology.

In order to better explore these factors, further analyses involving different combinations of regressors and response variables, with special regard to available resources for exploitation are necessary.

4 Conclusion and Future Work

The present work is a first study aimed at an integrated study of cyber-vulnerabilities through statistical modelling. Other statistical models are being developed for cyber-risk assessment, e.g. [1], and different models can be used to highlight specific aspects that are relevant for cyber-risk assessment.

The actual realisation of cyber-attacks does not only rely on technical features represented by attack vectors, but also on different information sources that can promote or mitigate cyber-attacks. It is plausible that access to information has a relevant role in this regard: as mentioned in [1], data regarding cyber-security are subject to limited disclosure and underestimation, both to adhere to security standards and to prevent reputational loss. On the other hand, open data provided by Organisations may be used not only to prevent cyber-incidents, but also to guide cyber-attackers.

Future work will further explore these issues, addressing the role of ranking of cyber-vulnerabilities dependent on multiple information sources and criteria, as well as relations between these rankings, in order to support decision-makers in both prioritisation and information disclosure.

References

1. Giudici, P., Raffinetti, E.: Cyber risk ordering with rank-based statistical models. AStA-Adv. Stat. Anal. **105**(3), 469–484 (2021)
2. Mell, P., Scarfone, K., Romanosky, S.: A complete guide to the common vulnerability scoring system version 2.0. In Published by FIRST-forum of incident response and security teams, **1**, p. 23 (2007)
3. Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sin., 639-650 (1994)

Hierarchical Bayesian models for analysing fish biomass data.

An application to Parapenaeus longirostris biomass data.

Rita Fici, Antonino Abbruzzo, Luigi Augugliaro and Giacomo Milisenda

Abstract The Mediterranean International Trawl Survey (MEDITS) programme provides spatially referenced ecological data. We adopted a hierarchical Bayesian model to analyse *Parapenaeus longirostris* biomass data. The model comprises three parts, each of which identifies: the variability due to the explanatory variables, the variability due to the spatial domain (seen as a Gaussian Process) and the irregular component modelled as white noise. The estimated parameters show that some seabed characteristics affect biomass quantity and that the estimated behaviour of the Gaussian Process changes over different groups of years.

Abstract *Il programma di ricerca internazionale Mediteranean International Trawl Survey (MEDITS) raccoglie dati ecologici identificati geograficamente. In questo lavoro, per analizzare dati sulla biomassa del Parapenaeus longirostris, è adottato un approccio Bayesiano gerarchico dove il modello di regressione spaziale è formato da tre componenti che dovrebbero racchiudere: la variabilità relativa alle variabili esplicative, quella dovuta alla componente spaziale (modellata attraverso un Processo Gaussiano), e una componente erratica modellata come un white noise. I parametri stimati indicano che alcune caratteristiche del fondale marino influenzano la quantità di biomassa, e che il Processo Gaussiano stimato cambia tra i diversi gruppi di anni analizzati.*

Key words: Gaussian Processes, Bayesian methods, spatial analysis, latent variables.

Antonino Abbruzzo, Luigi Augugliaro and Rita Fici,
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy. e-mail: antonino.abbruzzo@unipa.it; luigi.augugliaro@unipa.it; rita.fici@unipa.it
Giacomo Milisenda
Department of Integrative Marine Ecology (EMI), Stazione Zoologica Anton Dohrn, Sicily Marine Center, Palermo, Italy. e-mail: giacomo.milisenda@szn.it

1 Introduction

Several social, political, and economic players show a substantial interest in studying the marine ecosystem. The MEDITS project started in 1994 with the cooperation between several research institutes from the Mediterranean States. The target is to conduct a standard bottom trawl survey in which all the participants use the same sampling protocol and the same methodology [1]. The data concern the biomass quantity of *Parapenaeus longirostris* in the area extending from the southern coasts of Sicily. This fish species is one of the most critical in biological and economic terms. This paper aims to investigate the dependency structures between the biomass and some explanatory variables and study the covariance and autocovariance structures as a function of geographical coordinates. The hierarchical Bayesian model (Equation (1) in Section 2) captures these two aims by equating the response variable with the sum of three components: the effect of explanatory variables, the impact of latent variables, as a function of space, and a white noise.

2 Model specification

Let $\mathbf{Y}(\mathbf{s}) = \{Y(\mathbf{s}_1), Y(\mathbf{s}_2), \cdots, Y(\mathbf{s}_n)\}$ be an n-dimensional random vector of observations of biomass quantity for one species in the spatial points $\mathbf{s}_1, \ldots, \mathbf{s}_n$. We specify a regression equation as follows:

$$Y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta} + W(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \qquad \mathbf{s}_i \in \mathbb{R}^2; \tag{1}$$

where $\mathbf{x}(\mathbf{s}_i)$ are the geographically determined covariates, $\boldsymbol{\beta}$ is a vector of regression parameters, $W(\mathbf{s}_i)$ are the latent variable's effects incorporating the variability due to the space domain, and $\varepsilon(\mathbf{s}_i)$ are the independently and identically distributed *white noise*'s manifestations, with variability τ^2, representing the short scale randomness in the regionalized variable, also called *nugget effect* [2].

Assuming that the latent variable's effects follow a Gaussian Process (GP), $\mathbf{W}(\mathbf{s}) \sim GP(\mathbf{0}, \mathbf{C}_\theta)$, and $\boldsymbol{\varepsilon} \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \tau^2 \boldsymbol{I})$, the posterior parameter distribution is proportional to:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2) \times GP(\mathbf{w}|\mathbf{o}, \mathbf{C}_{\boldsymbol{\theta}}) \times N(\mathbf{y}|\mathbf{X}\beta + \mathbf{W}, \tau^2 \mathbf{I}), \tag{2}$$

where the first element is the joint prior distribution of the parameters, which depends on prior's choice. We consider as prior $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ where p is the number of explanatory variables. The choice of prior distributions for τ^2 and $\boldsymbol{\theta}$ are explained in the next paragraphs. The second element of the posterior in (2) is specified as a GP, with $\mathbf{C}_{\boldsymbol{\theta}}$ as the correlation matrix defined by a parametric covariance function. The last term in (2) represents the probability density distribution attributed to the analysed phenomenon, where the measurement errors or microscale variance depend on the parameter τ^2, while the mean is given by the sum of the process effects and the regression on explanatory variables [3].

Covariance matrix. Using Wackernagel's coregionalization approach [5], a possible specification of the entries in the covariance matrix is $\mathbf{C}_{\boldsymbol{\theta}}(\boldsymbol{s}_i,\boldsymbol{s}_j) = \sigma^2\rho(\boldsymbol{s}_i,\boldsymbol{s}_j;\phi,\nu)$, where $\sigma^2 = var[W(\boldsymbol{s})]$ is the space variance and $\rho(\cdot)$ represents the correlation function between $W(\boldsymbol{s}_i)$ and $W(\boldsymbol{s}_j)$. The specification of the parametric correlation function (which must return a symmetric and positive defined correlation matrix) constitutes an exciting point of reflection regarding the theoretical definition for complex relationships and their practical implementation. We used the correlation function formulated by Matérn [7]:

$$\rho(\boldsymbol{s}_i,\boldsymbol{s}_j;\phi,\nu) = \frac{(||\boldsymbol{s}_i-\boldsymbol{s}_j||\phi)^{\nu}K_{\nu}(||\boldsymbol{s}_i-\boldsymbol{s}_j||;\phi)}{2^{\nu-1}\Gamma(\nu)}, \tag{3}$$

where $||\boldsymbol{s}_i-\boldsymbol{s}_j||$ is the Euclidean distance between observed space coordinates $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$; ϕ tunes the spatial correlation decadence (the bigger ϕ, the faster the spatial correlation comes close to zero when the Euclidean distance increases); $\Gamma(\cdot)$ is the Gamma function and $K_{\nu}(\cdot)$ is the Bessel function of the third kind and order ν. If $\nu = 0.5$, the equation (3) is reduced to the exponential correlation function $\rho(\boldsymbol{s}_i,\boldsymbol{s}_j;\phi) = \exp[-\phi(||\boldsymbol{s}_i-\boldsymbol{s}_j||)]$.

Prior parameters choice. The choice of the Matérn correlation function in equation (3) implies the set of parameters $\theta = \{\sigma^2,\phi,\nu\}$. As recommended by Banerjee *et al.* [6], ϕ is given an informative prior distribution while a relatively vague prior is given to σ^2. The chosen prior for ϕ is a Uniform $(0.003, 0.008)$ and its starting value is 0.005. A decadence parameter of 0.003 or 0.009 means that after two km or after half a km the spatial correlation is close to zero. The parameter ν is chosen to follow a Uniform (0.1 , 0.9) and its starting value is 0.5. The priors for τ^2 and σ^2 follow two Inverse Gamma (IG) distributions. A density-based clustering algorithm [4] is used to choose their hyperparameters. The clusters are composed of at least ten observations less than 2.5 km away from their core point. The prior for τ^2 is an IG(0.07, 24.08) which gives an expected value equal to the median of the biomass's variances within clusters and a τ^2's variance equal to the variance of the clusters variances vector. The starting value of τ^2 is 337.66. The prior for σ^2 is an IG(3.062, 2143.75) which gives an expected value equal to the difference between the total response variance and the expected value of τ^2, and gives a big enough variance of σ^2. Its starting value is 700.
Inference is carried out by MCMC algorithms. The set of parameters $\{\sigma^2,\phi,\nu,\tau^2\}$ is estimated by a Metropolis-Hastings algorithm, where the proposal distributions are Normal with $50, 0.5, 0.2$ and 100 as variance parameters, respectively. A Gibbs-Sampler algorithm is used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{w}$. See [8] for details.

3 Data analysis

We analysed data related to the years in which the sampling was done during the summer (from June to August). Two different subsets of data are analysed, one given by the years from 2002 to 2005 and 2007, and the other given by years from 2011 to 2016 (2014 excluded). In each model comprise 1,005,000 iterations, a burn-in of 25% and a thinning value of 700. The final sample size has 1077 values for each parameter.

We use the Deviance Information Criterion (DIC) as a measure of model choice. The lower the DIC value, the better the model.

Here we show the results related to *Parapenaeus longirostris*'s biomass.

$\rho(\boldsymbol{s}_1;\boldsymbol{s}_j\|\boldsymbol{\theta})$	σ	τ	ϕ	ν	DIC
Esponenziale (2002-07)	(17.231 , 25.179)	(2.978 , 15.291)	(0.003 , 0.008)	0.5	654.0771
Esponenziale (2011-16)	(19.670 , 40.538)	(3.803 , 34.164)	(0.003 , 0.008)	0.5	1175.152
Matérn (2002-07)	(17.015 , 25.157)	(3.007 , 15.737)	(0.004 , 0.008)	(0.170 , 0.865)	656.6011
Matérn (2011-16)	(19.387 , 40.981)	(4.115 , 34.679)	(0.003 , 0.008)	(0.140 , 0.886)	1178.819

Table 1: 95% credible interval of parameters of variability for *Parapenaeus longirostris*'s biomass.

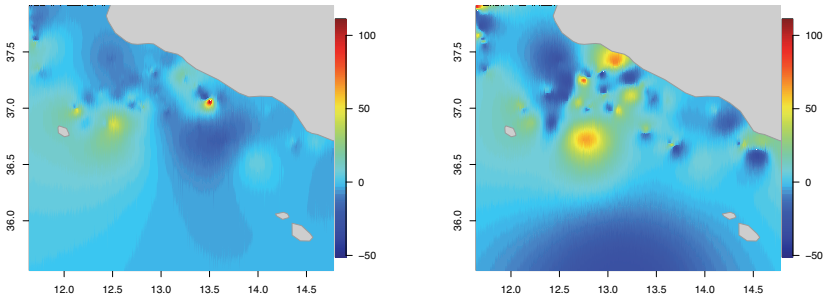
As shown in Table 1, for both groups of years, the value of the DIC is lower when an exponential correlation function is considered. It is possible to note that the estimated values for the parameter σ are bigger than the estimated values for τ: this suggests a bigger impact of the variability due to latent variables related to space than the variability due to the *nugget effect*. In all models, the estimated parameter

$\rho(\boldsymbol{s}_1;\boldsymbol{s}_j\|\boldsymbol{\theta})$	Depth	sq. Depth	Temperature	Seabed
Esponenziale (2002-07)	(-6.66 , 3.34)	(-9.24 , -0.61)	(-6.00 , 1.80)	(1.91 , 10.91)
Esponenziale (2011-16)	(-20.42 , -6.21)	(-10.15 , 1.68)	(-5.39 , 6.37)	(1.68 , 14.94)
Matérn (2002-07)	(-7.21 , 3.35)	(-9.26 , -0.43)	(-6.00 , 1.97)	(1.77 , 11.59)
Matérn (2011-16)	(-20.99 , -6.02)	(-9.97 , 1.77)	(-5.40 , 7.00)	(1.37 , 15.00)

Table 2: 95% credible interval of regression coefficients for *Parapenaeus longirostris*'s biomass.

for the slope of the seabed is significant. The depth is significant for the first group of years, while its quadratic component is significant for the second group of years. Temperature is never significant (Table 2).

Figure 1 shows the estimated components of the GP for the two groups of years. Blue and red colours indicate a negative and a positive effect of the GP on the biomass, respectively. W's distribution in the first group of years looks more homogeneous than W's distribution in the second group of years. Indeed, in Figure



(a) First group of years.

(b) Second group of years.

Fig. 1: *Parapenaeus longirostris*'s estimated component of GP.

(1a), excluding the only red-coloured zone, the change from one colour to another is slow. In Figure (1b) the changes from blue to yellow-coloured zones are more sudden than what happens in the first group of years (Figure 1a). Moreover, it is possible to observe darker blue-coloured zones. The difference between the two W's distributions could be due to bigger estimated values of the 95% credible interval for the spatial correlation decadence parameter, ϕ, in the second group of years (0.00314 , 0.00788) than in the first group (0.00305 , 0.00793).

4 Conclusions

Hierarchical Bayesian models have been applied to analyse ecological spatial data. The spatial effect is assumed to follow a Gaussian Process in which a parametric function gives the correlation between points. An informative prior distribution is given to the parameter related to the decadence of spatial correlation to assign a low probability to a correlation close to zero when the distance between the points is too small (less than half a kilometre). Previous sections present the results of applied models to *Parapenaeus longirostris*'s biomass data. Two analyses have been carried out, using a Matérn correlation function and an exponential correlation function. The comparison between them shows a better algorithm performance using the exponential function. The explanatory variables are the linear and the quadratic components of the depth, the temperature and the seabed slope. The seabed is always significant. Depth is significant with its linear component for some models and its quadratic component for others. Comparing the estimated Gaussian processes between the two groups of years is possible to note a stronger space correlation in the earliest years than in the latest years of observation.

We conclude with some final remarks regarding directions for future research. Analysing multivariate response data would increase the knowledge of the interactions between species and understanding the food web structure.

References

[1] International bottom trawl survey in the Mediterranean. **9**(2017)
[2] Finley, A. O., Banerjee, S. and Carlin, B. P.: spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models. Journal of statistical software, **19**(4) (2007): 1-24.
[3] Banerjee S. Finley A.O. Gelfand A.E. Datta, A.: On nearest-neighbor gaussian process models for massive spatial data. Wiley Interdisciplinary Reviews: Computational Statistics, **8**(5) (2016): 162-171.
[4] Hahsler, M., Piekenbrock and M., Doran, D. : dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software, **91**(1) (2019): 1-30.
[5] Wackernagel, H.: Multivariate Geostatistics. (1998)
[6] Banerjee, S., Carlin, B.P. and Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data. (2015)
[7] Matérn B.: Spatial Variation. (1960)
[8] Finley A. O., Banerjee S. and Gelfand A. E. : spBayes for Large Univariate and Multivariate Point-Referenced Spacio-Temporal Data Models. arXiv preprint arXiv:1310.8192 (2015)

Insights into the derivative-based method for nonlinear mediation models

Considerazioni sul metodo basato sulle derivate per modelli di mediazione non lineari

Claudio Rubino and Chiara Di Maria

Abstract Associational mediation analysis has generally relied on the linearity of models to estimate the indirect effect as a product of regression coefficients. Very few examples of generalisations to nonlinear settings have been proposed, including a derivative-based method that, however, is far from being widely spread among scholars. In this paper, we clarify some aspects of such an approach to nonlinear mediation models, which have not been addressed by the previous literature. In addition, we run a simulation study to compare confidence intervals for the indirect effect obtained through different approaches.

Abstract *L'analisi di mediazione in contesti associativi si è generalmente basata sulla linearità dei modelli per stimare l'effetto indiretto come prodotto di coefficienti di regressione. Tra i pochi esempi di generalizzazione a contesti non lineari proposti si annovera un metodo basato sulle derivate, che tuttavia è ben lontano dall'essere diffuso tra i ricercatori. In questo articolo chiariamo alcuni aspetti di questo approccio che non sono stati ancora affrontati dalla letteratura precedente. Inoltre, si effettua uno studio di simulazione per confrontare gli intervalli di confidenza per l'effetto indiretto ottenuti con diversi approcci.*

Key words: mediation analysis, derivative method, generalised linear models, bootstrap, Monte Carlo method, Bayesian statistics

1 Introduction

Mediation analysis is nowadays a widely spread approach to address settings where the effect of a variable on a response of interest is not entirely direct, but is transmitted by one or more intermediate variables called *mediators*. Mediation analysis was

Claudio Rubino, Chiara Di Maria
Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, Building 13, Palermo 90128, Italy. e-mail: claudio.rubino@unipa.it, chiara.dimaria@unipa.it

developed by social scientists [2] who used the path-analytic framework introduced by [12], subsequently further extended to structural equation models (SEM, [3]).

The main features of such an approach are its associational nature, in the sense that the direct and indirect effects do not have a causal interpretation, and its reliance on linear models. Indeed, when the mediator and the outcome models are linear (identity link functions and no interactions) the indirect effect can be estimated as the product of the coefficients lying on the paths connecting X to M and M to Y [2, 3, 8].

However, in real-world applications, it is frequent to deal with non-continuous variables requiring link functions different from the identity; consider, for example, a binary outcome denoting the presence or absence of a certain disease. Despite the widely spread presence of such variables in applied research, mediation analysis with nonlinear models has primarily been addressed in a causal framework [1, 5, 4]. As regards the associational one, very few approaches to deal with this issue have been proposed, based either on approximations [9], or on the less employed difference method [10].

An exception is given by a generalisation of the product method proposed in the '80s [11], and recently revived by [7, 6]. Although quite intuitive, this approach is not widely known and applied by practitioners, and for this reason it is not well developed, presenting theoretical shortcomings yet to be addressed. The aim of the present paper is to discuss some of these gaps and provide more insights into an approach having much potential. We first describe the method, then we move to some of its aspects to be deepened and finally we show the results of a simulation study concerning some of these issues.

2 Mediation analysis in nonlinear models

The basic idea underlying the method proposed by [11] is that the indirect effect can be interpreted as the variation in the outcome Y corresponding to a change in the exposure X through the variation in the mediator M. Such a definition can be formalised in terms of derivatives. Let $\hat{M}$ and $\hat{Y}$ denote the conditional expectations $\mathbb{E}[M \mid X]$ and $\mathbb{E}[Y \mid X, M]$, respectively, and consider the following models

$$g_1(\hat{M}) = \beta_0 + \beta_1 X \quad \rightarrow \quad \hat{M} = h_1(\beta_0 + \beta_1 X) \tag{1}$$

$$g_2(\hat{Y}) = \gamma_0 + \gamma_1 X + \gamma_2 M \quad \rightarrow \quad \hat{Y} = h_2(\gamma_0 + \gamma_1 X + \gamma_2 M), \tag{2}$$

where g_1 and g_2 are possibly non-linear link functions, connecting the conditional expectations of the mediator and the outcome to their linear predictors, and $h_k = g_k^{-1}$, $k \in \{1,2\}$. The indirect effect is then defined as

$$\frac{\partial \hat{Y}}{\partial M} \frac{\partial \hat{M}}{\partial X}, \tag{3}$$

i.e. the product of the derivative of Y on M and that of M on X. Notice that in the simple case of identity link functions, the indirect effect reduces to the traditional expression obtained via the product method $\beta_1\gamma_2$. In contrast, when at least one of the g functions differs from identity, the indirect effect does not assume a single value, since its expression depends on X and/or M. For this reason, [6] suggest to call the effect in Equation (3) *conditional indirect effect* (CIE). Thus, the researcher chooses some values of X of potential interest, say $x_1,\ldots,x_p$ and, if the expression of the indirect effect involves also the mediator, its values should be selected accordingly to those of X, as the predicted values corresponding to $x_1,\ldots,x_p$, obtained from the fitted model.

Although the main focus of this paper is on mediation analysis with generalised linear models (GLMs), it is worth remarking that the derivative-based approach can be used also in situations where the mediator or the outcome depends on nonlinear transformations of their regressors, such as X^2 or $\log(X)$, see [7] for some examples.

3 Potential issues

In this section, we discuss some aspects of the derivative-based method which have not been addressed or satisfactorily deepened in the previous literature on the topic.

Binary mediator

As already mentioned, when the expression of the indirect effect involves both X and M, the values of M should be fixed at the values it assumes in correspondence of the selected values of X, determined by the fitted model. This is quite straightforward, unless the mediator is binary, when some issues arise.

Consider a setting with a binary mediator and an outcome described by Equations (1)-(2), where g_1 is the logit link and g_2 is a generic function different from identity, say the logarithm to fix ideas. Therefore, applying the formula in Equation (3), the CIE is given by

$$\beta_1\gamma_2\frac{\exp(\beta_0+\beta_1X)}{(1+\exp(\beta_0+\beta_1X))^2}\exp(\gamma_0+\gamma_1X+\gamma_2M),$$

which, as can be seen, depends on both X and M. However, when coming to the estimation of such an effect, the problem of which value of the mediator to choose arises. Indeed, the mediator is binary, assuming only two values, while its predicted values from model (1) with a logit link are probabilities, ranging in the continuum from 0 to 1. Which values to select then? This issue is addressed neither by [7] nor by [6]. We believe that the most appropriate solution consistent with a data generating mechanism of this type is to include binary values of the mediator obtained by the corresponding expected probabilities $\hat{\pi}_{M|X}$ by mean of a cutoff c such that

$$\hat{M}|X = \begin{cases} 1 & \text{if } \hat{\pi}_{M|X} \geq c \\ 0 & \text{if } \hat{\pi}_{M|X} < c \end{cases}$$

A possible criterion for the choice of c could be that of selecting the value for which the sensitivity and specificity of the classification are equal.

Covariates

The models in Equations (1)-(2) are intentionally very simple, but real-word data generally require adjustment for covariates. The inclusion of covariates $\boldsymbol{Z}$ in the mediator and the outcome models impacts on the expression of the indirect effect, which can depend on the covariates' values in addition to those of X and M. [7] suggest to estimate the indirect effects conditional on the values of X and M setting the covariates to their mean values.

The authors do not address the scenario where the mediation models include interactions between the exposure or the mediator and the covariates, i.e. when the covariates act as effect modifiers. This simply makes the partial derivatives in (3) more complex, but does not add any conceptual difficulty.

Confidence intervals for the indirect effect

Estimation of confidence intervals (CIs) for the indirect effect is not straightforward even in the linear case, since the distribution of the product $\beta_1\gamma_2$ does not follow a Normal distribution although the two coefficient estimators are assumed to be Normal. This issue is potentially exacerbated in a nonlinear setting. [6] suggest to use non-parametric bootstrap or Monte-Carlo confidence intervals. The former relies on the resampling of statistical units and the estimation of the parameter of interest in each of these samples, in order to obtain the empirical distribution of the conditional indirect effect. The latter does not require resampling, but assumes that the model coefficients in (1)-(2) comes from a multivariate Normal: by generating multiple samples of regression parameters it is possible to obtain a random sample of indirect effects. In both cases, percentiles can be used to derive 95% confidence intervals which respect the potential asymmetry of the indirect effect's distribution.

Supported by a growing body of literature, see for example [13], we claim that another option for the estimation of CIs could be the Bayesian approach. Each parameter is endowed with an *a priori* distribution and an empirical distribution of the indirect effect is obtained via Monte Carlo Markov Chains (MCMC). To the best of our knowledge, no simulation studies have been run so far to compare the performance of these three approaches. This is the primary focus of the next section.

4 Simulation study

We ran a small simulation study to highlight the differences between bootstrap, Monte Carlo and Bayesian CIs in terms of coverage rates and average interval length. We considered six scenarios obtained by varying the sample size ($n =$

30, 100, 200) and the type of distribution/link function for the mediator and the outcome (binary/logit and Poisson/log). X was generated from a Normal with 0 mean and standard deviation 5, and the values selected to compute conditional indirect effects are the mean and $\pm 1sd$. Both the mediator and the outcome are generated from a Bernoulli in a scenario and from a Poisson in the other one, with expectations as in models (1)-(2) and link functions logit and log, respectively. Model coefficients were chosen arbitrarily. Bootstrap estimates are obtained from 1,000 samples, Monte Carlo from 1,000 draws of regression coefficients, assumed to be distributed as a multivariate Normal and Bayesian estimates derive from two chains of length 10,000 with burnin = 5,000, using sparse priors. The number of iterations was set to 500. Results are shown in Figure 1 and in Table 1.

It can be noticed that for both distributions, the coverage rate of Monte Carlo CIs is higher than that of the other types in the scenario with the smallest sample size, while coverage rates are approximately the same in the other cases. As expected, the average CI lengths and the differences between the three methods reduce as the sample size increases. More detailed studies could help to make possible differences among the approaches emerge.

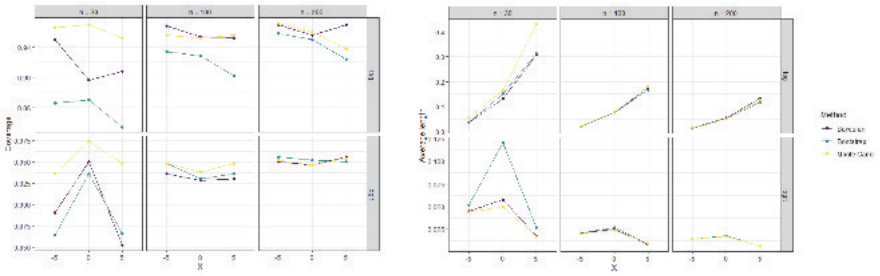


Fig. 1: Results of the simulation study: coverage rates on the left, average CI lengths on the right.

5 Conclusions

In this work, we deepened some aspects of the derivative-based method for mediation analysis in the presence of nonlinear models. A simulation study showed that bootstrap, Monte Carlo and Bayesian confidence intervals are all valid alternatives. The present work can be extended in several ways, for example focusing on the interpretation of conditional indirect effects when the exposure is categorical. In addition, it would be interesting to extend the proposed approach to settings with multiple mediators and to models with random effects.

Table 1: Results of the simulation study. For each method, the average coverage rates and interval lengths are reported.

Sample size	Link	X value	True eff.	Bayesian		Bootstrap		Monte Carlo	
				Cov. rate	Avg. length	Cov. rate	Avg. length	Cov. rate	Avg. length
30	logit	-5	[-0.005]	0.89	0.045	0.864	0.052	0.936	0.044
		0	[-0.007]	0.95	0.058	0.936	0.121	0.974	0.05
		5	[-0.002]	0.852	0.018	0.866	0.027	0.948	0.017
	log	-5	[-0.01]	0.95	0.037	0.866	0.036	0.966	0.054
		0	[-0.066]	0.896	0.13	0.87	0.15	0.97	0.167
		5	[-0.097]	0.908	0.307	0.834	0.313	0.952	0.43
100	logit	-5	[-0.005]	0.936	0.021	0.948	0.021	0.948	0.02
		0	[-0.007]	0.928	0.025	0.93	0.027	0.938	0.024
		5	[-0.002]	0.93	0.008	0.936	0.009	0.948	0.009
	log	-5	[-0.012]	0.968	0.021	0.934	0.019	0.956	0.021
		0	[-0.097]	0.954	0.075	0.928	0.076	0.952	0.077
		5	[-0.133]	0.952	0.175	0.902	0.167	0.956	0.181
200	logit	-5	[-0.005]	0.95	0.014	0.956	0.014	0.952	0.014
		0	[-0.007]	0.946	0.017	0.952	0.018	0.946	0.017
		5	[-0.002]	0.956	0.006	0.95	0.006	0.954	0.006
	log	-5	[-0.012]	0.97	0.014	0.958	0.013	0.972	0.014
		0	[-0.095]	0.956	0.054	0.95	0.051	0.96	0.052
		5	[-0.124]	0.97	0.133	0.924	0.118	0.938	0.124

References

1. Albert, J. M., Nelson, S.: Generalized Causal Mediation Analysis. Biom. **67**, 1028–1039 (2011)
2. Baron, R. M., Kenny, D. A.: The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. J. Personal. Soc. Psychol. **51**(6), 1173–1182 (1986)
3. Bollen, K. A.: Structural Equations with Latent Variables. Wiley, New York (1989)
4. Doretti, M., Raggi, M., Stanghellini, E.: Exact parametric causal mediation analysis for a binary outcome with a binary mediator. Stat. Methods Appl. **31**, 87–108 (2022)
5. Gaynor, S. M., Schwartz, J., Lin, X.: Mediation analysis for common binary outcomes. Stat. Med. **38**, 512–529 (2019)
6. Geldhof, G. J., Anthony, K. P., Selig, J. P., Mendez-Luck, C. A.: Accommodating binary and count variables in mediation: A case for conditional indirect effects. Int. J. Behav. Develop. **42**(2), 300–308 (2018)
7. Hayes, A. F., Preacher, K. J.: Quantifying and Testing Indirect Effects in Simple Mediation Models When the Constituent Paths Are Nonlinear. Multivar. Behav. Res. **45**, 627–660 (2010)
8. MacKinnon, D. P.: Introduction to Statistical Mediation Analysis. New York: Taylor and Francis Group (2008)
9. MacKinnon, D. P., Dwyer J. H.: Estimating mediated effects in prevention studies. Eval. Rev. **17**, 144–158 (1993)
10. Schluchter, M. D.: Flexible Approaches to Computing Mediated Effects in Generalized Linear Models: Generalized Estimating Equations and Bootstrapping. Multivar. Behav. Res. **43**(2), 268–288 (2008)
11. Stolzenberg, R. M.: The measurement and decomposition of causal effects in nonlinear and nonadditive models. Sociol. Methodol. **11**, 459–488 (1980)
    Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros. Psychol. Methods **18**(2), 137–150 (2013)
12. Wright, S.: The Method of Path Coefficients. Ann. Math. Stat. **5**(3), 161–215 (1934)
13. Yuan, Y., MacKinnon, D. P.: Bayesian Mediation Analysis. Psychol. Methods **14**(4), 301–322 (2009)

An exploration of Approximate Bayesian Computation (ABC) and dissimilarities

Un'esplorazione su Approximate Bayesian Computation (ABC) e sulle dissimilarità

Laura Bondi, Marco Bonetti and Raffaella Piccarreta

Abstract In this work we explore the use of dissimilarities among observations to measure the distance between two datasets. This is clearly relevant in the context of ABC, where observed and model-generated data need to be compared. In addition, we propose a new likelihood-free estimation procedure. Results of a simulation study on a simple model are presented.

Abstract *In questo lavoro esploriamo l'uso delle dissimilarità tra le osservazioni per misurare la distanza tra due dataset. Questo è chiaramente rilevante nel contesto di ABC, basato sul confronto tra dati osservati e dati generati da un modello. Proponiamo inoltre una nuova procedura di stima likelihood-free. Sono infine presentati i risultati delle simulazioni per un modello semplice.*

Key words: dissimilarities, ABC, likelihood-free, Wilcoxon-Mosler statistic

1 Introduction

Choosing the summary statistics and the distance function are fundamental issues for inference through Approximate Bayesian Computation (ABC) [10]. Ideally, the summary statistics should be sufficient statistics for the considered model. However, sufficient statistics are often unavailable or unknown, and this makes the choice of the summary statistic at the basis of retention of the generated parameter values one of the most delicate aspect of ABC. The literature on ABC offers several criteria to

Laura Bondi
MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, e-mail: laura.bondi@mrc-bsu.cam.ac.uk

Marco Bonetti
Department of Social and Political Sciences, Dondena Research Center, and Bocconi Institute for Data Science and Analytics, Università Bocconi, Milan, Italy, e-mail: marco.bonetti@unibocconi.it

Raffaella Piccarreta
Department of Decision Sciences, Dondena Research Center, and Bocconi Institute for Data Science and Analytics, Università Bocconi, Milan, Italy, e-mail: raffaella.piccarreta@unibocconi.it

select summary statistics or to reduce the dimensionality of the available summaries ([3], [10] and [5]).

Here we suggest a new direction to measure the dissimilarity between two datasets, based on the collection of the pairwise dissimilarities between their cases. Section 2 illustrates some techniques that can be used at this aim, based on a generic measure of dissimilarity and therefore applicable to different types of data. Section 3 describes a new estimation technique - alternative to ABC, but still likelihood-free - built from the same dissimilarity-based metric. The proposed estimator is based on calibration ideas, and is defined as the solution to a minimization problem. In Section 4 we apply the new dissimilarity-based estimation procedure to a simple model: the bivariate normal case. We close with some discussion in Section 5.

2 Dissimilarity-based criteria

Consider an iid sample $y = (y_1, \ldots, y_n)^T$, where each y_i contains information - possibly multivariate - on to the i-th individual. Given the summary statistic $S = S(y)$, in ABC one retains the values of the parameter θ generating a sample with a value of S *close* to the observed data, $S(y_{obs})$. For a fixed value of θ, we focus on the *whole* distribution $f_{S|\theta}(s|\theta)$ of $S(Y)$.

More generally, for a given y_{obs} and for a value y of the random vector $Y \sim f_Y(y|\theta)$, we define a quantity $T(S(y); S(y_{obs}))$ that can be used to measure how far y is from y_{obs}. Note that S can also be the identity transformation, thus $T(y; y_{obs})$. For example, one may set T to be the L^2-distance in Euclidean space. ABC retains the values of θ that generated a sample y such that $T(y, y_{obs}) \approx 0$, i.e. such that $S(y) \approx S(y_{obs})$. The selected parameter values represent a sample from the approximated posterior distribution, and can be *adjusted* to account for the discrepancy between simulated and observed statistics [2, 7].

We aim at introducing alternative possibilities for the definition of $T(y; y_{obs})$. Depending on the kind of data, one can define a one-dimensional dissimilarity measure between two observations y_i and y_j, $d_{ij} = D(y_i, y_j)$. For example, the dissimilarity might be computed as the L^2-distance, or by other distances defined on vector spaces (with the triangle inequality not being strictly necessary). The distribution of the dissimilarity between two randomly selected observations in a population is called the *interpoint distance distribution (IDD)* (see, e.g. [4]).

We propose to use the set of all pairwise dissimilarities $\{d_{ij}\}_{i,j=1,\ldots,n}$ to define a function T measuring the distance between two datasets, to be used in ABC. One can distinguish among three kinds of dissimilarities:

i) within-group dissimilarities within the observed data, i.e. all the pairwise dissimilarities between observed subjects' data;
ii) within-group dissimilarities within the generated data i.e. all the pairwise dissimilarities between generated subjects' data;
iii) between-group dissimilarities, i.e. all the pairwise dissimilarities between an observed and a generated subject's data.

A first approach consists in defining $T(y;y_{obs})$ using the Wilcoxon-Mosler statistic (WM), see [8], that contrasts within-group and between-group dissimilarities through a rank-based test statistic. Specifically, the set including all the dissimilarities (both the within- and the between-group) is sorted in ascending order and ranked, and the test statistic WM is defined as the sum of the ranks of the between-group dissimilarities:

$$\mathrm{WM} = \sum_{i\in\{obs\}} \sum_{j\in\{gen\}} \mathrm{rank}\,(d_{ij})\,, \tag{1}$$

When observed and model-generated data are similar, the ranks of the between-group dissimilarities should be placed at random. In particular, under the null scenario corresponding to the case when the true parameter (and model) generated the data, the pairwise dissimilarities are identically distributed - although not independent. One can show that under permutation of the $m+n$ group labels the statistic WM has expected value

$$E(\mathrm{WM}) = \frac{mn}{2}\left[\binom{m+n}{2}+1\right].$$

For our purposes, no testing of hypotheses, but only quantifying the distance between two datasets is needed. In particular, we define the distance as the squared difference between the observed WM and its expected value under the null hypothesis that observed and simulated data were generated by the same parameter value, i.e. $T(y;y_{\mathrm{obs}}) = [\mathrm{WM} - E(\mathrm{WM})]^2$.

Another approach to quantify the similarity between the observed and simulated data relies on the distance of the estimated cumulative distribution function of the within-group dissimilarities from the cumulative distribution function of the between-group dissimilarities, that can be calculated using one of many available distances, such as the Kolmogorv-Smirnov (KS) statistic. The similarity between model-generated and observed data implies similarity between the distributions of the between-group and of the within-group dissimilarities in the observed data, and a small KS distance.

In the next section we elaborate further on the use of these metrics in ABC, yielding a natural alternative estimator for θ.

3 A new ABC-inspired estimator

For any given θ and y_{obs}, there exists a whole distribution of the random variable $T(Y,y_{obs})$. We now exploit this noting that if T is a *good* measure of the distance of the random vector Y from the sample y_{obs}, its distribution over a range of values of θ can be used to estimate the true value of θ.

Specifically, we may estimate the quantile of order $\tau \in (0,1)$, of the distribution of T conditional to the parameter value θ, $q_\tau(T\,|\,\theta) = \mathrm{argmin}_a E[\rho_\tau(T-a)\,|\,\theta]$. Here,

$\rho_\tau(\cdot)$ is the so-called *check function* and it is given by $\rho_\tau(z) = z(\tau(1 - 1_{\{z<0\}}))$. This is known as quantile regression (see [6] for details). In particular, we perform quantile regression marginally for each parameter component θ (note that one may also apply the multivariate version of quantile regression) and we fit a local linear quantile regression curve $\widehat{q}_\tau(\theta)$, as proposed in [11].

Given the estimated conditional quantile of order τ of the statistic T given θ, we define a new estimator $\widehat{\theta}$ of θ as the solution of the following minimization problem:

$$\widehat{\theta}_\tau(y_{obs}) = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \widehat{q}_\tau(\theta; y_{obs}). \tag{2}$$

This construction is motivated by ABC. Indeed, one would only retain parameter values that produce y's with the *smallest* distance from y_{obs}. One may choose τ to be relatively close to zero (say, $\tau = 0.1$ or 0.05) , and estimate $\widehat{q}_\tau(\theta; y_{obs})$ through quantile regression from the ABC-produced samples (thus exploiting the computational effort already spent to perform ABC), or from samples generated over a grid of values of θ. Observe that as an estimator, $\widehat{\theta}_\tau = \widehat{\theta}_\tau(Y_{obs})$ with $Y_{obs} \sim f_Y(y;\theta^*)$, with θ^* the true value of θ. Also, $\widehat{\theta}_\tau$ clearly depends on the choice of τ.

Note that, if the objective function to be minimized in (2) is regular enough, thus admitting one minimum, it can be obtained by setting the (partial) derivatives of the objective function with respect to θ equal to zero. Clearly, the asymptotic properties, such as consistency and asymptotic normality, of the proposed estimators need to be investigated. Whilst not tackling this issue here, we believe that useful theoretical results can be found in the framework of the theory of estimating equations.

4 The bivariate normal model

We present an example of application of a dissimilarity-based metric in the framework of ABC, and we illustrate how the new estimator introduced in Section 3 can be computed. Let us consider the problem of estimating the centroid $\mu = (\mu_1, \mu_2)^T$ of a bivariate normal distribution with known covariance matrix Σ, $N_2(\mu, \Sigma)$. Let $y_{\text{obs}} = (y_{1i}, y_{2i})_{i=1,\ldots,n}$ be a sample of n observations from the distribution.

To evaluate whether our procedure effectively recovers μ, we set $\mu = (1,2)^T$ and $\Sigma = \text{diag}(2,3)$ and generate a sample of size $n = 100$. We set prior distributions for the unknown parameters $\mu_i \sim N(0,4)$, $i = 1,2$, and generate $K = 10{,}000$ datasets from the model, each one corresponding to a value of μ sampled from the prior distribution (and keeping Σ fixed). We define the pairwise dissimilarity between observations as the L^2-distance, and compute the WM statistic.

We are interested to assess how well the WM distance discriminates among the parameter values proposed by the prior distribution. Figure 1 shows the relationship between the values of the parameters and the distance from the observed data, marginally for the two components of μ.

The blue curves in the figure represent the estimates of the conditional quantiles of order $\tau = 0.1$. The plots in Figure 1 suggest that the metric based on the WM statistic is quite informative about the unknown parameter μ. The minimum dis-

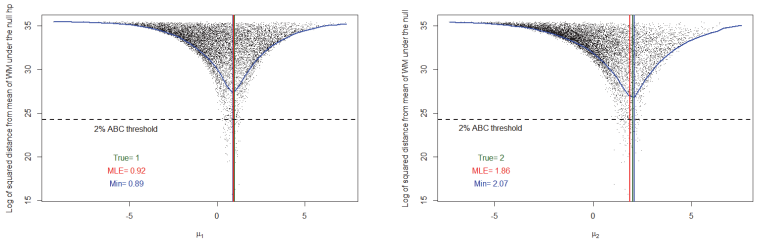
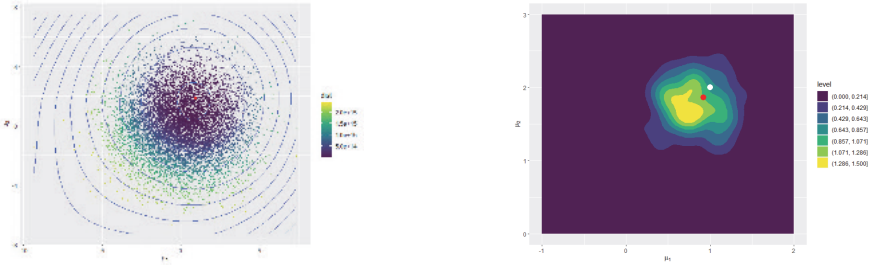


Fig. 1: Samples from the prior distribution of the two mean components, and Wilcoxon-Mosler distance (in log-scale) of each generated dataset from the observed data. The blue curve is the estimated conditional quantile of order $\tau = 0.1$.

tances are indeed reached for parameter values that are very close to the true ones $(\mu_1 = 1, \mu_2 = 2)$. The figure also shows the suggested estimator for each parameter component (blue vertical lines), i.e. the values for whom the minima of the quantile regression curves are reached (see Section 3), as well as the values of the true parameters (green vertical lines) and of the maximum likelihood estimates (red vertical lines). Note that some of the lines are not visible too clearly because of the overlapping with the others. Figures 2a and 2b report respectively the measure T, together with its level curves, as a function of μ_1 and μ_2 simultaneously, and the level curves of the joint posterior distribution for the parameters obtained through ABC when retaining the best 2% of the values.



(a) Bivariate sample from the prior distribution of μ colored according to the Wilcoxon-Mosler distance, whose level curves are also plotted.

(b) Joint posterior distribution obtained from ABC with retention rate of 0.02. The white dot shows the true data generating parameter and the red one the MLE.

Fig. 2: Posterior distributions for the bivariate normal model with known covariance.

Repeating the experiment with $n = 1000$ suggests that, as expected, the amount of *information* provided by the observed data increases with the sample size, and that the T-based criterion captures that fact.

5 Conclusions

While being only a proof-of-concept, this work suggests that the proposed estimation procedure is promising, and its properties and generalization to more complex models deserve to be investigated.

Importantly, criteria are available to suitably assess the extent of dissimilarity between cases and between two datasets based on the data specific characteristics. An interesting extension of the methods proposed in this article concerns the case of longitudinal data, i.e. when for each subject a sequence of states is observed over discrete time. Several dissimilarity measures among sequences can be defined. Among them, one can test the performance of measures exploiting the continuous nature of the data, such as correlation, as well as measures based on values discretized in bins. In the latter case, a very common choice is the Optimal Matching (OM) algorithm, introduced in molecular biology to study proteins and DNA sequences [9], and then extended to sociology [1]. Such alignment technique, based on the effort needed to transform a discrete sequence into another, can be used to construct a metric for ABC for sequence data.

References

[1] Abbott, A. "Sequence Analysis: New Methods for Old Ideas". In: *Annual Review of Sociology* 21 (1995), pp. 93–113.

[2] Beaumont, MA, Zhang, W, and Balding, DJ. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035.

[3] Blum, MGB et al. "A comparative review of dimension reduction methods in approximate bayesian computation". In: *Statistical Science* 28.2 (2013).

[4] Bonetti, M and Pagano, M. "The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection." In: *Statistics in Medicine* 24.5 (2005), pp. 753–73.

[5] Gutmann, MU et al. "Likelihood-free inference via classification". In: *Statistics and Computing* 28.2 (2018), pp. 411–425.

[6] Koenker, R. *Quantile Regression*. Cambridge University Press, 2005.

[7] Lintusaari, J et al. "Fundamentals and recent developments in approximate Bayesian computation". In: *Systematic Biology* 66.1 (2017), e66–e82.

[8] Mosler, K. *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach.* New York: Springer-Verlag, 2002.

[9] Sankoff, D., Kruskal, J., and Nerbonne, J. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. The David Hume Series. Cambridge University Press, 2000.

[10] Sisson, SA, Fan, Y, and Beaumont, M. *Handbook of Approximate Bayesian Computation*. New York: Chapman and Hall/CRC, 2018.

[11] Yu, K and Jones, MC. "Local Linear Quantile Regression". In: *Journal of the American Statistical Association* 93.441 (1998), pp. 228–237.

Advances in Categorical and Preference data

On the predictability of a class of ordinal data models

Sulla capacità previsiva di una classe modelli per dati ordinali

Rosaria Simone and Domenico Piccolo

Abstract The contribution aims at being a pilot study on the prediction performance for the class of mixture models with uncertainty, in order to assess if and to what extent the uncertainty specification provides an added value for model prediction. To this aim, a small simulation study is designed to assess prediction performance of competing models under miss-specification; further, a variable selection procedure based on prediction performance is outlined on a case study.
Abstract *Il contributo propone uno studio preliminare sulla capacità previsiva di una classe di modelli statistici per dati ordinali, usando sia dati simulati che dati reali. In particolare, il valore aggiunto dell'incertezza nella specificazione del modello viene potenziato con l'implementazione di una procedura di selezione delle variabili basata sulle performance predittive.*

Key words: CUB models, Rating Data, Prediction, Variable selection

1 Introduction

The contribution is meant as a pilot study on the added value of uncertainty specification on predictability of the class of mixture models with uncertainty for ratings. The reference modelling framework and the chosen indicators of prediction performance are recalled in Section 2. A pilot simulation experiment is planned in Section 3 to compare the sensitivity, under model mis-specification, of a well-known scoring rule for ordinal responses and of the normalized dissimilarity index used as a measure of prediction error of a given statistical model w.r.t. to a rating distribution. As

Simone R.
Department of Political Sciences, University of Naples Federico II e-mail: rosaria.simone@unina.it

Piccolo D.
Department of Political Sciences, University of Naples Federico II e-mail: dopiccol@unina.it

a by-product, the effect of covariate specification on prediction performance can be exploited to implement a variable selection procedure (see Section 4): comparisons with cumulative link models is reported for completeness.

2 Methods

The class of CUB mixture models allows to analyse both feeling and uncertainty of the rating process, assuming that the Binomial model is suitable for the latent sentiment as expressed on the discrete support, say with m ordered categories. It can be adjusted for over-dispersion by assuming the Beta-Binomial (CUBE), or mixed with a degenerate distribution to account for inflated frequencies (CUB with shelter): see [8] for an updated overview. Starting from the (shifted) Binomial distribution $b_r(\xi), r = 1,\ldots,m$, with probability parameter ξ, a CUB model for the rating response R is defined by its mixture with a discrete uniform distribution:

$$Pr(R = r|\boldsymbol{\theta}) = \pi_i\, b_r(\xi_i) + (1-\pi_i)\,\frac{1}{m} \tag{1}$$

where parameters π_i, ξ_i can be possibly linked to (row vectors of) subjects' covariates $\boldsymbol{y}_i, \boldsymbol{x}_i$ with a logit link $logit(\pi_i) = \boldsymbol{y}_i\boldsymbol{\beta}$, $logit(\xi_i) = \boldsymbol{x}_i\boldsymbol{\gamma}$. Otherwise, CUB models allow to estimate overall feeling $1-\xi$ and overall uncertainty weight $1-\pi$, meant as importance of the heterogeneity of the distribution not accounted for by the feeling model[1]. For this reason, CUB uncertainty is deemed to play a crucial role in enhancing both fitting and prediction performance of the feeling model. In case covariates are not specified in the model, then the (normalized) dissimilarity index $Diss(\boldsymbol{p},\boldsymbol{f}) = \frac{1}{2}\sum_{r=1}^{m}|f_r - p_r(\boldsymbol{\theta})| \in [0,1]$ allows to assess the extent by which a discrete model $\boldsymbol{p}$ matches category-wise an observed (relative) frequency distribution $\boldsymbol{f}$. For the purpose of the paper, the sensitivity of the Dissimilarity index under miss-specification will be tested when it is used as a measure of prediction accuracy of a model $\boldsymbol{p}$ estimated on a training set w.r.t. the frequency of a validation set.

2.1 The Ranked Probability score for prediction performance

We resort to the Rank Probability Score (RPS, [3, 7, 4]) as a scoring rule specifically designed for ordinal distributions to assess predictive performance of selected models. With the above notation, the RPS for a new observation $R^\star$, based on a fitted probability model $\boldsymbol{p} = (p_1,\ldots,p_m)$ for R, is defined as:

[1] If $p \geq 0$ and $q \geq 0$ covariates are specified for uncertainty and feeling parameters respectively, then we refer to CUB(p,q) model.

$$RPS_{R^\star} = \frac{1}{m-1}\sum_{k=1}^{m}\left(\sum_{s=1}^{k} p_s - \sum_{s=1}^{k} 1_{R^\star=s}\right)^2 \quad (2)$$

The model that has the lowest total RPS score, summed over all observations in the test set, is the one providing the best prediction performance. It is among the most popular prediction indicators for ordinal classifiers, as random forests and trees [5, 6].

3 A simulation study

This section is tailored to assess the sensitivity of prediction performance granted by the RPS score and the dissimilarity index between the model estimated on the training set and the frequency distribution of the test set, showing that the two measures of prediction error behave similarly. In particular, we have planned a simulation experiment to evaluate the effect of uncertainty specification on model-prediction, by sampling $n = 500$ observations from each generating model X for 100 times: Binomial, CUB, CUB with shelter, and CUBE for overdispersion. In the latter two cases, performances of Binomial with shelter and Beta-Binomial are reported for comparative purposes. The sample is then split into training and test set: after candidate models are fitted on the training set, the dissimilarity of each of them with the frequency distribution of both training and test set is computed. Finally, the RPS under each estimated model is computed for each observation in the test set, and then summed. The following list reports the average over simulation runs of both the dissimilarity index and of the RPS for each simulation scenario:

- Starting from a Binomial, with $m = 8$, it can be seen that uncertainty specification adjusts prediction performance of the feeling component, with an improvement that increases with the heterogeneity of the distribution (the relative gain amounts to 1.56% if $\pi = 0.7$ to 5.8% if $\pi = 0.3$ and to 8.32% if $\pi = 0.1$) (see Table 1);

	$X \sim Bin(0.7)$		$X \sim CUB(\pi = 0.7, \xi = 0.6)$		$X \sim CUB(\pi = 0.3, \xi = 0.6)$		$X \sim CUB(\pi = 0.1, \xi = 0.6)$	
Fit:	Binomial	CUB	Binomial	CUB	Binomial	CUB	Binomial	CUB
Average RPS	14.158	14.158	20.348	20.031	26.718	25.160	29.675	27.206
Diss - train set	0.042	0.042	0.131	0.045	0.278	0.047	0.346	0.045
Diss - test set	0.067	0.067	0.141	0.082	0.274	0.088	0.340	0.088

Table 1 Comparing prediction errors of Binomial and CUB models for data sampled from X

- Sampling from a CUBE, it can be seen that the sole over-dispersion specification does not improve prediction performances, whereas this goal is attained when mixing either Binomial or Beta-Binomial with uncertainty (see Table 2):
- Sampling from CUB with shelter at $c = 5$, with $m = 7$, it can be seen that the sole shelter specification does not improve prediction performance, whereas this goal is attained when mixing feeling with uncertainty (see Table 3):

	$CUBE(\pi=0.6,\xi=0.4,\phi=0.1);m=7$				$CUBE(\pi=0.2,\xi=0.4,\phi=0.1);m=7$			
Fit:	Binomial	CUB	Beta-Binomial	CUBE	Binomial	CUB	Beta-Binomial	CUBE
Average RPS	24.492	23.457	24.642	23.429	29.907	27.313	30.080	27.314
Diss - train set	0.238	0.056	0.058	0.042	29.907	27.313	30.080	27.314
Diss - test set	0.235	0.092	0.093	0.089	0.352	0.088	0.086	0.086

Table 2 Comparing prediction errors of Binomial, Beta-Binomial, CUB and CUBE models for data sampled from a CUBE model X

	$\pi^\star=0.3,\xi=0.1,\delta=0.1$				$\pi^\star=0.6,\xi=0.1,\delta=0.1$			
Fit:	Binomial	CUB	Bin + shelter	CUB with shelter	Binomial	CUB	Bin with shelter	CUB with shelter
Average RPS	29.477	26.949	29.514	26.76	25.005	23.026	25.019	22.869
Diss - train set	0.367	0.097	0.369	0.033	0.319	0.103	0.319	0.030
Diss - test set	0.363	0.134	0.365	0.090	0.321	0.135	0.321	0.086

Table 3 Comparing prediction errors of Binomial and CUB models, with and without shelter, for data sampled from a CUB with shelter X

4 On the prediction of subjective probability to survive

It is well-known that people are biased with respect to self-reported life expectancy and this affects financial behavior, pension participation and savings decision. For this reason, the determination of drivers of subjective probability to survive at given age is of strategic importance. To this aim, we consider a survey carried out in 2006 by ISFOL[2], and analyse a sample of $n = 20184$ validated answers to the question: '*In your opinion, what is the probability that you will reach age 90? Please provide a value between 0 (impossible event) and 100 (certain event).*' The continuous measurement was then discretized on a scale with $m = 7$ categories[3]. See [9] for a comparative discussion on fitting and explanatory performance of CUB and cumulative models on this dataset. The average RPS over 10 folds for the Binomial model without covariates amounts to $RPS_{Bin} = 233.545$, whereas $RPS_{CUB} = 229.036$, yielding a relevant decrease in prediction error. On a similar note, the dissimilarity between observed frequency distribution of the validation and model fitted on the training set, averaged over folds, decreases from $Diss(\boldsymbol{f}_{test}, Bin_{train}) = 0.161$ to $Diss(\boldsymbol{f}_{test}, CUB_{train}) = 0.088$. The first step of the proposed variable selection procedure is the identification of the variable that - for a one covariate model - performs the best. Table 4 reports results for competing models: Binomial with covariate, $CUB(0,1)$ with covariate only for feeling, $CUB(1,1)$ with the given covariate specified for both model components, and POM for comparison purposes, highlighting that uncertainty specification, beyond fitting and explanatory performance, improves prediction ability, or at least behaves equivalently to POM.

If we consider all covariates in model specification, we can verify if and to what extent prediction improves when switching from Binomial with covariates to $CUB(0,q)$ with covariates only for feeling and overall uncertainty, and similarly

[2] Institute for training of workers, Ministry of Labour and Welfare, Italy

[3] Class endpoints were set to $0,5,25,45,55,75,95,100$: the resulting relative frequency distribution $(0.095,0.132,0.150,0.166,0.193,0.144,0.120)$ is pretty heterogeneous.

	Gender	Age	No head	Married	Has children	Employed
Binomial	372.709	370.679	372.214	371.096	370.651	372.332
CUB$(0,1)$	351.419	350.436	351.252	350.639	350.334	351.240
CUB$(1,1)$	351.388	348.276	350.750	349.700	349.463	351.051
POM	351.353	349.693	350.944	350.041	349.673	351.059

Table 4 Average over 10-folds of the RPS score for competing models with given covariate

we can assess the gain of switching from CUB$(0,q)$ to CUB(q,q) with the same covariate set for both components: see Table 5.

Binomial	$CUB(0,q)$	$CUB(q,q)$	POM
425.9020	350.1574	348.0456	349.2173

Table 5 Average over 10-folds of the RPS score for competing models with all covariates specified either only for feeling or for both components

Finally, we report in Table 6 the forward selection results obtained by including, at each step, the variable for which the minimum RPS is attained: at each stage, the best performance corresponds to $CUB(p,q)$ with the same set of covariates for both feeling and uncertainty[4]: age, marital status, gender and having children are the most useful factors to predict subjective probability to survive.

	Age	Married	Female	Has children	No head	Work
Average RPS over folds	141.088	141.045	141.017	140.991	140.991	140.991

Table 6 RPS-based variable selection for responses of individuals living in southern Italy: covariates are listed with respect to their entering order in the procedure

For completeness, Table 7 reports the prediction performance for the best fitting model (which foresees an age effect for uncertainty together with differences implied by gender and marital status on feeling) compared to those of POM, Binomial, $CUB(0,q)$, CUB$(p,0)$ (for the latter two, we used only covariates for the corresponding component of the best model) [5].

	Binomial	CUB$(0,q)$	CUB$(p,0)$	CUB(p,q)	POM
Average RPS over folds	150.211	141.777	141.569	140.997	142.994

Table 7 Average RPS over folds for the best fitting CUB(p,q)

[4] For the sake of illustration, we report only results corresponding to respondents that live in Southern Italy ($n = 8107$).

[5] We have resorted to the fast estimation procedure and the corresponding best-subset variable selection available within the `R` library `FastCUB` [11].

5 Further developments

This preliminary research focuses on model-based prediction for rating data to implement a variable selection procedure: on the topics, see [10, 12] on the basis of random forest variable importance for numeric variables and ROC curves for binary outcomes, respectively; see [1] for a Bayesian research on the best predictive normal linear model. Further developments include comparisons between the best model attained with the proposed strategy and prediction achieved via model-based random forests based on CUB model regression trees [2], for instance.

References

1. Barbieri M.M., Berger, J.O. Optimal predictive model selection. Ann. Statist. **32**(3):870–897 (2004)
2. Cappelli C., Simone R., Di Iorio F. CUBREMOT: a model-based tree for ordinal responses. Expert Systems with Applications, **124**, 39–49 (2019)
3. Epstein E.S. A scoring system for probability forecasts of ranked categories, Journal of Applied Meteorology, **8**(6), 985-987 (1969)
4. Gneiting T. Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc, **102**(477):359–378 (2007)
5. Hornung R. Ordinal Forests. Journal of Classification **37**, 4–17 (2020)
6. Janitza S., Tutz G., Boulesteix A.L. Random forest for ordinal responses: Prediction and variable selection. Computational Statistics & Data Analysis, **96**, 57–73 (2016)
7. Murphy A. A note on the ranked probability score. Journal of Applied Meteorology, **10**, 155–15 (1971)
8. Piccolo D., Simone R. The class of CUB models: statistical foundations, inferential issues and empirical evidence. Statistical Method and Applications, **28**, 389–435 (2019)
9. Piccolo D., Simone R., Iannario M. Cumulative and CUB models for rating data: a comparative analysis. International Statistical Review, **87**(2), 207–236 (2019)
10. Sandri M., Zuccolotto P. Variable Selection Using Random Forests. In: Zani S., Cerioli A., Riani M., Vichi M. (eds.) Data Analysis, Classification and the Forward Search. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-35978-8_30 (2006)
11. Simone R. FastCUB: Fast EM and Best-Subset Selection for CUB Models for Rating Data. R package version 0.0.2. https://CRAN.R-project.org/package=FastCUB (2020)
12. Zhou X.H., Chen B., Xie Y.M., Tian F., Liu H. and Liang X. Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoporosis disease. Statistics in Medicine, **31**: 628–635 (2012)

Multivariate analysis of binary ordinal data using graphical models

Analisi multivariata di dati binari ordinali attraverso l'impiego di modelli grafici

Camilla Caroni, Fabio Alberto Comazzi, Andrea Deretti, Federico Castelletti

Abstract In this contribution we address the issue of inferring dependence relations between ordered binary variables. We propose a Bayesian model specification which assumes that ordinal data are generated by discretization of latent Gaussian data and that the joint density of the latent variables satisfies the conditional independencies imposed by a graphical model. We then propose an MCMC strategy for joint structural (graph) learning and parameter estimation.

Abstract *In questo contributo si considera il problema di stima di relazioni di dipendenza tra variabili binarie ordinali. Si propone un modello bayesiano, il quale assume che dati ordinali siano generati attraverso discretizzazione di osservazioni latenti gaussiane e che la distribuzione congiunta delle variabili latenti soddisfi relazioni di dipendenza imposte da un modello grafico. Si propone quindi un algoritmo di tipo MCMC per l'apprendimento della struttura grafica e l'inferenza sui parametri del modello.*

Key words: Structural learning, Ordinal data, Graphical model

Camilla Caroni
Politecnico di Milano, camilla.caroni@mail.polimi.it

Fabio Alberto Comazzi
Politecnico di Milano, fabioalberto.comazzi@mail.polimi.it

Andrea Deretti
Politecnico di Milano, andrea.deretti@mail.polimi.it

Federico Castelletti
Università Cattolica del Sacro Cuore, federico.castelletti@unicatt.it

1 Introduction

Modelling dependence relations between variables represents an important issue in many applied domains, and a variety of statistical methods have been proposed for this purpose. Typically the structure of dependencies is unknown and accordingly it must be inferred from the available data. To this end, probabilistic graphical models [6] adopt a graph-based representation of conditional independence relations among variables which are embedded in the joint density through a suitable graph-factorization. Specifically, a graph $\mathscr{G} = (V, E)$ consists of a set of nodes V associated to variables in the system, and a set of edges E, representing dependence relations. The goal is therefore to *learn* the graphical structure $\mathscr{G}$ given the data.

Most of the available methodologies however, both frequentist and Bayesian, deal under the assumption that Gaussian or categorical data are collected; see for instance [4], [5] and [7] for a Bayesian approach. More recently, a few methods for *mixed* (Gaussian and categorical) data have been proposed; see for instance [2] for a frequentist methodology. Still in the categorical framework, the majority of the literature has considered the case of *unordered* data which are represented as contingency tables of counts. In the Bayesian framework, methodologies based on suitable extensions of the Multinomial-Dirichlet model have been proposed for the analysis of unordered categorical data; see for instance [1] and references therein.

This contribution summarizes our first attempt to build a Bayesian framework for the multivariate analysis of categorical *ordinal* variables. For simplicity we assume all variables being *binary*, while in the last section of the work we provide some remarks for possible extensions to a general setting with ordered *polytomous* variables. Ordinal variables, namely variables whose levels are arranged according to a given ordering are quite common in many contexts, specifically in social sciences and psychology. In this framework, the adoption of standard methods for the analysis of categorical unordered data, although possible, is clearly not satisfactory, since the whole information encoded in the ordering of the data is lost.

The key assumption of our method is that the observed categorical variables $X_1, \dots X_q$ are generated by *discretization* of latent Gaussian random variables $Z_1, \dots,$ Z_q, whose joint density satisfies the conditional independence relations imposed by the graph $\mathscr{G}$. The main advantage of the proposed methodology is that we are able to infer both the underlying network structure as well as the *sign* of the dependence relation between pairs of variables. Finally, being fully Bayesian, a coherent quantification of the uncertainty around parameter estimates is provided.

2 Model formulation

In this section we introduce our model specification, which is based on the assumption that ordinal binary data are generated by thresholding of latent Gaussian observations.

Let $(X_1,\ldots,X_q)$ be a collection of ordinal (binary) variables such that $X_j \in \{0,1\}$ for each $j=1,\ldots,q$. Let also $(Z_1,\ldots,Z_q)$ be q continuous variables, each associated with one of the q binary variables. We assume that each binary variable is obtained by *discretization* of its latent counterpart as

$$X_j = \begin{cases} 0 & \text{if } Z_j < \theta_0^{(j)} \\ 1 & \text{if } Z_j \geq \theta_0^{(j)} \end{cases} \tag{1}$$

where Z_j is the (latent) Gaussian random variable associated with X_j and $\theta_0^{(j)} \in (-\infty,\infty)$ represents an unknown cut-off parameter. Observed data then consist of $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ where $\mathbf{x}^{(i)} = \left(x_1^{(i)},\ldots,x_q^{(i)}\right)^\top$ denotes the i-th realization of the random vector $(X_1,\ldots X_q)$, row of the (n,q) data matrix $\mathbb{X}$.

In what follows we assume that the joint distribution of the latent variables is multivariate Gaussian and that the model-parameter (covariance matrix Σ) satisfies the constraints imposed by a graphical model $\mathscr{G}$. Specifically, we rely on decomposable undirected graphical models (UGs) [3] and write

$$\begin{aligned} (Z_1,\ldots,Z_q) \mid \Sigma,\mathscr{G} &\sim \mathscr{N}_q(\mathbf{0},\Sigma) \\ \Sigma \mid \mathscr{G} &\sim \mathrm{HIW}(b,D), \end{aligned} \tag{2}$$

where HIW denotes the *Hyper Inverse Wishart* prior distribution. In addition, under the decomposable UG $\mathscr{G}$, the joint density of $(Z_1,\ldots,Z_q)$ factorizes as

$$p(\mathbf{z} \mid \Sigma,\mathscr{G}) \propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_C^T\Sigma_C^{-1}\mathbf{z}_C\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_S^T\Sigma_S^{-1}\mathbf{z}_S\right\}}, \tag{3}$$

where $\mathscr{C}$ and $\mathscr{S}$ denotes the sets of cliques and separators respectively of $\mathscr{G}$. We refer the reader to [3] for full details.

Because of (1), the *augmented* density of $(X_1,\ldots,X_q)$ and $(Z_1,\ldots,Z_q)$ can be written as

$$\begin{aligned} p(\mathbf{x},\mathbf{z} \mid \Sigma,\Theta,\mathscr{G}) &= p(\mathbf{z} \mid \Sigma,\mathscr{G})p(\mathbf{x} \mid \mathbf{z},\Theta) \\ &= p(\mathbf{z} \mid \Sigma,\mathscr{G})\mathbf{1}\left\{\mathbf{z} \in C(\mathbf{x},\Theta)\right\}, \end{aligned} \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $C(\mathbf{x},\Theta)$ is defined as

$$C(\mathbf{x},\Theta) = \left[\theta_{x_1-1}^{(1)},\theta_{x_1}^{(1)}\right) \times \left[\theta_{x_2-1}^{(2)},\theta_{x_2}^{(2)}\right) \times \cdots \times \left[\theta_{x_q-1}^{(q)},\theta_{x_q}^{(q)}\right),$$

and for each $j=1,\ldots,q$ we adopt the notation $\theta_{-1}^{(j)} = -\infty$, $\theta_1^{(j)} = +\infty$. Given (3) we can write explicitly

$$p(\mathbf{x},\mathbf{z} \mid \Sigma,\Theta,\mathscr{G}) \propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_C^T\Sigma_C^{-1}\mathbf{z}_C\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{z}_S^T\Sigma_S^{-1}\mathbf{z}_S\right\}} \mathbf{1}\left\{\mathbf{z} \in C(\mathbf{x},\Theta)\right\}. \tag{5}$$

Consider now the (n,q) data matrix $\mathbb{X}$ collecting n i.i.d. (binary) observations $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ from (4) and the corresponding (n,q) latent data matrix $\mathbb{Z}$. Then, we can write the *augmented likelihood* as

$$p(\mathbb{X},\mathbb{Z} \mid \Sigma,\Theta,\mathscr{G}) = p(\mathbb{Z} \mid \Sigma,\mathscr{G}) \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C\big(\mathbf{x}^{(i)},\Theta\big)\right\}$$
$$\propto \frac{\prod_{C\in\mathscr{C}} |\Sigma_C|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_C^{-1}S_C)\right\}}{\prod_{S\in\mathscr{S}} |\Sigma_S|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_S^{-1}S_S)\right\}} \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)},\Theta)\right\},$$

where $S = \mathbb{Z}^T\mathbb{Z}$ and S_C denotes the sub-matrix of S with columns and rows indexed by $C \subseteq \{1,\ldots,q\}$.

Now remember that $\theta_0^{(j)}$, $j = 1,\ldots,q$, are (unknown) random thresholds linking the latent data to the binary observations. As a prior distribution we then assume

$$\theta_0^{(j)} \overset{\text{iid}}{\sim} \mathscr{N}(0,\tau^2), \quad j = 1,\ldots,q. \tag{6}$$

We complete our model specification by assigning a prior to graph $\mathscr{G}$, for any $\mathscr{G} \in \mathbb{S}_q$, the space of decomposable UGs on q nodes. Specifically, let $A_{u,v}$ be the 0-1 (u,v)-element of the lower triangular adjacency matrix of $\mathscr{G}$. We assign hierarchically

$$A_{u,v} \mid \pi \overset{iid}{\sim} \mathrm{Ber}(\pi),$$
$$\pi \sim \mathrm{Beta}(a,b),$$

which corresponds to a multiplicity-correction prior on graph $\mathscr{G}$; see also [8].

3 Posterior inference

Starting from the model formulation introduced in the previous section we can consider the posterior distribution

$$p(\Sigma,\Theta,\mathscr{G},\mathbb{Z} \mid \mathbb{X}) = p(\mathbb{Z} \mid \Sigma,\mathscr{G}) \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)},\Theta)\right\} p(\Sigma \mid \mathscr{G}) p(\mathscr{G}) \prod_{j=1}^{q} p(\theta_0^{(j)}),$$

where the latent data $\mathbb{Z}$, since unobserved, are also included among the "parameters" of the model. Because direct sampling from the latter distribution is not possible, we propose to implement a block Gibbs sampler scheme coupled with a Metropolis Hasting step to sample from the full conditional distributions of the model parameters. For convenience, we consider the two sets of parameters $(\Sigma,\mathscr{G})$ and $(\mathbb{Z},\Theta)$.

To start with, the full conditional of $(\Sigma,\mathscr{G})$ can be written as

$$\begin{aligned} p(\Sigma,\mathscr{G} \mid \Theta,\mathbb{Z},\mathbb{X}) &= p(\Sigma,\mathscr{G} \mid \mathbb{Z}) \\ &\propto p(\mathbb{Z} \mid \Sigma,\mathscr{G})p(\Sigma \mid \mathscr{G})p(\mathscr{G}), \end{aligned} \tag{7}$$

where $p(\Sigma,\mathscr{G} \mid \Theta,\mathbb{Z},\mathbb{X})$ reduces to $p(\Sigma,\mathscr{G} \mid \mathbb{Z})$ because Σ and $\mathscr{G}$ are independent of $\mathbb{X}$ and Θ given $\mathbb{Z}$, that is once the latent data are "observed". The latter expression corresponds to a joint posterior of $(\Sigma,\mathscr{G})$ in a Gaussian graphical model with a HIW (conjugate) prior; accordingly, direct sampling from $p(\Sigma,\mathscr{G} \mid \mathbb{Z})$ is possible following the MCMC scheme presented in [9].

The full conditional of $(\mathbb{Z},\Theta)$ can be instead written as

$$p(\mathbb{Z},\Theta \mid \Sigma,\mathscr{G},\mathbb{X}) = p(\mathbb{Z} \mid \Theta,\Sigma,\mathscr{G},\mathbb{X})p(\Theta \mid \Sigma,\mathscr{G},\mathbb{X}). \tag{8}$$

Specifically, we can write the first term as

$$p(\mathbb{Z} \mid \Theta,\Sigma,\mathscr{G},\mathbb{X}) \propto \prod_{i=1}^{n} d\mathscr{N}_q(\mathbf{z}^{(i)} \mid \mathbf{0},\Sigma)\mathbf{1}\left\{\mathbf{z}^{(i)} \in C(\mathbf{x}^{(i)},\Theta)\right\},$$

where $d\mathscr{N}_q(\cdot)$ denotes the density function of the multivariate Normal distribution $\mathscr{N}_q(\mathbf{0},\Sigma)$. Accordingly, we can sample each latent observation $\mathbf{z}^{(i)}$, $i = 1,\dots n$, independently from a suitable multivariate Normal distribution truncated at the region $C(\mathbf{x}^{(i)},\Theta)$. The second term can be written as

$$p(\Theta \mid \Sigma,\mathscr{G},\mathbb{X}) \propto \prod_{i=1}^{n} \left\{ \Phi_q(\theta^{(1)}_{x_{ij}},\dots,\theta^{(q)}_{x_{ij}} \mid \Sigma) - \Phi_q(\theta^{(1)}_{x_{ij}-1},\dots,\theta^{(q)}_{x_{ij}-1} \mid \Sigma) \right\} \cdot \prod_{j=1}^{q} p(\theta_0^{(j)}),$$

where $\Phi_q(\cdot)$ is the c.d.f. of $\mathscr{N}_q(\mathbf{0},\Sigma)$. For $j = 1,\dots,q$, we can update $\theta_0^{(j)}$ sequentially using a Metropolis-Hastings scheme based on the following steps:

- draw $\left(\theta_0^{(j)}\right)^*$ from a suitable proposal distribution $q\left(\left(\theta_0^{(j)}\right)^* \mid \theta_0^{(j)}\right)$, for instance $\mathscr{N}\left(\theta_0^{(j)},\sigma_0^2\right)$;
- given the current value of $\theta_0^{(j)}$ accept $(\theta_0^{(j)})^*$ with probability

$$\alpha_j = \min\left\{ 1\,;\ \frac{p\left(\left(\theta_0^{(j)}\right)^*,\ \theta_0^{-j} \mid \Sigma,\mathscr{G},\mathbb{X}\right)}{p\left(\theta_0^{(j)},\ \theta_0^{-j} \mid \Sigma,\mathscr{G},\mathbb{X}\right)} \cdot \frac{q\left(\theta_0^{(j)} \mid \left(\theta_0^{-j}\right)^*\right)}{q\left(\left(\theta_0^{-j}\right)^* \mid \theta_0^{(j)}\right)} \right\}$$

where $\theta_0^{-j} = \left\{\theta_0^{(k)}, k \neq j\right\}$ denotes the collection of all cut-offs excluding the j-th.

4 Conclusions

We have presented a Bayesian methodology for learning dependence relations among multivariate ordinal binary data. Our method assumes that the ordinal cate-

gorical data are generated by thresholding of latent Gaussian data whose sampling distribution is Markov w.r.t. to a graph. Main advantage of the proposed method is that, differently from alternative methods based on Multinomial-Dirichlet models, we can provide a correlation-type measure between each pair of ordinal variables, which is encoded in the covariance matrix of the latent variables, Σ.

An extension of the method to ordinal variables $X_1, \ldots, X_q$, each with an arbitrary number of levels is possible and requires a suitable adaptation of the proposed framework. In particular, assuming that $X_j \in \{0, 1, \ldots, K_j\}$, one could introduce for each variable X_j a collection of cut-offs $\theta_0^{(j)}, \ldots, \theta_{K_j-1}^{(j)}$ such that

$$X_j = \begin{cases} 0 & \text{if } -\infty < Z_j \leq \theta_0^{(j)} \\ 1 & \text{if } \theta_0^{(j)} < Z_j \leq \theta_1^{(j)} \\ \ldots & \\ K_j & \text{if } \theta_{K_j-1}^{(j)} < Z_j < +\infty \end{cases} \tag{9}$$

where Z_j is the latent variable associated with X_j, and priors to the new cut-off parameters can be assigned independently following Equation (6).

We remark that this extension may augment the computation cost of the proposed algorithm, especially in high-dimensional (large q) settings, and specifically in the update of the latent data which requires sampling from truncated multivariate Normal distributions. Efficient computational implementations are currently under investigation.

References

1. Castelletti, F., Peluso, S.: Equivalence class selection of categorical graphical models. Computational Statistics & Data Analysis **164**, 107304 (2021)
2. Cheng, J., Li, T., Levina, E., Zhu, J.: High-Dimensional Mixed Graphical Models. Journal of Computational and Graphical Statistics **26**, 367–378 (2017)
3. Dawid, A. P., Lauritzen, S. L.: Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. The Annals of Statistics **21**, 272–1317 (1993)
4. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**, 432–441 (2008)
5. Kalisch, M., Bühlmann, P.: Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. Journal of Machine Learning Research **8**, 613–636 (2007)
6. Lauritzen, S. L.: Graphical Models, Oxford University Press (1996)
7. Mohammadi, A., Wit, E. C.: Bayesian Structure Learning in Sparse Gaussian Graphical Models. Bayesian Analysis **10**, 109–138 (2015)
8. Scott. J. G., Berger, J. O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. The Annals of Statistics **38**, 2587–2619 (2010)
9. Wang, H., Li, S. Z.: Efficient Gaussian graphical model determination under G-Wishart prior distributions. Electronic Journal of Statistics **6**, 168–198 (2012)

Multinomial Thompson Sampling for adaptive experiments with rating scales

Multinomial Thompson Sampling per esperimenti adattivi con scale di rating

Nina Deliu

Abstract Bandit algorithms such as Thompson Sampling (TS) have been put forth for decades as useful for conducting adaptively-randomized experiments. By skewing the allocation ratio towards superior arms, they can substantially improve participants' welfare with respect to particular outcomes of interest. For example, as we illustrate in this work, they may use participants' ratings for understanding and assigning promising text messages for managing mental health issues more often. However, model-based algorithms such as TS, typically assume binary or normal models, which may lead to suboptimal performances in categorical rating scale outcomes. Guided by our field experiment, we extend the application of TS to rating scale data and show its improved performance in a number of synthetic experiments.
Abstract *Gli algoritmi di tipo multi-armed bandits, quali il Thompson Sampling (TS), rappresentano un metodo efficace per la conduzione di esperimenti adattivi. Adattando automaticamente il rapporto di allocazione verso il braccio superiore (quando questo esiste), essi possono migliorare sostanzialmente sia il welfare dei partecipanti–anche semplicemente rilevando le loro preferenze–che l'efficacia e il costo di uno studio. Tuttavia, algoritmi come il TS assumono generalemnte un modello binario o normale per la variabile di risposta, portando a risultati non ottimali in caso di dati con scale di rating. Motivati da un esperimento reale sulla salute mentale, in questo lavoro si propone un'estensione dell'algoritmo TS per scale di valutazione e si studiano la sue performances in una serie di esperimenti sintetici.*

Key words: Adaptive experiments, Thompson Sampling, Multi-armed bandits, Rating Scales, Multinomial Model, Dirichlet Distribution

---

Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF), Sapienza Università di Roma; e-mail: nina.deliu@uniroma1.it

1 Introduction

Adaptively-randomized experiments have the potential to enhance participants' welfare while collecting high-quality data, resulting in a more flexible, efficient, and ethical alternative compared to traditional fixed studies [1]. In such experiments, allocation ratios are skewed towards more efficient or informative arms, with the goal of assigning superior arms to as many participants as possible.

Multi-armed bandit (MAB) algorithms [2] have been argued for decades as useful to adaptively randomize experiments with the aim of optimizing an outcome of interest. In MAB problems, an *agent* chooses at each time t one of the available *arms* for which a *reward* is then provided. The goal is to maximize the cumulative reward, or equivalently minimize regret, under uncertainty on the best arm. As rewards are generally stochastic, and they might vary between arms, the agent must balance *exploring* each arm to gain information with *exploiting* the information gained so far by choosing arms with the higher expected reward. *Thompson Sampling* (TS) is a highly-interpretable randomized MAB strategy that allocates arms in proportion to their posterior mean reward. Due to its empirical and theoretical optimality [3], it has received renewed attention in recent years and have been successfully applied in a wide variety of domains, going from recommendation to education [4].

In many real-world applications, arms are a set of recommendations or explanations, that are allocated according to users' preferences. In this work, for example, we are motivated by mental health experiments we deployed on the Amazon Mechanical Turk [5], in which participants were asked to rate two types of messages in terms of how helpful they would be for managing their mood. Results of a primary experiment (MTurk I), based on a fixed design with an equal allocation ratio for gaining benchmark knowledge, are reported in **Fig. 1**.

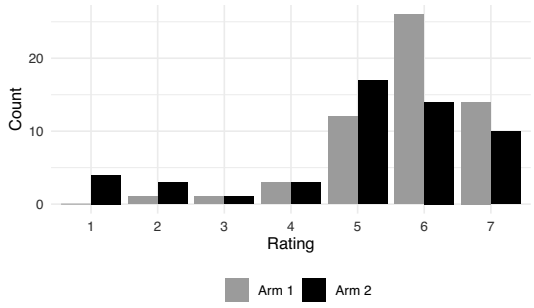


Fig. 1 Arms' ratings distribution in the two-armed *MTurk I* experiment, based on a 7-point Likert scale. A small superiority of Arm 1 vs Arm 2 is shown: sample means $\hat{\mu}_1 = 5.81$ vs $\hat{\mu}_2 = 5.08$, sample standard deviation $\hat{\sigma}_1 = 1.04$ vs $\hat{\sigma}_2 = 1.72$, for a sample size $N = 110$.

Subsequent experiments [6] have then been conducted adaptively with TS and a Gaussian model, in absence of more adequate modeling procedures for rating data within TS. Notably, despite the vast array of MAB and TS variants, most of them assume either binary or Gaussian rewards, and a shortage of MAB methodologies to handle rating scale data persists. The general practices in such cases consist in either dichotomizing the ordinal reward with a threshold (often arbitrary) rule and

then using a binary model, or using a normal model directly on the rating outcomes, as in our MTurk study. Such practices have been long recognized as suboptimal in terms of both reward efficiency [7] and statistical inference [8].

In this work, we extend the applicability of TS to rating scale data, introducing the *Multinomial TS*, which can be easily implemented with the *Dirichlet* conjugate family [9]. Then, guided by our motivating experiment, we evaluate its performance compared to the standard Gaussian TS in a number of synthetic experiments.

2 Problem Setting and Methods

Experimental set-up We consider a two-armed T-time horizon experiment, in which participants are accrued in a fully-sequential way, with a total experiment sample size $N = T$. At accrual, each participant $t = 1, \dots, T$ is assigned one of the two available arms, say $A_t \in \{1,2\}$, and an outcome or reward $Y_t(A_t)$ associated with that arm is subsequently observed before the next time $t+1$. Arms are drawn according to a policy $\boldsymbol{\pi_t} \doteq \{\pi_{t,k}, k = 1,2\}$, where $\pi_{t,k}$ is the allocation probability of arm k at time t. Given the history of selected arms and associated rewards $\mathscr{H}_t \doteq \{A_\tau, Y_\tau(A_\tau), \tau = 1, \dots, t\}$, the goal of an adaptive experiment may be to find an allocation policy so as to maximize the expected cumulative reward over T.

In line with MTurk I, we consider a 7-points rating scale outcome, i.e., $Y_t \in \{1,2,\dots,7\}$ for all t's, with higher values indicating better outcomes. We assume they are drawn independently from a fixed distribution that depends on the arm but not on t (*stochastic stationary bandits*; [2]), with the following conditional mean:

$$\mathbb{E}(Y_{t,i}(A_{t,i})|\mathscr{H}_{t-1}) = \mathbb{E}(Y_t|A_t) = \mu_1 \mathbb{I}(A_t = 1) + \mu_2 \mathbb{I}(A_t = 2), \tag{1}$$

with (β_1, β_2) the unknown arm parameters. In the next sections we will make an additional assumption on rewards distribution, maintaining the parameters unknown.

Thompson Sampling Rooted in a Bayesian framework, TS allocates arms in proportion to their posterior probability of being associated with the maximum expected reward at each round t. In a two-armed setting, denoting by π_{t1}^{TS} TS's allocation probability of arm 1 at step t, and considering Eq. (1), we have that:

$$\pi_{t1}^{\text{TS}} = \mathbb{P}\Big(\mathbb{E}\big(Y_t(A_t = 1)\big) \geq \mathbb{E}\big(Y_t(A_t = 0)\big)|\mathscr{H}_{t-1}\Big) = \mathbb{P}\Big(\mu_{t1} \geq \mu_{t2}|\mathscr{H}_{t-1}\Big), \quad \forall t.$$

The typical way for implementing TS, involves drawing at each round t a sample from the posterior distribution of each of the unknown arms' parameters in Eq. (1), say $\tilde{\mu}_{tk}$, with $k = 1,2$, and then selecting the arm associated with the highest posterior estimated mean reward $\mathbb{E}(\tilde{Y}_t|A_t = k) = \tilde{\mu}_{tk}$, i.e., $\tilde{a}_t \doteq \text{argmax}_{k=1,2}\mathbb{E}(\tilde{Y}_t|A_t = k) = \text{argmax}_{k=1,2}\tilde{\mu}_{tk}$. A conjugate family is generally assumed for the reward variable, with the most common ones being the Binomial and the Gaussian [3].

2.1 Our Proposal: Multinomial Thompson Sampling

The multinomial distribution is the extension of the binomial distribution for categorical outcomes with more than two response categories [10]. Given J mutually exclusive categories of an outcome Y, among which one and only one category is observed at each time t (i.e., $\sum_{j=1}^{J}\mathbb{I}(Y_t=j)=1$ and $\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbb{I}(Y_t=j)=T$), it models the probability of counts $X_{tj}=\sum_{\tau=1}^{t}\mathbb{I}(Y_t=j), \forall j\in[1,J], \forall t\in[1,T]$. At a given time t, the probability mass function of a multinomial, denoted by Multinom$(t;\boldsymbol{p})$, with $\boldsymbol{p}=(p_1,\dots,p_J)=\Big(\mathbb{P}(Y_t=1),\dots,\mathbb{P}(Y_t=J)\Big)$ the unknown parameters, is given by:

$$f(x_{t1},\dots,x_{tJ};t;p_1,\dots,p_J)=\left(\frac{t!}{\prod_{j=1}^{J}x_{tj}!}\right)\prod_{j=1}^{M}p_j^{x_{tj}}, \tag{2}$$

where $t=\sum_{j=1}^{J}x_{tj}$. For a multinomial distribution, we have that $\mathbb{E}(X_{tj})=tp_j$. In case of ordinal, real valued categories (e.g., $j=1,\dots,7$ as in our rating scales), we can compute the mean of the outcome Y_t at each t, as $\mu_t=\mathbb{E}(Y_t)=\sum_{j=1}^{J}jp_j$.

Eq. (2) can be also expressed using the gamma function Γ, directly showing its resemblance to the Dirichlet distribution, which is its conjugate prior. Given $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_J)$, with $\alpha_j>0, \forall j\in[1,J]$, this is denoted by Dir$(\boldsymbol{\alpha})$, and models our beliefs/knowledge on the unknown parameters as:

$$f(p_1,\dots,p_J;\alpha_1,\dots,\alpha_J)=\left(\frac{\prod_{j=1}^{J}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{J}\alpha_j)}\right)\prod_{j=1}^{M}p_j^{\alpha_j-1}, \tag{3}$$

where $\boldsymbol{p}$ belongs to the standard $J-1$ simplex ($\sum_{i=1}^{J}p_j=1$ and $p_j\geq 0, \forall j\in[1,J]$).

We then update our beliefs on the parameters based on the observed outcomes, by updating their posterior distribution as $\alpha_j\leftarrow\alpha_j+c_{tj}$, $\forall t$'s and $\forall j$'s, where $c_{tj}=\mathbb{I}(y_t=j)$ states whether category j was observed at time t or not (see Algorithm 1).

Algorithm 1 Multinomial TS pseudocode

Input: Time horizon T, prior parameters $\boldsymbol{\alpha_k}=(\alpha_{1k},\dots,\alpha_{Jk})$, with J the overall number of categories of the rating scale outcome, and $k=1,2$ the study arm.

for $t=1,2,\dots,T$ **do**
  for $k=1,2$ **do**
    Sample $\tilde{\boldsymbol{p}}_{\boldsymbol{k}}=(\tilde{p}_{1k},\dots,\tilde{p}_{Jk})\sim\text{Dir}(\alpha_{1k},\dots,\alpha_{Jk})$
    Compute $\tilde{\mu}_{tk}=\sum_{j=1}^{J}j\tilde{p}_{jk}$
  end for
  Select arm $\tilde{a}_t=\text{argmax}_{k=1,2}\tilde{\mu}_{tk}$ and observe the reward y_t.
  for $j=1,\dots,J$ **do**
    Compute $c_{tj}=\mathbb{I}(y_t=j)$
  end for
  for $k=1,2$ **do**
    Update posteriors: $(\alpha_{1k},\dots,\alpha_{Jk})\leftarrow(\alpha_{1k},\dots,\alpha_{Jk})+(c_{t1},\dots,c_{tJ})$
  end for
end for

3 Synthetic Experiments and Results

For our empirical evaluation, we focus on the set-up introduced in Section 2, with $T = 1000$, and simulation scenarios defined according to the MTurk I motivational study in **Fig. 1**. We evaluate the proposed Multinomial TS with equal Dirichlet priors for the two arms with parameters $\alpha_{jk} = 1$ for $j = 1, \ldots, 7$ and $k = 1, 2$, and compare it to a Gaussian TS with variances set according to the results of the MTurk I experiment and Normal $N(0, 10)$ priors. We also illustrate the behavior of an *Oracle* that always allocates the best arm–when one exists–or both of them with equal probability–when they are identical. We look at standard MAB performance measures.

Regret and Optimal Arm Allocation under a unique optimal arm The unique optimal arm scenario is simulated with $\mu_1 = 5.81 > \mu_2 = 5.08$ ($\sigma_1 = 1.04$, $\sigma_2 = 1.72$), and individual scale frequencies given by the MTurk sample estimates: $p_1 = (0.00, 0.02, 0.02, 0.05, 0.21, 0.45, 0.24)$ and $p_2 = (0.08, 0.06, 0.02, 0.06, 0.33, 0.27, 0.19)$. Results in **Fig. 2** show the increased performance of the proposed Multinomial TS over Normal TS.

Arm Allocation under identical arms In addition to the aforementioned scenario reflecting the MTurk experiment, we now evaluate the behavior of the proposed strategy in an identical arms scenario for understanding how it balances the allocation of one arm over the other one, when no one should be preferred. This scenario is simulated with $\mu_1 = \mu_2 = 4$ ($\sigma_1 = \sigma_2 = 2$) and symmetric scale frequencies: $p_1 = p_2 = (0.02, 0.09, 0.23, 0.31, 0.23, 0.09, 0.02)$. The latter resembles a Gaussian distribution for ensuring higher comparability between the two TS alternatives. Results in **Fig. 2** show that Multinomial TS is more balanced in arms' allocation compared to a Gaussian TS, which can be very extreme–allocating one arm almost the totality of the times–even when this is not superior.

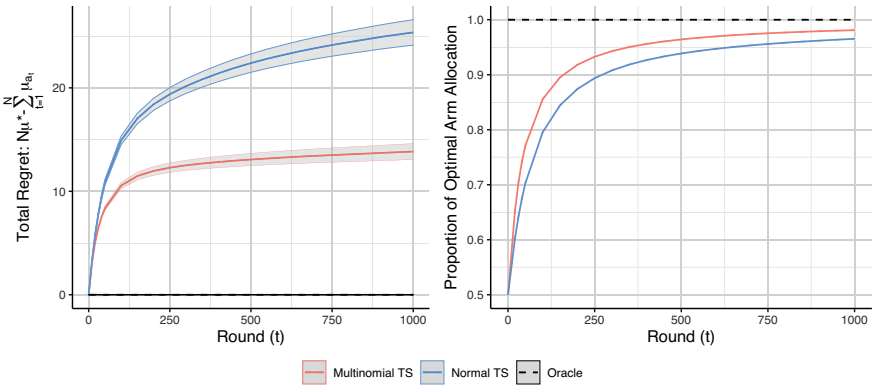


Fig. 2 Regret and proportion of optimal arm allocation in the proposed Multinomial TS vs a Gaussian TS. Values are obtained by averaging across 10^4 independent TS trajectories.

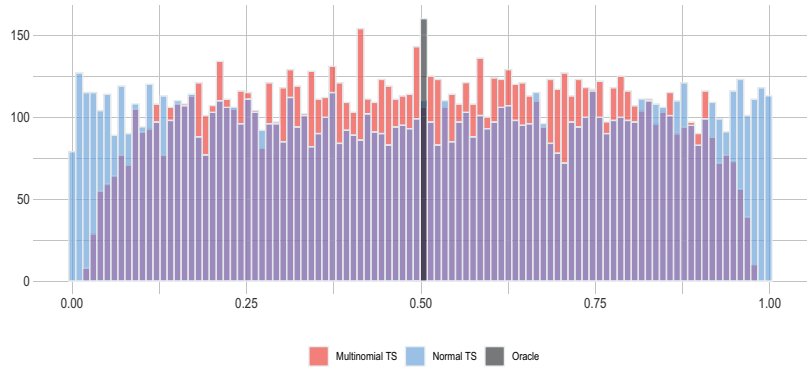


Fig. 3 Empirical allocation of one of the two identical arms (Arm 1). Values are obtained by averaging across 10^4 independent TS trajectories.

4 Conclusion

In this work, motivated by the MTurk field experiment, we extended the applicability of TS to rating scales data, introducing the Multinomial TS. We demonstrated that it can outperform the widely used TS with a Gaussian model in scenarios with a unique optimal arm, and that it is a more balanced solution–that can translate into inferential advantages, due to a lower imbalance in the allocation (see e.g., [6])–when arms are identical. Further work is required to understand how the proposed version would behave in a multi-armed setting or under non-stationarities.

References

1. Bothwell, L. E., Avorn, J., Khan, N. F., & Kesselheim, A. S. (2018). Adaptive design clinical trials: a review of the literature and ClinicalTrials. gov. *BMJ open*, 8(2), e018320.
2. Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.
3. Agrawal, S., & Goyal, N. (2013). Further optimal regret bounds for Thompson Sampling. *In Artificial Intelligence and Statistics* (pp. 99-107). PMLR.
4. Williams, J. J., Rafferty, A., Tingley, D., Ang, A., Lasecki, W. S., & Kim, J. (2018). Enhancing Online Problems Through Instructor-Centered Tools for Randomized Experiments. *CHI2018, 36th Annual ACM Conference on Human Factors in Computing Systems*.
5. Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1), 1-23.
6. Deliu, N., Williams, J. J., & Villar, S. S. (2021). Efficient Inference Without Trading-off Regret in Bandits: An Allocation Probability Test for Thompson Sampling. *arXiv:2111.00137*.
7. Williamson, S. F., & Villar, S. S. (2020). A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, 76(1), 197-209.
8. Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, 332(7549), 1080.
9. Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). *Continuous multivariate distributions, Volume 1: Models and applications* (Vol. 1). John Wiley & Sons.
10. Agresti, A. (2013). *Categorical data analysis. Third edition.* Wiley series in Probability and statistics. Wiley.

Ranking extraction in nested partially ordered data systems

Estrazione di ranking in sistemi annidati di dati parzialmente ordinati

Marco Fattore, Barbara Cavalletti, Matteo Corsi, Alessandro Avellone

Abstract We propose a novel procedure for extracting rankings from nested systems of partially ordered data, a kind of data structure often encountered in multi-indicator system analysis and multi-criteria decision making, e.g. when units or alternatives are pairwise compared or when preferences are expressed within different contexts, or in view of different goals, which are themselves partially ordered. We develop the procedure through a real example, pertaining to the prioritization of public interventions, to support small companies in the food sector.

Abstract *L'articolo presenta un nuova procedura per l'estrazione di ranking da sistemi annidati di insiemi parzialmente ordinati, una struttura dati tipica nell'analisi di sistemi multi-indicatore e in processi di decision-making e prioritizzazione, quando oggetti, alternative o preferenze vengono confrontati in diversi contesti, o rispetto a obiettivi differenti, a loro volta parzialmente ordinati. La procedure viene sviluppata attraverso un esempio reale, relativo alla prioritizzazione degli investimenti pubblici, per il supporto alle piccole imprese, nel settore agro-alimentare.*

Key words: Nested posets, Partially ordered set, Preferences, Prioritization, Ranking extraction

---

Marco Fattore
University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 - Milano e-mail: marco.fattore@unimib.it

Barbara Cavalletti
University of Genova, Via Balbi 5, 16126 Genova e-mail: barbara.cavalletti@unige.it

Matteo Corsi
University of Genova, Via Balbi 5, 16126 Genova e-mail: matteo.corsi@edu.unige.it

Alessandro Avellone
University of MIlano-Bicocca, Piazza dell'Ateneo Nuovo, 1 20126 - Milano e-mail: alessandro.avellone@unimib.it

1 Introduction

The issue of ranking extraction from partially ordered data occurs in a variety of fields and contexts, usually for prioritization and decision-making purposes [2]. Here, we address it considering a particular but typical setting, where partially ordered data are *nested* in a two-level hierarchy. More precisely, the data structure we consider is composed of a higher level *partially ordered set* (*poset* for short) π_s on k objects $s_1, \ldots, s_k$ (e.g., strategies or goals to be pursued with different priorities) and of k lower level posets $\pi_a^1, \ldots \pi_a^k$, *on the same objects* $a_1 \ldots, a_h$, each poset being associated to an element of π_s (e.g. possible actions, partially ordered by importance, to be taken in view of the strategies or the goals). Technically, the problem is to produce a ranking of $a_1 \ldots, a_h$, taking into account both the information provided by their different lower level orderings and that comprised in the higher level partial order of $s_1, \ldots, s_k$. We want to do this in a *purely ordinal way*, i.e. avoiding the use of numerical weights and of metric or aggregative procedures that, although often employed, would be inconsistent in an ordinal setting.

2 Real example: prioritization of public economic interventions

To make the issue more concrete, we consider the following real case. Liguria Region, in the North of Italy, launched a survey among small producers of typical regional food products, to identify the key strategic areas they deemed more relevant to improve, for their economic sustainability, among the following five: 1 - *Management&Accounting*, 2 - *Product Quality*, 3 - *Marketing*, 4 - *Distribution&Sales channels* and 5 - *Innovation&Technology*. Producers were also asked to compare the relevance of five possible public actions, for improving the strategic areas, namely: 1 - *Traininig*, 2 - *Creation of networks*, 3 - *Providing technical/ICT support*, 4 - *Implementing product certifications* and 5 - *Promoting exchange meetings*. Based on the expressed preferences, to each producer a priority poset of the strategic areas and five priority posets (one, for each area) of the actions have been associated, as exemplified in Figure 1. Given the priorities expressed, how can a ranking of actions be generated, for each producer? (Notice that here we do not consider how to synthesize the priorities expressed by all the producers, into a global ranking and just focus on the issue of getting single action rankings for single producers).

3 The ranking procedure

We now describe the ranking extraction procedure and, to ease the exposition, we do this by applying it to the economic intervention example, just introduced. The procedure is organized into three main steps: (i) turning the input data into a multi-indicator system (below called the “ranking MIS”), where the strategic areas play

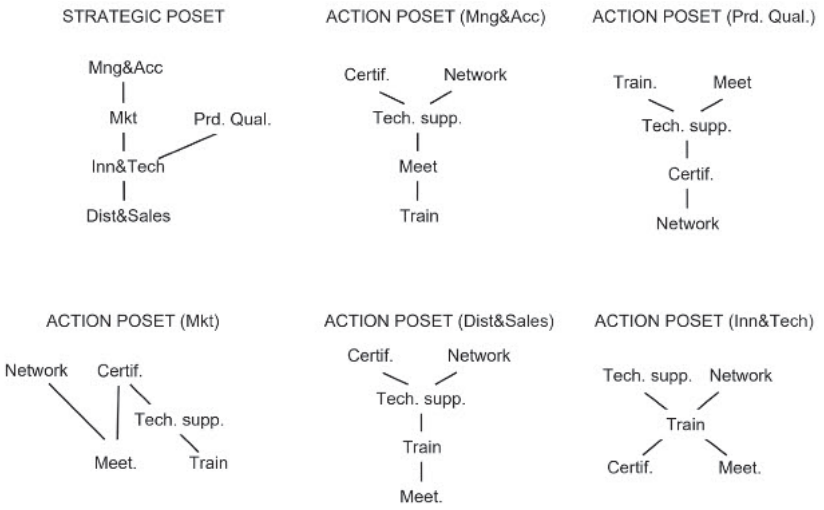


Fig. 1 Top left panel: Poset of strategic areas; other panels: posets of the actions for each strategic area. The diagrams must be read from top to bottom; if a downward edge path exists between two nodes, than the higher is "preferred" to the lower one; otherwise, the two nodes are *incomparable* and no priority between them is explicitly stated.

the role of "indicators" which order the actions as less or more relevant; (ii) turning the MIS into a poset of actions, incorporating the priorities expressed in the "strategic" poset and (iii) finally extracting a ranking out of it. The key feature of the procedure is that these steps are taken consistently with the ordinal nature of the input data. In particular, this holds true for the way the priorities on the strategic areas (i.e., the relevance of the "indicators" in the ranking MIS) are accounted for, without introducing any numerical weights, as shown below.

1. *Building the ranking MIS*. For each action poset in Figure 1, a ranking of the actions is extracted, by using the *SVD*-based algorithm described in [5], which ranks poset elements based on their so-called *mutual ranking probabilities* (*mrps* for short) and where the *mrp* p_{ij}^r $(r = 1,\ldots,k)$ of action j over action i in the action poset π_a^r is defined as

$$p_{ij}^r = \frac{|\ell \in \Omega(\pi_a^r) : \textit{action } j \textit{ is ranked higher than action } i \textit{ in } \ell|}{|\Omega(\pi_a^r)|} \qquad (1)$$

   denoting by ℓ a *linear extension* of π_a^r, by $\Omega(\pi_a^r)$ the set of all of its linear extensions and by $|A|$ (for any finite set A) the cardinality of A. Collecting the ranks of each of the five actions in each of the five rankings (one for each strategic area), the MIS reported in Table 1 is obtained.
2. *Turning the ranking MIS into a poset*. The MIS of Table 1 is turned into a poset by comparing action ranking profiles (i.e. MIS rows) component-wise, i.e. stating that action a_j is *more relevant* than action a_i (written $a_i \lhd a_j$) if and only if

Table 1 Multi-indicator system of the actions (rows), built from their rankings, relative to the five strategic areas (columns). The integer in entry ij is the position of action a_i, in the ranking relative to the strategic area s_j; lower scores mean greater relevance, i.e. higher positions in the ranking (notice that some actions are tied, in some rankings)

	Mng&Acc	Prd. Qual.	Mkt	Dist&Sales	Inn&Tech
Training	1	5	2	2	3
Networks	5	1	4	5	5
Tech. Support	3	3	3	3	5
Certifications	5	2	5	5	1
Meetings	2	5	1	1	1

its position in the five rankings is not lower and, in at least one, is higher than the corresponding position of a_i. We apply this condition to all of the $5^5 = 3125$ possible ranking profiles that can be built considering five indicators (the strategic areas) on five degrees (the possible ranking positions of five actions). This way, a big "component-wise" poset π_{cmp} is built, comprising 3125 elements, five of which are actually observed in the data. The reason why this larger poset is built is to provide a "context", where to set the evaluation of the relative importance of the actions, whose comparison outside of it would be much less informative. Poset π_{cmp}, however, treats strategic areas as equally important. To account for the priorities among them, as expressed by the strategic poset, we follow the *extension* principle introduced in [4] and generate a new ranking profile poset π_{ext}, by turning some incomparabilities of π_{cmp} into comparabilities, according to the following natural criterion. Let a_i and a_j be two actions, whose ranking profiles in the component-wise poset are incomparable due to conflicting scores on only two areas, e.g. $a_i = 23421$ and $a_j = 32421$ (so, the two actions are tied on the last three rankings). Suppose that, in the strategic poset, area s_2 is deemed more important than area s_1; then a_j, being placed higher on the second ranking (recall that lower scores correspond to higher positions), is deemed more relevant than a_i and the original incomparability between these profiles is resolved, by putting $a_i \lhd a_j$. By performing this operation for each pair of ranking profiles in π_{cmp}, many new comparabilities are added to it (and none is removed), producing the *extended* poset π_{ext}, which provides the new evaluation context incorporating the information comprised in the partial order of strategic areas. Notice that the extension criterion adopted here is weaker than that described in [4], which would add much more comparabilities to the component-wise poset. This choice aims at reducing as much as possible the degree of arbitrariness, somehow unavoidable in any "weighting" procedure.

3. *Extracting the final ranking*. To rank the observed actions, i.e. their corresponding ranking profiles, we compute their mutual ranking probabilities *within* π_{ext}, getting the matrix reported in Table 2. By computing the SVD of this matrix, as in Step 1 (see again [5]), each action is assigned a *dominance* score, which is then used to get the final ranking, which turns out to be as follows (in parenthesis, the scores):

$$\overset{(0.19)}{Meet.} \lhd \overset{(0.40)}{Train.} \lhd \overset{(0.45)}{Certif.} \lhd \overset{(0.53)}{Tech.\ supp.} \lhd \overset{(0.56)}{Networks}\ .$$

Being just interested in the ranking, the absolute value of the dominance scores is of less interest; however, it is noticeable that action *Meetings* has a very low score, compared to those of the other actions, showing that it is considered of far less impact. In any case, here we do not provide any interpretation to the results, the data being used only to describe and exemplify the procedure.

The whole ranking procedure is represented in Figure 2.

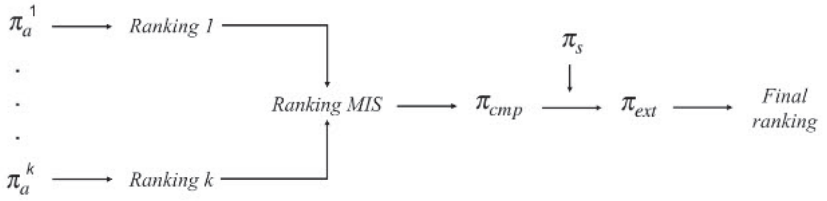


Fig. 2 The schematic flow of the ranking procedure.

Remark. The above procedure involves some heavy computations, for the construction of the extended poset π_{ext}. To lessen the computational burden, a random subset of 805 profiles, out of 3125 (comprising the observed ones), has been generated, getting a random subposet of π_{ext}, over which the mutual ranking probabilities have been computed. This introduces some probabilistic uncertainty in the results; but while this may affect the precise numerical expressions of the dominance scores, it does not alter the big picture and, in particular, the final ranking which is likely to be fairly robust against the sampling process.

Table 2 Mutual ranking probabilities of the actions, computed in π_{ext}

	Training	Networks	Tech. supp.	Certif.	Meet.
Training	1.00	0.91	0.86	0.68	0.00
Networks	0.09	1.00	0.38	0.17	0.00
Tech. support	0.14	0.62	1.00	0.26	0.00
Certifications	0.39	0.83	0.74	1.00	0.00
Meetings	1.00	1.00	1.00	1.00	1.00

4 Software implementation

While the ranking procedure is new, its implementation is essentially a composition of existing algorithms available in the `R` language, namely in packages `parsec` or

`POSetR` [1]. The first provides various functions to manipulate posets, while the second implements efficient algorithms for the computations of the mutual ranking probabilities, which are involved in the computation of the rankings, in the first and the last step of the procedure. The only function not available yet in the two mentioned packages is that providing the extension of the component-wise poset, which is however easy to implement with basic `R` functions.

5 Conclusion

In this paper, we have introduced a new purely ordinal procedure for generating rankings, from nested poset systems, a kind of data structure often encountered in multi-criteria decision-making and in synthetic indicator construction. The entire ranking process exploits just the ordinal information comprised in the data, without requiring exogenous weights and avoiding aggregative and compensative approaches, that would be inconsistent with the relational and non-numerical nature of the inputs. The procedure shows how complex systems of ordinal and partially ordered data can be consistently treated, by using the right mathematical setting, i.e. order theory [3], which provides a firm basis to this novel and growing area of multidimensional data analysis. The ranking procedure can be applied to more complex datasets, than that used here for exemplification purposes, although its natural application field remains that of the typical indicator systems used in socio-economics (or even in psychometry) and not that of the large datasets of big data analysis. As usual with combinatoric structures, the computational issues are not trivial. The development of new and more efficient algorithms to treat posets is, thus, a main research avenue, together with the theoretical deepening of the mathematical properties of the entire process and of its sensitivity to the unavoidable approximations involved in the computations.

References

1. Arcagni A., Avellone A., Fattore M. (2021) “POSetR: a new computationally efficient R package for partially ordered data”, *Book of Short Papers - SIS2021*, Pearson.
2. Bruggemann R., Patil G. P. (2011) *Ranking and Prioritization for Multi-indicator Systems*, Springer.
3. Davey B. A., Priestley B. H. (2002) *Introduction to Lattices and Order*, CUP.
4. Fattore M., Arcagni A. (2018) “A reduced posetic approach to the measurement of multidimensional ordinal deprivation”, *Social Indicators Research*, 136(3), 1053-1070.
5. Fattore M., Arcagni A (2020) “Ranking extraction in ordinal multi-indicator systems”, *Book of Short Papers - SIS 2020* - Pearson.

Towards the definition of distance measures in the preference-approval structures

Un nuovo approccio per definire misure di distanza nell'ambito dei preference-approval

Alessandro Albano, Mariangela Sciandra and Antonella Plaia

Abstract The task of combining preference rankings and approval voting is a relevant issue in social choice theory. The preference-approval voting (PAV) analyses the preferences of a group of individuals over a set of items. The main difference with the classical approaches for preference data consists in introducing, in addition to the ranking of candidates, a further distinction; candidates are subsetted in "acceptable" and "unacceptable", or also in "good set" and "bad set" (a way to express the approval/disapproval). This work introduces the definition of a new measure to quantify disagreement between preference-approval profiles. For each pair of alternatives, we consider the two possible disagreements in approvals and rankings and merge them through a function $h(\cdot)$ increasing in each component. We show that our approach allows to emphasize particularly those cases in which both ranking and approval show simultaneously the maximum discordance. .

Abstract *Il modello preference-approval (PAV) analizza le preferenze di un gruppo di invidui su un insieme di elementi. Questo modello può essere considerato una generalizzazione dei classici studi di preferenza, poichè oltre all'ordinamento dei candidati (rankings) viene fatta una distanzione tra candidati "accettabili" e candidati "non accettabili" (approval). In questo lavoro proponiamo una nuova distanza per misurare il grado di disaccordo tra i profili di preference-approval. Per ogni coppia di candidati, consideriamo congiuntamente le discordanze tra approval e rankings e le combiniamo attraverso una funzione $h(\cdot)$ crescente in ogni componente. Il nostro approccio, contrariamente alla letteratura, enfatizza in particolar-*

Alessandro Albano
Department of Economics, Business and Statistics - University of Palermo e-mail: alessandro.albano@unipa.it

Mariangela Sciandra
Department of Economics, Business and Statistics - University of Palermo e-mail: mariangela.sciandra@unipa.it

Antonella Plaia
Department of Economics, Business and Statistics - University of Palermo e-mail: antonella.plaia@unipa.it

mente i casi in cui occorre contemporanemente massima discordanza tra ranking e approval.

Key words: Preference rankings, Approval Voting, Preference-Approvals, Kemeny distance

1 Introduction

In social choice theory, preference rankings and approval voting are two popular ways to collect the preferences of a group of subjects on a set of elements. Preference rankings order the candidates from best to worst without distinguishing between acceptable and unacceptable. That is, if a is ranked before b, we can only infer that a is preferred over b, but we cannot infer anything about their absolute acceptability. In contrast, the approval voting system consists of drawing an ideal line separating the set of "good" objects from the set of "bad" objects without considering the ordering of the objects. Rankings and approvals are related, but they are basically different types of information and cannot be inferred from each other. However, in some situations, it is appropriate to gather both information simultaneously. For example, suppose a committee has to express their preferences over some candidates based on their scientific production. In this case, preference-approvals are helpful. Indeed, the committee outcome needs to include both candidates ranking and the division between "qualified" (good) and "unqualified" (bad).
Extensions to preference rankings mainly focus on the definition of weighted distances Plaia et al. 2018; Albano and Plaia 2021, while less effort has been devoted to the preference-approval settings. Brams and Sanver (2009) faced, for the first time, the issue of combining rankings and approval in a hybrid voting system called Preference-Approval Voting (PAV). In the PAV system, raters express candidate rankings and draw a line between good and bad candidates. In their study, the authors explain the various desirable properties that PAV meets and analyze the candidates who can and cannot win with this system. The problem of measuring agreement between preference-approval profiles is an open issue in the literature, distance-based approaches Erdamar et al. 2014; Dong et al. 2021 have gained increasing attention over the last years. The state-of-the-art methods compute the ranking distance and the approval distance marginally and eventually aggregate them by their linear combination. In this paper, we propose a new approach to define the distance between preference-approval profiles, simultaneously taking into account the disagreement in ranking and approval for each pair of candidates. Finally, our proposal is compared with approaches in the literature to highlight the differences.

2 Notation

Suppose n voters are asked to order m different alternatives, the ranking π is a mapping function from the set of items $X = \{x_1, \ldots, x_m\}$ to the set of ranks $\pi = (\pi(x_1), \pi(x_2), \ldots, \pi(x_m))$, where $\pi(x_i)$ is the rank given by a judge to item x_i. If the m items are ranked in m different ranks, a complete (full) ranking or linear ordering is achieved (Cook, 2006). In certain cases, some items could receive the same preference, then a tied ranking or a weak ordering is obtained.
In the framework of PAV modelling, each preference ranking, π, is paired with an approval vector, A. For any given set X of candidates, we define approvals by partitioning X into G, the set of good alternatives, and $U = X \backslash G$ the set of unacceptable alternatives, where G and U can be empty sets.
We represent a voter's preference-approval profile by a top-down order of candidates with a horizontal bar: candidates above the bar are approved, those below rejected.

$$\begin{array}{c} x_1 \\ \underline{x_2} \\ x_3 \\ x_4. \end{array}$$

The previous representation indicates that the voter's two top-ranked candidates, x_1 and x_2, are approved, and the voter's two bottom-ranked candidates, x_3 and x_4, are disapproved. The preference-approval profile is codified as follows:

$$\pi_1 = (1,2,3,4) \quad A_1 = (1,1,0,0).$$

3 Marginal distance-based approach

Measuring agreement between preference-approval profiles has been so far solved by a linear combination of a ranking distance and an approval distance calculated marginally. Erdamar et al. (2014) proposed to measure disagreement between preferences using the Kemeny distance (Kemeny, 1959), d_k, and to measure approval disagreement using the Hamming distance, d_H, (Hamming, 1950). The two distances are normalized into [0,1] (d_R, d_A), and eventually aggregated in a final preference-approval distance d_λ:

$$d_\lambda = \gamma\, d_R(\pi, \pi^*) + (1-\gamma) d_A(A, A^*), \tag{1}$$

where π, π^* are two m-size rankings, A, A^* are the two m-size approval vectors, and $\gamma \in [0,1]$ is a parameter used to control the relative relevance of the two components. Similarly, Dong et al. (2021) defined a new distance d_{PAS} using an L_1 metric both to compare preferences and approvals:

$$d_{PAS} = \sum_{i=1}^{m} \gamma \left|\pi(x_i) - \pi^*(x_i)\right| + (1-\gamma)\left|A_1(x_i) - A_2^*(x_i)\right|. \tag{2}$$

The two measures share the same "marginal" approach that consists of deriving the final distance through the additive effect of the sub-components involved. In other words, d_λ, d_{PAS} perfectly depend on the values of d_K, d_H, L_1 computed marginally.

4 Pairwise approach: d_{AP}

This section illustrates our proposal, d_{AP}. Given two preference-approval profiles (π, A) and (π^*, A^*) on m alternatives, the problem of measuring the disagreement between them is broken down into $m(m-1)/2$ easier sub-problems, i.e. one for each pair of items. Thus, the total distance d_{AP} is obtained by multiplying the two possible disagreements, in approvals and in preferences, for each pair of items $i, j \ \forall i > j \in [1, \ldots, m]$.
The ranking disagreement, for each pair of items, is computed by creating a *score matrix* of a ranking. A rank vector π with m objects can be transformed into a symmetric $m \times m$ score matrix, whose elements a_{ij} are defined by:

$$a_{ij} = \begin{cases} 1 & \text{if } i \text{ is preferred to } j \\ 0 & \text{if } i = j \text{ or } i \text{ is tied with } j \\ -1 & \text{if } j \text{ is preferred to } i. \end{cases} \tag{3}$$

The disagreement in approval, for each pair of items, is measured using an L_1 metric between the 2-dimensional approval vectors.

$$p_{ij} = |A_i - A_i^*| + |A_j - A_j^*|. \tag{4}$$

The formula of the aggregated distance d_{AP} between two m-size preference approval profiles $(\pi, A), (\pi^*, A^*)$ with score matrix a_{ij}, and b_{ij} respectively (defined as in Eq.3), and p_{ij} (defines ad in Eq.4):

$$d_{AP} = \sum_{i<j} h(\left|a_{ij} - b_{ij}\right|, p_{ij}), \tag{5}$$

where $h(\cdot)$, increasing in each component, merges both disagreements.

$$h(x,y) = \begin{cases} 0 & \text{if } x = 0, y = 0 \\ x \cdot \gamma & \text{if } x \neq 0, y = 0 \\ y \cdot \gamma & \text{if } x = 0, y \neq 0 \\ x \cdot y & \text{if } x \neq 0, y \neq 0, \end{cases} \tag{6}$$

where $0 < \gamma < 1$. It can be proved that $d_{AP} \in [0, 2m(m-1)]$.

4.1 Illustrative example

This section clarifies the practical impact of the pairwise distance, d_{AP}, and highlights the differences with the marginal approaches d_λ, d_{PAS}. Let us consider four preference-approval profiles as defined in table 1.

(π_1, A_1)	(π_2, A_2)
$x_1\ x_2\ x_3$	x_3
	$x_1\ x_2$

(π_3, A_3)	(π_4, A_4)
x_1	x_2
x_2	x_1
x_3	x_3

Table 1: Preference-approval profiles

The preference-approvals are codified as follows:

$$I = [\pi_1 = (1,1,1),\ A_1 = (1,1,1)]$$

$$II = [\pi_2 = (2,2,1),\ A_2 = (0,0,1]$$

$$III = [\pi_3 = (1,2,3),\ A_3 = (1,0,0)]$$

$$IV = [\pi_4 = (2,1,3),\ A_4 = (0,1,0)].$$

Table 2 compares the distances between the preference-approval profiles *I-II* and the profiles *III-IV*, assuming $\gamma = 0.5$.

Preference-	*Rank-dist*		*Approval-dist*		*Aggregated-dist*		
Approval	d_K	L_1	L_1	d_H	d_λ	d_{PAS}	d_{AP}
I-II	2	2	2	2	0.5	2	**3**
III-IV	2	2	2	2	0.5	2	**5**

Table 2: Preference-approval distances, $\gamma = 0.5$

It should be noted that:

- $d_\lambda(I,II) = d_\lambda(III,IV)$
- $d_{PAS}(I,II) = d_{PAS}(III,IV)$
- $d_{AP}(I,II) < d_{AP}(III,IV)$.

This is due to d_{AP} being sensitive to *double disagreements*. A double disagreement occurs when two profiles agree neither on the ranking nor on the approval of two items. In order to get a clearer view of the d_{AP} distance, let us break down the calculation into its sub-components.

$$d_{AP}(I,II) = d_{AP}[(\pi_1,A_1),(\pi_2,A_2)] = \sum_{i<j} h(\left|a_{ij}-b_{ij}\right|, p_{ij}) =$$
$$= h(0,2)+h(1,1)+h(1,1) = 1+1+1 = 3$$

$$d_{AP}(III,IV) = d_{AP}[(\pi_3,A_3),(\pi_4,A_4)] = \sum_{i<j} h(\left|a_{ij}-b_{ij}\right|, p_{ij}) =$$
$$= \mathbf{h(2,2)}+h(0,1)+h(0,1) = \mathbf{4}+0.5+0.5 = 5$$

The profile pair $(III\text{-}IV)$ exhibits a double disagreement on x_1 and x_2, **h(2,2)**, that is, there is maximum discordance in both components simultaneously. In other words, the presence of a double disagreement heavily amplifies the final d_{AP} distance.
Note that in the case of completely disagreement between two subjects, the three distances d_λ, d_{PAS}, d_{AP} reach their maximum value:

$$\begin{array}{cc} x_1 & \overline{x_4} \\ x_2 & x_3 \\ x_3 & x_2 \\ \underline{x_4} & x_1. \end{array}$$

5 Conclusions

This work introduces a new approach for computing the distance between preference-approval profiles. We consider the two possible disagreements in approvals and preferences for each pair of alternatives and merge them through a function $h(\cdot)$ increasing in each component. The aggregated distance, d_{AP}, differs from the distances in the literature in his sensitivity to double disagreements. In fact, our approach strongly penalizes the cases of maximum discordance in both components simultaneously. Therefore, while the state-of-the-art approaches only depend on the marginal distances between rankings and approvals, our proposal also depends on how many double disagreements occur. Future works should be devoted to the definition of a *consensus ranking-approval* defined as the closest preference-approval profile (i.e. with the minimum distance or maximum correlation) to the whole set of profiles.

References

Albano, A. and Plaia, A. (2021). Element weighted kemeny distance for ranking data. *Electronic Journal of Applied Statistical Analysis*, 14(1):117–145.

Brams, S. J. and Sanver, M. R. (2009). Voting systems that combine approval and preference. In *The mathematics of preference, choice and order*, pages 215–237. Springer.

Cook, W. D. (2006). Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of operational research*, 172(2):369–385.

Dong, Y., Li, Y., He, Y., and Chen, X. (2021). Preference–approval structures in group decision making: Axiomatic distance and aggregation. *Decision Analysis*.

Erdamar, B., García-Lapresta, J. L., Pérez-Román, D., and Sanver, M. R. (2014). Measuring consensus in a preference-approval context. *Information Fusion*, 17:14–21.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, 88(4):577–591.

Plaia, A., Sciandra, M., and Buscemi, S. (2018). Consensus measures for preference rankings with ties: an approach based on position weighted kemeny distance. *Advances in Statistical Modelling of Ordinal Data*, page 171.

Covid-19 Assessment and Evaluation 1

Covid-19 impact assessment and inequality decomposition methods

La valutazione dell'impatto del COVID-19 e i metodi di scomposizione della disuguaglianza

Federico Attili and Michele Costa

Abstract The first aggregate assessment of the impact of COVID-19 is usually made on the basis of a comparison of averages. Our goal is to combine the comparison between the averages with inequality decomposition methods. In particular, we refer to three measures of inequality between subgroups that allow an effective assessment of the inequality factors. Our analysis compares 2019 and 2020 IT-SILC consumption micro data in order to assess the impact of COVID-19 with respect to main inequality factors such as age, gender and geography.

Abstract *La prima valutazione aggregata dell'impatto del COVID-19 viene usualmente effettuata sulla base di un confronto tra medie. Il nostro obiettivo è quello di affiancare a tale confronto i metodi per la scomposizione della disuguaglianza. In particolare, facciamo ricorso a tre misure della disuguaglianza tra i gruppi che consentono una valutazione efficace dei fattori di disuguaglianza. Analizziamo i micro dati IT-SILC sui consumi per il 2019 e il 2020 per valutare l'impatto del COVID-19 rispetto a fattori chiave quali età, genere e ripartizione territoriale.*

Key words: Inequality decomposition, Gini index, Covid-19 impact

1 Introduction

The severity of the COVID-19 pandemic represented a very strong shock with serious health, psychological, social and economic consequences. The need to evaluate its effects on inequality has immediately been recognized. It led to a thriving literature (see e.g. [7] for a review), to which we wish to contribute by including the methods related to inequality decomposition in the methodological toolbox of researchers and policy makers.

Federico Attili, Michele Costa
Department of Economics, University of Bologna, e-mail: federico.attili2@unibo.it; michele.costa@unibo.it

COVID-19 impacted Italian households consumption in 2020 through different channels. The principal ones are income reduction [4], lockdown, restrictions, behavioural changes ([5], [6]). Except for lockdown, which flattened household consumption to unprecedented levels in all the society strata, the other channels shaped the distribution of consumption (and income) with heterogeneous intensity along its deciles and subgroups. To fully understand the heterogeneous changes that have taken place, it is therefore important to measure the inequality between subgroups at all points of the distribution.

We investigate household consumption during summer 2019 and 2020. From July to September 2020 the government removed the shutdown of most activities and many of the restrictions on social distancing and travels. By focusing on summer consumption, we improve the identification of the (potentially structural) effects of income losses and behavioural changes from 2019 and 2020.

With respect to three different partitions, which define subgroups based on age, gender and geographical factor, we compare how different can be the picture of household consumption inequality when considering decompositions that follow different paradigms. We show that the increase of age and gender gaps is much smaller if measured by comparing the whole subgroup distributions; the impact of COVID-19 on household consumption is instead homogeneous among the Italian regions and the different points of their distributions, therefore decompositions of inequality agree on modest reduction of geographical divide in consumption.

2 Inequality factors and inequality decomposition

Traditionally, the most direct and immediate way to measure the relevance of an inequality factor and the gap between subgroups j and h, refers to the (relative) difference between the subgroups means, such as $\Delta\bar{y}_{jh} = (\bar{y}_j - \bar{y}_h)/\bar{y}_h$ which provides an aggregate indicator. A more detailed framework can be developed on the basis of the comparisons between n quantiles of subgroup j (y_{ji}, $i = 1,\ldots,n$) and the corresponding quantiles of subgroup h (y_{hi}) that is $\Delta y_{jhi} = (y_{ji} - y_{hi})/y_{hi}$ with their synthesis given by $d_{jh} = \sum_{i=1}^{n} |\Delta y_{jhi}|/n$. The extension to the case of more than two subgroups involves a series of binary comparisons and their synthesis.

In this paper, we want to combine the differences between the means with the methods for the inequality decomposition, where the inequality between plays a key role in evaluating the importance of the underlying inequality factors and the gap between the subgroups. We focus on the Gini index G and on three of its decompositions. Our first reference [2] is the decomposition proposed by Bhattacharia and Mahalanobis, where the inequality between k subgroups is evaluated on the basis of the subgroup means as

$$G_b = \sum_{j=1}^{k} \sum_{h=1}^{k} p_j p_h |\bar{y}_j - \bar{y}_h|/2\bar{y} \qquad j \neq h. \tag{1}$$

where p_j is the share of population in subgroup j. A further relevant contribution to the Gini index decomposition literature [8] is proposed by Yitzhaky and Lerman, who suggest to evaluate the inequality between subgroups as

$$G_b = 2cov(\bar{y}_j, \bar{F}_j)/\bar{y} \tag{2}$$

where F indicates the cumulative distribution.

Alongside these two well-known decompositions of the Gini index, our third reference [1] is represented by a recent and innovative proposal by Attili, which, not being as well known as the previous ones, deserves some introductory notes. The main intuition of Attili's decomposition is to measure the between component by comparing the income of individuals belonging to different subgroups instead of comparing subgroup means. We present the paradigm of the decomposition by a simple scheme involving four units divided in two groups (Fig. 1). The Gini index of the four values normalises the sum of all the absolute differences indicated by the arrows. The vertical and horizontal arrows refer to the inequality within and between, respectively; the decomposition is finalised disentangling contributions to within and between inequality from the diagonal arrows. E.g., for the diagonal difference $(8-2)=6$, we split it in $(8-3)+(3-2)=6$, attributing the first (vertical) difference to inequality within and the second (horizontal) difference to inequality between. In the case of grey diagonals, the two differences are rescaled so that their sum equates the diagonal difference. This is achieved by a rescaling factor, calculated dividing the diagonal difference by the sum of the vertical and the horizontal one. Likewise, black arrows are rescaled by 1.

As in the example, the decomposition of G in within and between components singles out the sum of vertical from the sum of horizontal differences, where each difference is weighted by the sum of the related rescaling factors. Generalising to the case of n quantiles instead of two units for each subroup, the between component reads

$$G_b = \sum_{j=1}^{k}\sum_{h=1}^{k}\sum_{i=1}^{n} \left|y_{ji} - y_{hi}\right| w_{jhi}/(2\bar{y}n^2k^2) \tag{3}$$

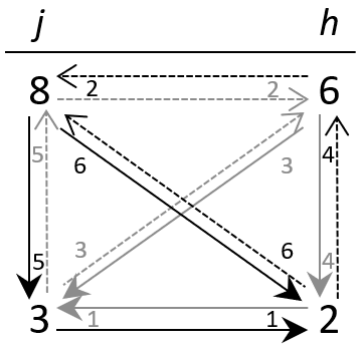


Fig. 1 Illustration of the Gini index in the case of four units divided in two groups. Arrows represent absolute pairwise differences between the four values.

It compares each quantile of subgroup j (y_{ji}) with the corresponding quantile of the other subgroups (y_{hi}) and the weights w_{jhi} ensure correspondence with the Gini index in a fully informative within-between decomposition scheme. A comprehensive presentation is given in Attili. Here we just point out that this measure of G_b is strongly correlated with the *distance between income distributions* proposed by [3], which observes a rigorous axiomatic approach.

3 Data and results

Our analysis is based on the EU-SILC survey for Italian households in 2019 and 2020. We focus on surveys carried out in July, August and September (4650 and 6964 households in 2019 and in 2020, respectively). Household consumption is equivalised for the number of members by using the square root equivalence scale.

In Table 1 we investigate three different bi-partitions based on age, gender and geographical factor. The first panel compares Young-(< 65) and Old-headed(≥ 65) households. In 2019 the former are richer than the latter but in the extreme deciles. Consumption of all groups and deciles decreases in 2020, but COVID-19 more evidently reduces rich old-headed households consumption. Higher values notoriously have impact on the means, so the difference between subgroup means ($\Delta\bar{y}_{jh}$) increases by 71.1% from 2019 to 2020. However, some of the deciles get closer in 2020 (see columns (C)-(D)), which is captured by much smaller increase of d_{jh} (+3.6%). The second panel explores gender divide in households consumption. In 2019 Male-headed households are richer than Female-headed households; again, extreme deciles are an exception. In the same deciles gender divide increases a lot in 2020. As before, $\Delta\bar{y}_{jh}$ is more sensible (+155.8) than d_{jh} (+99.7) to fluctuations of richer households. As for territorial divide, which is investigated in the third panel, the impact of COVID-19 on household consumption is stronger for households in the Centre and the North, but it is distributed quite homogeneously across deciles. Therefore, the reduction of the distance between subgroups is measured at around 7% by both $\Delta\bar{y}_{jh}$ and d_{jh}.

Table 2 reports Gini index decompositions for the three bi-partitions. Consistently with previous results, the increase of age divide is above 70% using eq. (1)-(2), but is more than ten times lower using eq. (3). A similar pattern involves gender divide, where the relevance of the inequality factor increases by 153.2% and 313.2% using eq. (1)-(2), and just by 57.4% with eq. (3). Territorial divide is the most relevant inequality factor and causes from 7.7% to 20.8% of total inequality in 2019. The three decompositions substantially agree on small reduction of the between component share (column (E)), thus suggesting a slightly smaller territorial divide.

To sum up, when the means are influenced by variations in top deciles, and when deciles vary heterogeneously, the change of the difference between subgroup means does not fully account for the evolution of the distance between the subgroup distributions. The latter is better captured by the average change of the differences

Table 1 Italian households monthly consumption. Deciles and means of the subgroups defined by age, gender and geographical factors. Columns (C) and (D) compare different groups in 2019 and in 2020, while column (E) evaluates the 2019-2020 dynamic.

	(A)		(B)		(C)	(D)	(E)
	Young		Old		(A-B)/(B)		(D-C)/(C)
	2019	2020	2019	2020	2019	2020	
d_1	591	584	592	572	-0.2%	2.2%	
d_2	879	853	844	823	4.2%	3.7%	
d_5	1470	1404	1322	1293	11.2%	8.6%	
d_9	2700	2639	2548	2455	6.0%	7.5%	
d_{10}	4035	3953	4273	3723	-5.6%	6.2%	
d_{jh}					7.3%	7.5%	3.6%
mean	1784	1731	1705	1604	4.6%	7.9%	71.1%
	Man		Woman				
	2019	2020	2019	2020	2019	2020	
d_1	590	608	593	533	-0.5%	14.2%	
d_2	873	867	852	789	2.5%	10.0%	
d_5	1438	1386	1368	1310	5.1%	5.8%	
d_9	2688	2635	2581	2469	4.2%	6.7%	
d_{10}	4088	3948	4171	3721	-2.0%	6.1%	
d_{jh}					3.8%	7.6%	99.7%
mean	1772	1722	1727	1614	2.6%	6.7%	155.8%
	Centre and North		South and Islands				
	2019	2020	2019	2020	2019	2020	
d_1	676	663	501	489	34.9%	35.5%	
d_2	981	947	716	695	37.1%	36.4%	
d_5	1585	1499	1122	1083	41.2%	38.3%	
d_9	2868	2779	2083	2061	37.7%	34.9%	
d_{10}	4397	4117	3253	3101	35.2%	32.7%	
d_{jh}					38.6%	36.0%	-6.6%
mean	1392	1356	1926	1838	38.4%	35.5%	-7.4%

between quantiles (d_{jh}). When Attili's between component evolution differs from the others, it does so in the proportion of d_{jh} to $\Delta\bar{y}_{jh}$.

4 Conclusions

We analyze the COVID-19 impact on inequality with respect to three factors: age, gender and geography. Age and gender gaps increase, while territorial divide (which is confirmed as the most important inequality factor in Italy) slightly decreases.

The assessment of these dynamics strongly depends on the employed measure. If it is based on averages, the evaluation reflects an aggregate effect, which may be at-

Table 2 Evolution of the between component from 2019 to 2020 for three partitions by age, gender and geographical factor. A, BM, YL stand for the three Gini index decompositions compared here.

		(A)		(B)		(C)	(D)	(E)
		$G \cdot 100$		$G_b \cdot 100$		G_b/G		(D-C)/(C)
		2019	2020	2019	2020	2019	2020	
Age	A	30.2	29.7	1.3	1.4	4.4%	4.7%	7.0%
	BM	30.2	29.7	1.0	1.7	3.4%	5.8%	72.7%
	YL	30.2	29.7	0.1	0.1	0.3%	0.5%	78.4%
Gender	A	30.2	29.7	0.8	1.2	2.6%	4.2%	57.4%
	BM	30.2	29.7	0.6	1.4	1.9%	4.9%	153.2%
	YL	30.2	29.7	0.0	0.1	0.1%	0.3%	313.2%
Geography	A	30.2	29.7	4.5	4.2	14.8%	14.2%	-3.9%
	BM	30.2	29.7	6.6	6.2	21.8%	20.8%	-4.5%
	YL	30.2	29.7	2.3	2.0	7.7%	6.9%	-10.2%

tributable to some aspects of the distribution only. This is the case of age and gender gaps in Italian household consumption: the average change is driven by the group of households consuming more. When, on the other hand, the measure of variation refers to the whole distribution, we have a more complete representation of the intervening dynamics. Inequality decomposition methods include both mean-based measures, such as [2] and [8], and measures referring to the whole distribution, such as [1], and represent an excellent choice for measuring the effects on inequality of phenomena such as COVID-19.

References

1. Attili, F.: Within-between decomposition of the Gini index: a novel proposal. University of Bologna, Department of Economics working paper n.1167, 1–33 (2021)
2. Bhattacharya, B. and Mahalanobis, B.: Regional disparities in household consumption in India. Journal of the American Statistical Association, **62**, 143–161 (1967)
3. Ebert, U.: Measures of distance between income distributions. Journal of Economic Theory, **32**, 266–274 (1984)
4. Gallo, G., Raitano, M.: SOS incomes: Simulated effects of COVID-19 and emergency benefits on individual and household income distribution in Italy. ECINEQ, Society for the Study of Economic Inequality, **566**, 1–36 (2020)
5. Hasan, S., Islam, M.A., Bodrud-Doza, M.: Crisis perception and consumption pattern during COVID-19: do demographic factors make differences? Heliyon, **7**, 1–11 (2021)
6. Marques Santos, A., Madrid, C., Haegeman, K., Rainoldi, A. . Behavioural changes in tourism in times of Covid-19. Luxembourg: European Commission, **22**, 121-147 (2020)
7. Stantcheva, S.: Inequalities in the times of a pandemic. DP 16856, CEPR, London (2022)
8. Yitzhaki, S., Lerman, R.: Income stratification and income inequality. Review of Income and Wealth, **37**, 313–329 (1991)

Multiversal methods for model selection: COVID-19 vaccine coverage and relative risk reduction

Metodi multiversali per la selezione dei modelli: la copertura vaccinale da COVID-19 e la riduzione relativa del rischio

Venera Tomaselli, Giulio Giacomo Cantone

Abstract Multiversal methods are used for assessment of misspecification risk in the scientific research. In this study, the Relative Risk Reduction after full COVID-19 vaccination is estimated in populations subjected to lockdown measures. While the multiverse model suggests a positive estimate of the effect of vaccination on risk reduction, apparently trivial alternative analyses in model specification reduce the estimate. Absurdly, some model specifications support a significant negative estimate. This is regarded as a case of Janus effect. Causes of Janus effect are identified in adoption of both Oxford Stringency Index as control variable and Negative Binomial instead of Quasi Poisson model. In some literature, these alternatives are characterised as equivalent. The multiversal methodology is further proposed to enhance countermeasures of Janus effect and risk of misspecification.

Abstract *I metodi multiversali sono utilizzati per l'accertamento del rischio di cattiva specificazione nella ricerca scientifica. In questo studio, la riduzione del rischio relativo a seguito del ciclo completo di vaccinazione anti COVID-19 è stimata in popolazioni soggette a misure di lockdown. Il modello multiversale indica una stima positiva dell'effetto della vaccinazione sulla riduzione del rischio, ma specificazioni del modello alternative, apparentemente ininfluenti, ne riducono la stima. Per assurdo, alcune specificazioni del modello indicano una stima negativa significativa. Questo è considerato un caso di effetto Giano, le cui cause sono identificate nell'adozione dell'Oxford Stringency Index come variabile di controllo congiuntamente all'adozione del modello Binomiale Negativo, invece del modello Quasi Poisson. In letteratura, queste alternative sono talvolta considerate come equivalenti. La metodologia multiversale è infine proposta per definire misure di contrasto rispetto all'effetto Giano ed al rischio di cattiva specificazione.*

Venera Tomaselli
Department of Political and Social Sciences, University of Catania, 8, Vittorio Emanuele II, 95131 Catania, e-mail: venera.tomaselli@unict.it, mobile: 39-3478056127, *(corresponding author)*

Giulio Giacomo Cantone
Department of Physics and Astronomy "E. Majorana", University of Catania, 64, S. Sofia, 95123 Catania, e-mail: giulio.cantone@phd.unict.it

Key words: multiverse methods, specification curve analysis, relative risk reduction, COVID-19 vaccine coverage.

1 Introduction

Multiversal methods are statistical and graphical tools for assessment of risk of misspecification in a study or in a meta-analysis. They share relevant features:

- an intellectual debt to original methodologies for sensitivity analysis [9] and their computational intensive developments;
- authors invested in improving methodological robustness of scientific studies, mostly as a reaction to the crisis of credibility of null hypothesis testing in experimental social science and clinical medicine [8, 17];
- instead of rejecting null hypothesis in hypothesis testing, lowering α, or limiting analytical freedom of the researcher [15], multiversal methods enhance structural variability in analytical choice of the researcher, the so-called *degrees of freedom of the research* in [5], to reduce arbitrariness of findings only to robust results.

All multiversal methods generate a 'multiverse' of estimates $\hat{\beta}$ in a relation between predictor and outcome ($x \rightarrow y$). Each estimate is associated to a p-value.

The methods differ in the way the multiverse is generated:

- Vibration models [12] are focused on model specification through identification of an extensive set of control variables Z, of which the presence or absence is considered to have equivalent weight in the model selection strategy. This approach has been criticised for being prone to fall into M-collider bias [3].
- Multiverse analysis [14] and Specification Curve analysis [13] are focused on alternating equivalent operationalizations of the empirical measures (e.g., using Mercalli instead of Richter scale) or equivalent regression models (e.g., logit *vs* probit).
- Young and Holsteen [18] use a multiverse for estimation of model variance, that is the variability in estimation across different modeling of the underlying problem. Model variance is positively correlated to the risk of elicitating false-positive associations of variables.

In the present study, a multiverse has been generated to model the effectiveness of relative risk reduction ($\hat{RRR}$) in people full vaccinated against COVID-19 in 2021, that is, the inferred reduction of likelihood to die infected by COVID-19 and not recovered, after full vaccination.

Vaccine effectiveness is a special case of confirmatory ecological inference, since the estimate of $\hat{RRR}$ is a function of the relative frequency of full vaccinated people in a country (*vaccine coverage*) as a predictor x, and the observed outcome y is the variety over time in the number of deaths with infection. A reduction of y when vaccine coverage raises, would raise the $\hat{RRR}$, too:

$$R\hat{R}R = 1 - e^{\hat{\beta}_x}, \tag{1}$$

where $\hat{\beta}_x$ is the estimate of the coefficient in the Poissonian regression.

The two salient features for an observational study on $R\hat{R}R$ are:

1. The link function is modeled after a member of Poisson family.
2. y (and, possibly x) not only variate over time but variate also across territories G, for a set of fixed over time determinants.

The variance across G can be accounted in two ways:

1. by Within estimation, that is, demeaning y across territories $((y \mid G) - \bar{y_G})$;
2. by a pooled strategy, that is, controlling *per* an identified set of covariates fixed within G. The most obvious covariate would be the recorded population of the territory for that year.

This covariance structure induces overdispersion in y, compared to the assumption for finite samples of Poisson model: $\frac{s^2(y)}{\bar{y}} \sim 1$. Two Poissonian models account for overdispersion: Negative Binomial and Quasi Poisson. For both of them, it holds the assumption for location of a Poissonian model: $\mathbb{E}(Y) = \mu$.

In Negative Binomial, the variance is not equal to the location and it assumes to be corrected through a quadratic relation to an overdispersion parameter θ:

$$\sigma^2(Y_{NB}) = \mu + (\theta \cdot \mu)^2 \tag{2}$$

differently, in Quasi Poisson, the correction is linear:

$$\sigma^2(Y_{QP}) = \theta \cdot \mu \tag{3}$$

Even if the literature sometimes dismisses this point [4], identification for the of link function of θ in Poissonian panels is not trivial. The multiverse for the present study shows that, conditioned on other choices, the results may be extremely sensible to this decision, to the point that alternative models would justify totally opposite scientific findings. Multiverse methodology helps to check one's own models and possibly to visualize the risk of misinformed decision making.

2 Materials

The data sources are the "Our World in Data COVID-19" vaccination dataset [11], the COVID-19 DataHub[6] and the Oxford COVID-19 Government Response Tracker (OxCGRT) [7].

Oxford's Stringency Index (STRG) is a composite measure in the unit interval of the overall pandemic-countering measures (not necessarily limitations) activated in the countries, while Lockdown (LCKD) is a record of the pandemic limitations to mobility in an ordinal scale. Conditions of sample eligibility for a country are:

- $>$ 350 daily reports on deaths with COVID-19 in 2021

- > 1 daily report on vaccine coverage.

174 eligible countries were sampled for a sum of $63,510$ vectors of observations over 365 days (t), even if the sample had missing data on vaccine coverage (x_t) and other covariates. All the missing data before the first chronological recorded $x_t = 0$ are natural zeros. Other missing data are true *missing* and can eventually be substituted by linear interpolation. The choice whether to interpolate or not is exemplary about how a multiverse follows organically the necessary decision making for data analysis.

2.1 Multiverse Model

The multiverse combines:

- raw y_t *vs.* smoothed y_t, mobile averages at day ± 1;
- raw x_t *vs.* natural zeros-filled x_t *vs.* zero fill plus interpolation of x_t;
- 10 lags for x_{t-lag}: from 13 days to 22 days before y_t.

Adopting a multiversal approach, the combined models to estimate the relative risk reduction $\hat{RRR}$ are the following eight:

- QPW-STRG: Quasi Poisson, Within Estimation, control for STRG;
- QPW-LCKD: Quasi Poisson, Within Estimation, control for LCKD;
- NBW-STRG: Negative Binomial, Within Estimation, control for STRG;
- NBW-LCKD: Negative Binomial, Within Estimation, control for LCKD;
- QPFP-STRG: Quasi Poisson, control for the logarithm of population, for the % of people older than 65, and for STRG;
- QPFP-LCKD: Quasi Poisson, control for the logarithm of population, for the % of people older than 65, and for LCKD;
- NBFP-STRG: Negative Binomial, control for the logarithm of population, for the % of people older than 65, and for STRG;
- NBFP-LCKD: Negative Binomial, control for the logarithm of population, for the % of people older than 65, and for LCKD.

All the models account for a total of 480 combinations of estimates of $\hat{RRR}$ (see 1) with p-values $> .05$ or $< .05$. STRG and LCKD, while relatively invariant across time, were still lagged of 14 days.

3 Results and discussion

Smoothing the deaths and vibrating the lags of vaccine coverage held no impact on $\hat{RRR}$. However, the coexistence in the same multiverse of red and blue estimates in Fig. 1 implies a Janus effect in the multiverse model, that is a severe risk of

misspecification given seemingly equivalent analytical alternatives [12]. Figure 1 shows the source for Janus effect in two choices: STRG and Negative Binomial.

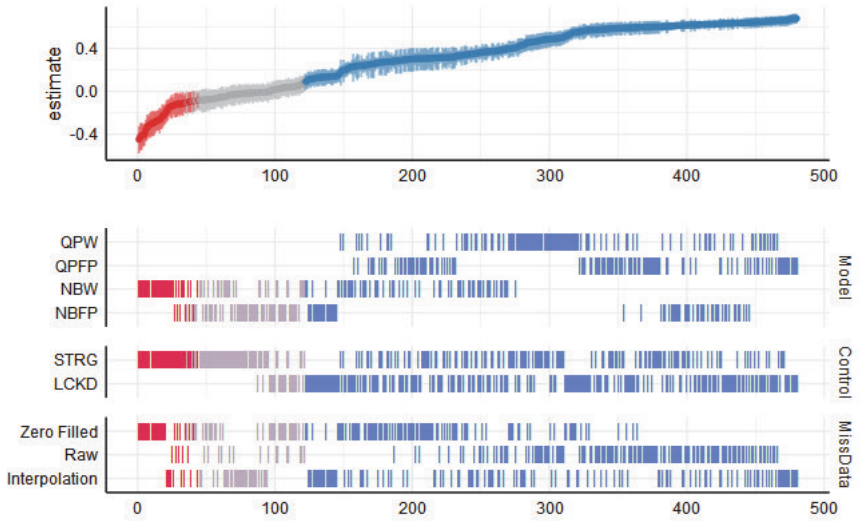


Fig. 1 Specification Curve: in grey the $R\hat{R}R$ estimates associated with not significant p-values $> .05$, in blue the significant (p-values $< .05$) and positive $R\hat{R}R$ estimates, and in red the significant (p-values $< .05$) and negative $R\hat{R}R$ estimates.

An explanation of why Negative Binomial is biased for panel data is that it is very sensible to bias of incidental parameters [1]. NB-models amplify by an exponent the estimated effect in observations with small n, compared to observations with a much larger n [16]. In this case, the Janus effect would be triggered by the control for a composite Oxford Stringency Index. The argument here would be to avoid composite indexes as controls in regressions.

Specification Curve also helps to measure the impact of operationalizations on raw data: filling the natural zeros, while logically appropriate, would induce a statistical bias, that is then corrected only after interpolation of the other missing values.

According to original proponents [13], the slope of the Specification Curve is informative about the robustness of the results. In particular the flat sections are the true *statistical signal* in the multiverse: the existence of flat section implies that as much as the researcher's decisions allow a certain model variance [18], the variance in the estimate will still be minimized within the flat section. In other words, a flat section is informative of convergence of methods towards a likely estimate. If this is true, then the best estimate of $R\hat{R}R$, controlled by mobility limitations, should lie between .45 and .55, since it is the flattest section in Figure 1.

References

[1] Allison PD, Waterman RP (2002) Fixed–Effects Negative Binomial Regression Models. Sociological Methodology 32(1):247–265. 10.1111/1467-9531.00117

[2] Benjamin DJ, Berger JO, Johannesson M, et al (2018) Redefine statistical significance. Nature Human Behaviour 2(1):6–10. 10.1038/s41562-017-0189-z

[3] Del Giudice M, Gangestad SW (2021) A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. Advances in Methods and Practices in Psychological Science 4(1):2515245920954,925. 10.1177/2515245920954925

[4] Gardner W, Mulvey EP, Shaw EC (1995) Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychological Bulletin 118(3):392–404. 10.1037/0033-2909.118.3.392

[5] Gelman A, Loken E (2014) The statistical crisis in science: data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. American Scientist 102(6):460–466

[6] Guidotti E, Ardia D (2020) COVID-19 Data Hub. Journal of Open Source Software 5(51):2376. 10.21105/joss.02376

[7] Hale T, Angrist N, Goldszmidt R, et al (2021) A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour 5(4):529–538. 10.1038/s41562-021-01079-8

[8] Halsey LG, Curran-Everett D, Vowler SL, et al (2015) The fickle P value generates irreproducible results. Nature Methods 12(3):179–185

[9] Leamer EE (1985) Sensitivity Analyses Would Help. The American Economic Review 75(3):308–313

[10] Masur PK, Scharkow M (2020) specr: Conducting and Visualizing Specification Curve Analyses (Version 0.2.2). https://CRAN.R-project.org/package=specr

[11] Mathieu E, Ritchie H, Ortiz-Ospina E, et al (2021) A global database of COVID-19 vaccinations. Nature Human Behaviour 5(7):947–953. 10.1038/s41562-021-01122-8

[12] Patel CJ, Burford B, Ioannidis JPA (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. Journal of Clinical Epidemiology 68(9):1046–1058. 10.1016/j.jclinepi.2015.05.029

[13] Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. Nature Human Behaviour 4(11):1208–1214. 10.1038/s41562-020-0912-z

[14] Steegen S, Tuerlinckx F, Gelman A, et al (2016) Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science 11(5):702–712. 10.1177/1745691616658637

[15] Trafimow D, Amrhein V, Areshenkoff CN, et al (2018) Manipulating the Alpha Level Cannot Cure Significance Testing. Frontiers in Psychology 9:699. 10.3389/fpsyg.2018.00699

[16] Ver Hoef JM, Boveng PL (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88(11):2766–2772. 10.1890/07-0043.1

[17] Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond "$p < 0.05$". The American Statistician 73(sup1):1–19

[18] Young C, Holsteen K (2015) Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. Sociological Methods & Research 46(1):3–40. 10.1177/0049124115610347

Efficiency and feasibility of two stage sampling designs for estimating SARS-CoV-2 epidemic

Efficienza e fattibilità di disegno di campionamento a due stadi per la stima dell'epidemia di SARS-CoV-2

Pietro Demetrio Falorsi, Vincenzo Nardelli and Giuseppe Arbia

Abstract. Spatial information and aggregate data on the number of verified infected, either hospitalized or in compulsory quarantine, are utilized to improve a standard two-stage sampling design broadly adopted for studying human populations. The basic strategy, is based on spatially balanced sampling. The conditions for their efficiency are analytically proven. The relative advantages are shown through a simulation study.

Abstract. *Le informazioni spaziali e i dati aggregati sul numero di infetti ricoverati in ospedale o in quarantena obbligatoria, vengono utilizzati per migliorare un disegno standard di campionamento a due stadi ampiamente adottato per lo studio delle popolazioni umane. La strategia di base si basa su un campionamento spazialmente bilanciato. Le condizioni per la sua efficienza definite. I vantaggi relativi sono mostrati attraverso uno studio di simulazione.*

Key words: Local cube method, Anticipated Variance, Spatial Correlation.

1 Introduction

The aim of this paper is to improve the current practice in epidemic data collection. This work examines two-stage sampling, a technique widely used by the National Statistical Institutes to conduct surveys involving direct interviewing (such as large-scale household surveys) in estimating the critical parameters of the Sars-Cov2 epidemic. We first define the conditions for achieving a theoretical efficiency under a

[1] Pietro Demetrio Falorsi, International Consultant – Roma - piero.falorsi@gmail.com
Vincenzo Nardelli, Università degli Studi di Milano-Bicocca - v.nardelli2@campus.unimib.it
Giuseppe Arbia, Università Cattolica del Sacro Cuore – Roma – giuseppe.arbia@unicatt.it

super-population model which foresees the phenomenon's spatial correlation. We then examine practical sampling choices that allow us to approach the theoretical efficiency but are feasible in practice.

Spatially balanced sampling in the first stage (Grafström et al., 2012; Grafström and Lundström, 2013; Tillé, 2020) and a simple random sampling in the second stage is the basic sampling design here examined. This strategy merge efficiency and feasibility. Since the phenomenon of infection is positively correlated, spatial sampling allows gaining efficiency by spreading the first-stage sample over space; moreover, the balancing strategy leverages aggregated open data on the number of people hospitalized or in a compulsory quarantine. Simple random sampling in the second stage ensures the feasibility of the proposed strategy because the sampling mechanism is simple and does not require information at the unit level. We carried out a comparative analysis of different sampling designs to verify the effectiveness of the proposed strategy using simulated data.

2 The basic sampling framework

We define U as the population of individuals of size N. Suppose that U can be partitioned into M subpopulations, called clusters, denoted as $U_1, \dots, U_i, \dots, U_M$. The clusters are the first-stage sampling units such as neighbourhoods, hospitals or cities. The set of clusters is symbolically represented as $U_{\mathrm{I}} = \{U_1, \dots, U_i, \dots, U_M\}$. Cluster U_i has N_i individuals, being $N = \sum_{i \in N_i} N_i$. Let v_{ij} $(i = 1, \dots, M;\ j = 1, \dots, N_i)$ be the value of the variable v on the verified status of infection for person j in cluster U_i, where $v_{ij} = 1$ if person ij has a verified state of infection and $v_{ij} = 0$ otherwise. The people with $v_{ij} = 1$ could be either hospitalized or in compulsory quarantine. Let $\mathcal{V}_i = \sum_{j=1}^{N_i} v_{ij}$ *and* $\mathcal{V} = \sum_{i=1}^{M} \mathcal{V}_i$ be the known totals of the verified infected people in cluster U_i and in population U. In most situations, public health authorities know the aggregate quantities $\mathcal{V}_i$ and often disseminate such amounts as open data.

Let y_{ij} be the value of variable $\mathcal{y}$ for person j in cluster U_i, where $\mathcal{y}$ is equal to 1 if the person is infected and 0 otherwise. If $v_{ij} = 1$, $y_{ij} = 1$; and if $v_{ij} = 0$, it is possible that either $y_{ij} = 1$ (an infected person for whom the infection has not yet been verified) or $y_{ij} = 0$ (a healthy person). The target parameter, Y, is the total number of infected people, that is,

$$Y = \sum_{i \in U_{\mathrm{I}}} \sum_{j \in U_i} y_{ij} = \sum_{i \in U_{\mathrm{I}}} Y_i \qquad (2.1)$$

where Y_i indicates the number of infected people of cluster i.

We select a sample S using a two-stage sampling design without replacement. A first-stage sample, S_{I}, of fixed size m is selected without replacement from U_{I}, with inclusion probabilities $\pi_{\mathrm{I}i}$ $(i = 1,2 \dots, M)$. A standard solution is to select cluster U_i with Probability Proportional to Size (PPS): $\pi_{\mathrm{I}i} = m\frac{N_i}{N}$.

A second-stage sample, $S_{\mathrm{II}i}$, of fixed size, $\bar{n}$, is drawn from each sampled cluster U_i by *drawing the units* without replacement and with equal probabilities. The second-

stage inclusion probability, $\pi_{\mathrm{II}i}$, of people in the sampled cluster U_i is $\pi_{\mathrm{II}i} = \frac{\bar{n}}{N_i}$. The final inclusion probability of person *j* being selected from cluster U_i is $\pi_{ij} = \pi_{\mathrm{I}i}\pi_{\mathrm{II}i} = m\frac{N_i}{N}\frac{\bar{n}}{N_i} = m\frac{\bar{n}}{N}$. The sampling process is *self-weighting* in the sense that all the units in U have an equal probability of being selected, irrespective of their cluster. The Horvitz-Thompson estimator of Y is

$$\hat{Y} = \sum_{i\in S_{\mathrm{I}}}\sum_{j\in S_{\mathrm{II}i}}\frac{1}{\pi_{ij}}y_{ij} \qquad (2.2).$$

The sample S_{I} is approximately balanced on the d_{I} variables if $\sum_{i\in S_{\mathrm{I}}}\frac{d_{\mathrm{I}i}}{\pi_{\mathrm{I}i}} \cong \sum_{i\in U_{\mathrm{I}}} d_{\mathrm{I}i}$. A well spread sampling avoids the selection of neighbouring units and it is said to be spatially balanced. As for the algorithms that carry out the spreading of the sample, Grafström (2012) introduced the Local Pivotal (LP) method, which enables the selection of unequal probability samples that are well spread over the population. The method uses the distance between units to create small joint inclusion probabilities for nearby units, forcing the samples to be well dispersed. Grafström and Tillé (2013) proposed the Local Cube (LC) method, which selects a sample that is both spread out geographically and balanced in terms of auxiliary variables, LC is an extension of the LP method.

3 Efficiency vs feasibility

Consider the following generalized linear model *M*:

$$y_{ij} = \tilde{y}_{ij} + u_{ij}$$

where $\tilde{y}_{ij} = \Pr(y_{ij} = 1)$, and u_{ij} is a random error where $E_M(u_{ij}) = 0$, $V_M(u_{ij}) = \sigma_u^2$, $Cov_M(u_{ij}, u_{\ell k}) = \sigma_u^2\rho_{ij,\ell k}$, in which $E_M(\cdot)$, $V_M(\cdot)$, and $Cov_M(\cdot,\cdot)$ are expectation, variance and covariance under the model and σ_u^2 is the homoscedastic variance of errors. The correlations $\rho_{ij,\ell k}$ are supposed to be decreasing in function of the distance $\delta_{ij,\ell k}$ between units ij and ℓk. The probability $\tilde{y}_{ij}$ can be modelled as a Lipschitz continuous function $\tilde{y}_{ij} = g(\boldsymbol{x}_{ij})$, where $\boldsymbol{x}_{ij}$ is a column vector of auxiliary variables specific for unity ij. The accuracy of the proposed sampling strategies may be measured by the anticipated variance given by

$$AV(\hat{Y}) = E_P E_M(\hat{Y} - Y)^2 = A + B + C + D + E - V_M(Y)$$

where $E_P(\cdot)$ denotes the expectation over repeated sampling, and

$$A = \sigma_u^2\sum_{i\in U_{\mathrm{I}}}\sum_{\ell\neq i}\frac{N_i}{\pi_{\mathrm{I}i}}\frac{N_\ell}{\pi_{\mathrm{I}\ell}}\pi_{\mathrm{I}i,\mathrm{I}\ell}\bar{\rho}_{i,\ell} \qquad B = V_{\mathrm{I}P}\left(\sum_{i\in S_{\mathrm{I}}}\tilde{Y}_i\frac{1}{\pi_{\mathrm{I}i}} - \sum_{i\in U_{\mathrm{I}}}\tilde{Y}_i\right),$$

$$C = \sum_{i\in U_{\mathrm{I}}}\sum_{j\in U_i}\sum_{k\neq j}\sigma_u^2\rho_{ij,ik}\frac{1}{\pi_{\mathrm{I}i}}\frac{\pi_{\mathrm{II}ij,\mathrm{II}ik}}{\pi_{\mathrm{II}i}^2},$$

$$D = \sum_{i\in U_{\mathrm{I}}}\frac{1}{\pi_{\mathrm{I}i}}N_i\left(\frac{N_i - \bar{n}}{\bar{n}}\right)\sigma_{\mathrm{II}\tilde{y}}^2, E\sum_{i\in U_{\mathrm{I}}}\frac{1}{\pi_{\mathrm{I}i}}\frac{N_i}{\bar{n}}\sigma_u^2,$$

where $V_{1P}(.)$ denotes the first-stage sampling variance, $\pi_{1i,1\ell}$ is the joint inclusion probability of selecting clusters U_i and U_ℓ in the first-stage sampling, $\pi_{2i|j,2i|k}$ is the second-stage joint inclusion probability of selecting units i and j in cluster U_i,

$$\text{and } \tilde{Y}_i = \sum_{j\in U_i} \tilde{y}_{ij} = \sum_{j\in U_i} g(\boldsymbol{x}_{ij}) \quad , \qquad \sigma^2_{2i|\tilde{y}_i} = \frac{1}{N_i - 1}\sum_{j\in U_i}\left(\tilde{y}_{ij} - \frac{\tilde{Y}_i}{N_i}\right)^2 ,$$

$$\bar{\rho}_i = \frac{1}{N_i(N_i-1)}\sum_{j\in U_i}\sum_{k\neq i}\rho_{ij,ik} , \qquad \bar{\rho}_{i,\ell} = \frac{1}{N_i N_\ell}\sum_{j\in U_i}\sum_{k\in U_\ell}\rho_{ij,\ell k}.$$

If we adopt a sampling algorithm for the first stage sampling to ensure that joint inclusion probabilities $\pi_{1i,1\ell}$ are small whenever $\bar{\rho}_{i,\ell}$ is large, we minimize the term A. As demonstrated in theorem 1 in Grafström and Lunsdström (2013), if the first-stage sampling is well-spread on the totals $\boldsymbol{d}_{1i} = \sum_{j\in U_i}\boldsymbol{x}_{ij}$, and we balance on the same totals, then we obtain the balancing on the theoretical unknown values $\tilde{Y}_i$. In this way the term B tends to be negligible. We may diminish the term C geographically spreading the second-stage sampling in the clusters. If we adopted a spatial sampling mechanism geographically spreading the sampling in each unit, the joint inclusion probabilities $\pi_{2i|j,2i|k}$ become very small when the units are close, and the correlation $\rho_{ij,ik}$ is high. Eventually, if we spread geographically the second stage sampling on the $\boldsymbol{x}_{ij}$ values and balance on the same variables, then we obtain the balancing on the theoretical unknown values $g(\boldsymbol{x}_{ij})$. In this way the term D tends to be negligible. In synthesis, taking as fixed the first-order inclusion probabilities π_{1i} and $\pi_{2i|i}$, the maximum efficiency is obtained by spreading and balancing each of the two stages of sampling selection so the term E becomes the dominant term of the AV.

Spreading and balancing in the second stage may not be practicable. Therefore, a feasible strategy should ensure that the second stage selection is carried out autonomously in each sample cluster. In this case, it would be better to adopt a simple SRSWOR design. Spreading and balancing the second-stage sampling would necessitate the availability of a central population register containing as many records of the geographical and balancing variables as there are units in the population. Realistically, this is not the case for most situations that we may encounter in practice.

A feasible strategy is to individuate a $\boldsymbol{z}_i$ vector of known auxiliary variables at the cluster level reasonably correlated to the dissemination of pandemic. We may then define a suboptimal model $y_{ij} = \tilde{y}_{(z)i} + u_{(z)ij}$, where $\tilde{y}_{(z)i} = h(\boldsymbol{z}_i)$ denotes a Lipschitz continuous function, which returns the same value for all the people in the same cluster, and the residual $u_{(z)ij}$ is given by $u_{(z)ij} = u_{ij} + g(\boldsymbol{x}_{ij}) - h(\boldsymbol{z}_i)$.

We can spread the first-stage sampling directly on the z_i variables and balance on the same totals to have a balancing effect on the unknown function $h(z_i)$, and hence also have a small $\pi_{1i,1\ell}$ when $\bar{\rho}_{i,\ell}$ is large. With this strategy we have

$$AV(\hat{Y}) \cong R_z + C + D + E - V_M(Y),$$

$$\text{where } R_z = V_{1P}\left[\sum_{i\in S_1}\left(\tilde{Y}_i - N_i h(z_i)\right)\frac{1}{\pi_{1i}} - \sum_{i\in U_1}\left(\tilde{Y}_i - N_i h(z_i)\right)\right].$$

More the terms $N_i h(z_i)$ are close to the theoretical values $\tilde{Y}_i$ more the variance R_z of residuals is close to zero, and we gain in the efficiency.

In the context of the pandemic, possible and reasonable choices of z_i are:

$$z_i = \bar{V}_i \ (3.1); \quad z_i = k_i \quad (3.2); \quad \text{and} \quad z_i = (\bar{V}_i, k_i')' \quad (3.3)$$

where $\bar{\mathcal{V}}_i = \mathcal{V}_i / N_i$ and k_i is the vector of the geographical coordinates of the cluster U_i.

4 Empirical results

We test the proposed sampling methodologies on a generated dataset representing an artificial population. in which people are free to move around the territory randomly and, according to their social network, can meet other people and eventually become infected. Each individual of the population can be classified as susceptible, infected or removed according to the framework known as the "SIR model". We considered six categories of individuals: susceptible (S), exposed to the virus (E), infected with symptoms (I) and without symptoms (A), and removed from the population either because the individual is healed (R) or dead (D). In the present experiment, we considered the subjects belonging to group (I) to be "known" infected (identified by a positive swab from health screening) and those belonging to group A to be "unknown infected", i.e., those who were not aware of being infected. The simulation algorithm is detailed described in Alleva et al. (2022) and available in the epidsample R package (Nardelli, 2020).

In the experiment we consider the two stage PPS sampling design illustrated in Section 2. The following six schemas of selection are examined: (1) Fixed-size Probability Proportional to Size (FPPS); (2) the LP method; (3) the LC method Based on Verified Infected (LCBV) where the balancing variables are based on specification 3.3a of the z variables; (4) the LC method Based on Verified Infected (LCBG); the balancing is based on specification 3.3b of the z variables; (5) the LC method Based on Verified infected and on Geography (LCBVG) where the balancing is based on specification 3.3c of the z variables; (6) the LC method Based on the $g(\cdot)$ Model at the unit level which best explains the generation of the spread infection (LCBBM). The model g(·), considered in the LCBBM design is based on the following variables: the number of people, the total distance and number of travels, the total number of contacts and the rate of the known infected.

Using the artificial population generated as described previously, we simulated a sample survey at multiple time points during different phase of the epidemic and lockdown periods. For each combination of the simulation parameters, we repeated the Monte Carlo exercise for 10,000 runs so as to ensure the convergence of the sampling methods. We simulated second-stage sampling with a Monte Carlo experiment for different sample cells with different numbers of people sampled in each cell. For simplicity, in Table 1 we report the results for day 15 (3 people for 80 cells) and display the true value, the estimate obtained as the mean of the simulations, the relative bias expressed in absolute terms (RAB) and the standard error (SE). The results of the simulations obtained using other parameter combinations do not add any further insight with respect to those presented here. All the sampling methods considered produce unbiased estimates, as expected. Furthermore, they show consistency, although with very different convergence speeds and a significantly lower standard error (SE) than proportional sampling. The sampling balanced with

model at the unit level which best explains the generation of the spread infection (LCBBM) is the most efficient and displays the lowest standard error. However, as discussed in the previous section, this method collides with feasibility so the second best method is the balanced one on verified infected (LCBV) which has a very similar standard error.

Table 1. Simulation results for day 15 (3 people – 80 cells)

Sampling method	True value	Estimate	RAB	SE
FPPS	1,035	1,048	0.0125	0.37
LP	1,035	1,035	0.0001	0.34
LCBV	1,035	1,034	0.0006	0.25
LCBG	1,035	1,037	0.0015	0.34
LCBVG	1,035	1,028	0.0071	0.32
LCBBM	1,035	1,031	0,0032	0.23

5 Conclusion

The suggested technique has the benefit of being more efficient than the benchmarking case of a regular PPS sampling design while maintaining feasibility. With spatially correlated data, we created an explicit definition of the expected variance. Furthermore, we identified the requirements for increasing efficiency in the two-stage sample design that takes use of the available auxiliary variables, resulting in a promising empirical outcome. This methodology might be applied to the tracing of COVID-19 variants. Future methodological advances will concentrate on determining the relative impact of the AV components on both actual and artificially generated data.

References

1. Alleva, G., Arbia, G., Falorsi, P. D., Nardelli, V. & Zuliani, A. (2022). A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design. Journal of Official Statistics, Accepted.
2. Grafström A, Lundström NL, Schelin L. (2012). Spatially balanced sampling through the pivotal method Biometrics, 68, 2, 514-20.
3. Grafström A, Lundström. N.L, (2013). Why Well Spread Balanced Samples Are Balanced. Open Journal of Statistics, 3, 1, 36-41
4. Grafström, A. and Tille´, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals". Environmetrics, 24, 120–131.
5. Nardelli, V. (2020) Epidsampler R Package - https://vincnardelli.github.io/epidsampler
6. Tillé, Y. (2020). Sampling and estimation from finite populations. John Wiley & Sons.

Evaluating the impacts of Covid-19 on the overall Italian death process via Functional Data Analysis

Valutazione degli impatti dell'epidemia di Covid-19 sul processo di mortalità in Italia, via Functional Data Analysis

Riccardo Scimone, Alessandra Menafoglio, Laura M. Sangalli, and Piercesare Secchi

Abstract We propose a novel approach to modeling daily death counts in Italian administrative units, namely Provinces and Municipalities. In particular, starting from data collected by ISTAT during the period 2011-2020, we look at the mortality density along single years, in each province, embedding the analysis in a compositional framework, whose geometry provides a natural amplitude alignment, allows for comparison among differently populated Provinces, and gives interesting insights on the temporal dynamic of the death phenomenon during the years. The majority of the analysis is carried out in an infinite dimensional compositional framework, namely in the Hilbert space B^2 of strictly positive density functions. Combining functional linear models in this Hilbert space with spatial techniques developed in the framework of Functional Geostatistics, we are able to precisely quantify how the structure of the overall death process has been affected by the pandemic waves during 2020. All the results illustrated in this talk are based on [1].

Abstract *Proponiamo un approccio originale alla modellazione del numero di morti giornaliere nelle unità amministrative Italiane, ovvero Province e Comuni. In particolare, a partire dai dati raccolti da ISTAT negli anni dal 2011 al 2020, ci con-*

Riccardo Scimone
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
Center for Analysis, Decision and Society, Human Technopole
e-mail: riccardo.scimone@polimi.it

Alessandra Menafoglio
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
e-mail: alessandra.menafoglio@polimi.it

Laura M. Sangalli
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
e-mail: laura.sangalli@polimi.it

Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy
Center for Analysis, Decision and Society, Human Technopole
e-mail: piercesare.secchi@polimi.it

centriamo sulle densità di mortalità nei singoli anni, in ogni provincia, inserendo l'analisi in un contesto di analisi dati composizionale, la cui geometria consente un naturale allineamento, permette di confrontare province con popolosità variabile e apre interessanti prospettive sulla dinamica temporale della mortalità negli anni. Gran parte dell'analisi viene svolta in un contesto composizionale e infinito dimensionale, ed in particolare nello spazio di Hilbert B^2, costituito da funzioni di densità strettamente positive. La combinazione di modelli lineari funzionali in tale spazio di Hilbert, unita all'applicazione di tecniche di statistica spaziale provenienti dall'ambito della Geostatistica per dati funzionali, ci permette di quantificare come la struttura del processo di mortalità sia stata colpita ed alterata dalle ondate pandemiche. Tutti i risultati illustrati in questa presentazione sono basati sul lavoro [1].

Key words: Compositional Data Analysis, Functional Statistics, Bayes Space, Functional Geostatistics

1 Introduction

In order to properly analyze the impacts of the Covid-19 pandemic on the Italian mortality process, we deem appropriate to focus the analysis not only on deaths directly attributed to the disease and its variants, rather favoring a broader point of view by looking at death counts for all causes, as collected by the Italian Statistical Institute[1]. These data contain information about any shock which the Italian death process underwent, be it directly caused by the pandemic, an indirect consequence of it, or even totally unrelated to Covid-19. Several practical reasons also drive this choice: daily death counts from all causes are very high quality data, being scarcely affected by problems of missing data or controversial attribution of the cause of death; in addition they have fine temporal and spatial granularity, being collected at the Italian municipality level. We show in Fig. 1 the death counts referring to people more than 70 years old, at the province level, for the years spanning from 2017 to 2020: we will focus on this elderly class for the talk, because this is the most affected by the effects of Covid-19; analyses and insights regarding other age classes are available at https://github.com/RiccardoScimone/Mortality-densities-italy-analysis.git. A somewhat stable seasonal behavior can be identified, during 2017-2019, by a simple visual inspection: in particular, higher death tolls characterize the coldest and hottest period of the year. On the other hand, this stability is clearly broken during 2020, in which many provinces present abnormal behaviors in correspondence of the first and second waves: for instance Bergamo, drawn in orange, is dramatically affected only by the first wave, Naples and Rome only by the second one, while Milan faced both.

[1] All the data object of our work can be freely downloaded at https://www.istat.it/it/archivio/240401

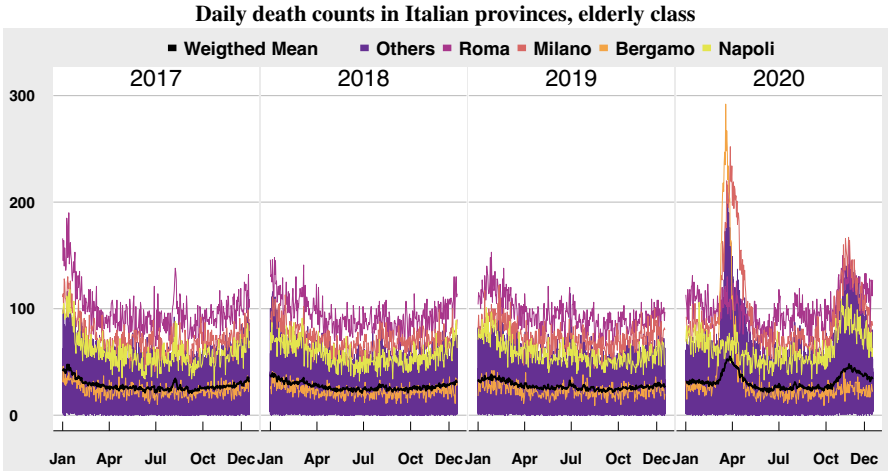


Fig. 1 We show the total amount of deaths in each day, in each year, for each Italian province, referring only to the elderly class (70+ years), it being the most affected by the pandemic. All provinces are drawn in purple, but for some particularly significant provinces which are highlighted in different colors.

Rather than directly analyzing death counts we center our analysis on the empirical distributions

$$p_{iyt} = \frac{d_{iyt}}{\sum_t d_{iyt}} \qquad \text{for } t = 1, \ldots, 365,$$

where d_{iyt} stands for the death count registered in province (or municipality) i in the day t during year y. The collection $\{p_{iy}\}_{iy}$ of discrete distributions can be seen as the empirical distribution functions of the absolutely continuous random variables T_{iy}, each one modeling the time of death of a person who passed away in province i during year y. Analyzing the $\{p_{iy}\}_{iy}$ corresponds to a *compositional data analysis* framework ([6, 7]), i.e., analysis of vectors whose component are parts of a given total. We show in Fig. 2 such family, for the elderly class in Italian Provinces, during 2017-2019. Is it clear that such representation provides a natural alignment with respect to the raw death counts shown in Fig. 1. Moreover, looking at densities highlights different temporal dynamics: the most striking example is the density corresponding to the province of Bergamo, whose peak during the first pandemic wave appears much sharper and dramatic with respect to bigger provinces as Milan, while in the raw counts representation this was less evident.

2 Summary of results

In this talk we will review the main results of our analysis, which is extensively described in [1]. Here follows a short list of our key results:

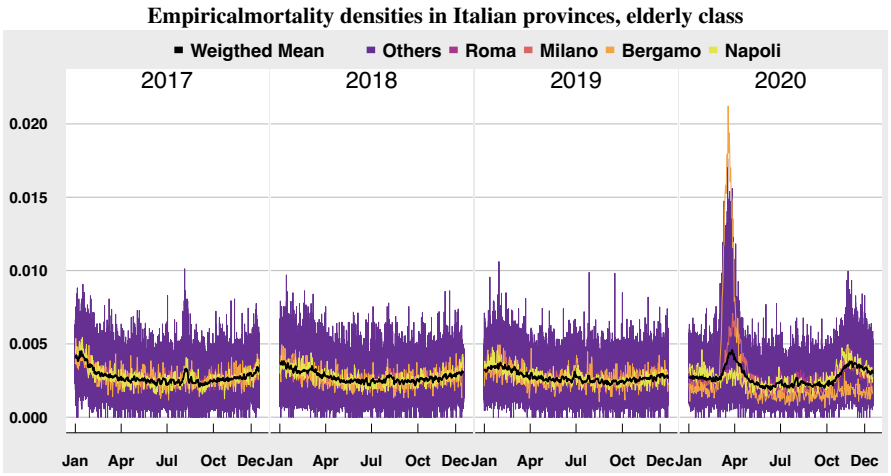


Fig. 2 We plot the $p_{iy_{iy}}$ for Italian provinces, in purple, but for some selesected provinces which are represented with different colors. The black solid line is the weighted mean density, where the weight for each province has been set to be proportional to its population.

- Starting from the $\{p_{iy}\}_{iy}$, we deliver an exploratory analysis based on the *Wasserstein distance* ([5]) for probability density functions, using it as a tool to investigate the stability of the death density process along the years.
- We go beyond the metric embedding, developing the compositional analysis in the Hilbert space B^2 of density functions ([2, 4, 3]). In particular, we smooth the $\{p_{iy}\}_{iy}$ via *compositional splines* ([8, 9]), formulating then a family of function-on-function linear models ([14]) in B^2 to decouple the mortality densities in a *predictable* and an *unpredictable* component.
- We use the previous decoupling as a starting point for a spatial analysis: via functional variography ([10, 11]) and Principal Component Analysis in the B^2 space ([13]) we are able to consistently attribute to the pandemic waves the perturbation of the spatial correlation structure of the death process.
- Finally, we use *spatial downscaling* techniques ([12]) to perform analysis and anomaly detection at the very granular scale of Italian municipalities.

References

1. Scimone, R., Menafoglio, A., Sangalli, L.M., Secchi,P. (2021): A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. Spatial Statistics, doi: https://doi.org/10.1016/j.spasta.2021.100541
2. Egozcue, J., L. Díaz–Barrero, J. & Pawlowsky-Glahn, V. (2006) Hilbert Space of Probability Density Functions Based on Aitchison Geometry. *Acta Mathematica Sinica, English Series*. **22** pp. 1175-1182
3. Pawlowsky-Glahn, V., Egozcue, J. & Boogaart, K. (2014) Bayes Hilbert Spaces. *Australia And New Zealand Journal Of Statistics*. **56** pp. 171-194

4. Van den Boogaart, K., Egozcue, J. & Pawlowsky-Glahn, V. (2010) Bayes linear spaces *SORT*. **34** pp. 201- 222
5. Villani, C. (2003) Topics in Optimal Transportation. (American Mathematical Society)
6. Aitchison, J. (1982) The Statistical Analysis of Compositional Data. *Journal Of The Royal Statistical Society. Series B (Methodological)*. **44** pp. 139-177
7. Aitchison, J. (1986) The Statistical Analysis of Compositional Data. (Chapman & Hall, Ltd. London)
8. Machalová, J., Hron, K. & Monti, G. (2015) Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal Of Applied Statistics*. **43**
9. Machalová, J., Talská, R., Hron, K. & Gába, A. (2020) Compositional splines for representation of density functions. *Computational Statistics*. pp. 1-34
10. Nerini, D., Monestiez, P. & Manté, C. (2010) Cokriging for spatial functional data. *Journal Of Multivariate Analysis*. **101**, 409-418
11. Menafoglio, A., Secchi, P. & Dalla Rosa, M. (2013) A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal Of Statistics*. **7**, 2209-2240
12. Goovaerts, P. (2008) Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*. **40**, 101-128
13. Hron, K., Menafoglio, A., Templ, M., Hrůzová, K. & Filzmoser, P. (2016) Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*. **94** pp. 330-350
14. Ramsay, J. & Silverman, B. (2005) Functional Data Analysis. (Springer)

Developing countries, migration and migrants

Domestic Violence in Africa: A Glance Through the DHS Survey

La Violenza Domestica in Africa: Uno Sguardo Attraverso l'indagine DHS

Micaela Arcaio,[1*] Daria Mendola,[1] Anna Maria Parroco[1]

Abstract Recent data states that 33 per cent of women in Sub-Saharan Africa are survivors of domestic violence. This work aims at assessing the association between women's characteristics, their environment, and their history of violence in fifteen African countries. Three kinds of violence were explored: physical, emotional, and sexual, each one exerted by the current partner. The data are from the Demographic and Health Survey, in which a special module assessed domestic violence in female respondents. Using three independent logistic regression models, we found that experiencing violence of any kind is associated with a history of past violence, and while women empowerment is not protective, partners with high education are less likely to be violent.

Abstract *Dati recenti provano che il 33% delle donne in Africa Sub-Sahariana è vittima di violenza del partner. Questo studio mira a valutare l'associazione tra le caratteristiche delle donne, il loro ambiente, e la loro storia di violenza in quindici paesi africani. In particolare, lo studio si focalizza su tre tipi di violenza: fisica, emotiva e sessuale. I dati derivano dalla Demographic and Health Survey, in cui un modulo speciale indaga la violenza domestica tra le intervistate. Usando tre modelli indipendenti di regressione logistica, emerge che subire abusi di qualsiasi tipo è associato con un passato di violenza e, nonostante l'empowerment femminile non sia un fattore protettivo, partner con un elevato livello di istruzione sono meno violenti.*

Key words: Abuse; Demographic and Health Survey; intimate partner violence

[1] Department of Psychology, Educational Science and Human Movement (SPPEFF), University of Palermo, Italy; Micaela Arcaio, University of Palermo; email: micaela.arcaio@unipa.it
Daria Mendola, University of Palermo; email: daria.mendola@unipa.it
Anna Maria Parroco, University of Palermo; email: annamaria.parroco@unipa.it
*corresponding author

1 Domestic Violence

Domestic or intimate partner violence (IPV) is defined by the United Nations as a behavioural model of relationships in which one of the partners seeks to either obtain or keep power and control over the other [1]. There are different kinds of domestic violence: physical abuse refers to causing harm; emotional abuse points out not only to verbal abuse but also to social isolation from friends and family; finally, sexual abuse involves forcing a partner into sexual acts without their explicit consent [1]. All countries in the world experience IPV but at different rates and women are overall the most targeted group. Sub-Saharan Africa has the second-highest rate (33%) of lifetime IPV for women aged 15-49, following the 35% in Southern Asia [2].

The analysis of IPV requires more than the study of victims' characteristics [3]. Several studies control for education, both of victims and their partners, with people with higher education being less likely to either perpetrate or suffer IPV [4, 5]. Wealth is also considered a protective factor [6]. There is also evidence of a complex interplay between context and personal characteristics: e.g., as women's access to education and the labour market increases, so does IPV within that nation [7]. Religious women are more protected than atheists, given they can rely on a broader social network; however, there is no clear evidence of differential effects among religious affiliations [12]. IPV is also associated with women's approval of wife-beating [4, 8, 9]. Furthermore, history of violence decidedly plays an important role: women who have a history of violence between parents – i.e., their fathers used violence on their mothers – are much more likely to be abused later in life, due to assimilation of acceptability of this kind of aggressions and perpetuation of the same patterns of violence [10,11]; similarly, men who had violent fathers or witnessed violence become perpetrators as well [6].

2 Data and Methods

This paper focuses on violence against women by their heterosexual partners. Data are drawn from the Demographic and Health Survey (DHS), a nationally representative household survey, covering over 90 countries and 40 years. In particular, we focussed on fifteen surveys in Africa for which the module on domestic violence was administered: Angola, Burundi, Cameroon, Chad, Ethiopia, Gabon, Kenya, Liberia, Mali, Malawi, Rwanda, Senegal, Togo, Zambia, and Zimbabwe. Surveys range from 2015 to 2018, with some exceptions for some countries where available data were less recent.

Information on IPV is collected only among women ever partnered, selected at random in those households involved in the main survey. Questions are both referred to violence perpetrated by the respondent's current partner and to past violent experiences. We restricted our analyses to a sample of almost 40,000

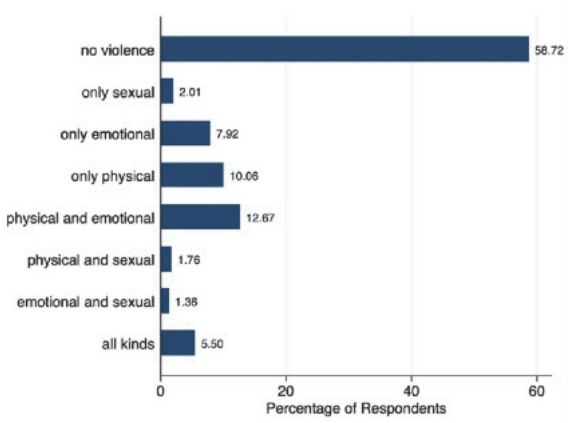
currently partnered women, aged 31 years on average, pooled across the 15 countries.

Domestic violence, our response variable, is assessed via three indicators, made dichotomous:

- "Physical Violence", which takes value 1 if the respondent has been pushed, shook, slapped, punched, threatened at gunpoint by her partner;
- "Emotional Violence", which takes value 1 if the respondent's partner has ever humiliated, threatened with physical harm or insulted her;
- "Sexual Violence", which takes value 1 if the respondent has ever been forced into sexual acts by her partner.

Figure 1 shows the percentages of women who might have experienced any form of the three types of IPV by their partner. More than 40% of respondents have experienced at least one form of abuse by their partner, with almost 6% of them having experienced all three and more than 15% having experienced two.

Figure 1: *Occurrence of the three kinds of violence – percentages*



Given the nature of the phenomenon, and of the response variables, three logistic regression models were defined to assess the effect of the selected regressors, which are the same for all models to improve comparability. These were classified into five different dimensions: 1) *History of Violence*: father used violence on mother; respondent has ever been raped (not by her current partner); no. of people who used violence on the respondent (except for partner); 2) *Respondent's characteristics*: education level, employment status, age, age at first intercourse, religious affiliation, no. of justifications for wife-beating;[1] 3) *Partner characteristics*: partner's age (in terms of age difference among partners), education level and no. of partner's control

[1] Her husband is justified in beating her for: burning the food, arguing, going out without telling him, neglecting her children, refusing sex.

issues;[1] 4) *Household characteristics*: she is the head of the household, quintiles of household's wealth, no. of living children under five years of age; 5) *Context variables*: whether the respondent lives in a rural/urban setting, and the country she lives in. These last country dummies account for regional specificity.

3 Results

The models provide the odds for women to have experienced at least one physical, emotional, or sexual violence by their current partner (estimates in Table 1). History of violence plays a crucial role in predicting every kind of IPV [10,6,11]. Noticeably, women are almost twice as likely to be physically abused if their father harmed their mother, and thrice as likely to be sexually violated by their partners if they had been previously raped. The greater the number of abusers in women's lives the higher the odds to be abused by the current partner.

As for the respondent's characteristics, older women are more likely to experience violence, while there are no significant effects of their education, nor of their age at first intercourse. However, as debated above, women who work are 10% more likely to experience physical or emotional abuse. The religious environment also matters despite but not for every kind of violence: atheist women are 25% more likely than Christian women to be physically abused by their partners, whereas Muslim women are 25% and 30% less likely to be physically or emotionally abused than Christian women respectively. Interestingly, the number of justifications that women provide for wife-beating does contribute to increasing the odds of each form of violence.

Partners with a higher education level are also associated with women being 20% less likely to be physically or emotionally abused, and almost 30% less likely to be sexually abused. Those who tend to control their partners more are more likely to act violently. When he is older than her, the odds of being abused lower by about 2% per year but only for physical and sexual violence.

When it comes to household characteristics, women in the richest quintile are less likely to be either physically or sexually abused. Moreover, when the respondent is herself the head of her household, her odds of experiencing physical abuse lower by 10%. The presence of young children affects the odds of physical violence, by increasing it by about 5% for each child. Finally, women who live in rural areas are less likely to report physical and emotional abuse than those who live in a city.

Furthermore, other things being equal, violence is not equally spread across countries, with Burundi being the riskiest country for physical and sexual violence and Senegal for the emotional violence (with respect to Rwanda, assumed as reference).

[1] The respondent's partner is jealous, accuses her of unfaithfulness, does not allow her to visit female friends, insists on knowing where she is at all times, does not trust her.

Table 1: Logit models for the presence of violence acted by the partner (odds ratio estimates)

		Physical Violence	*Emotional Violence*	*Sexual Violence*
History of violence	***Father used violence on mother***	1.924***	1.633***	1.446***
	She has ever been raped	1.517***	1.776***	3.052***
	No. of abusers in her life (except for partner)	1.285***	1.337***	1.217***
Her characteristics [b]	***Age***	1.543***	1.742***	1.567***
	In paid work	1.128***	1.145***	1.042
	Religion (ref. Christian)			
	Atheist	1.267**	1.094	0.845
	Muslim	0.766***	0.687***	0.930
	Traditional/Animist	1.157	1.073	1.337
	Other	0.726	0.784	0.376***
	No. woman's justification for wife-beating	1.089***	1.049***	1.108***
His characteristics	***He is older (age difference)***	0.987***	1.002	0.990***
	Education (ref. None)			
	Primary	1.046	1.055	1.024
	Secondary	0.986	0.919	0.904
	Higher	0.797**	0.774**	0.690**
	No. of control issues	1.924***	1.633***	1.446***
Household characteristics	***She is head of the house***	0.876***	0.947	0.899
	Wealth Index (ref. Poorest)			
	Poorer	1.001	0.995	0.991
	Middle	0.970	1.079	0.980
	Richer	0.881**	0.955	0.893
	Richest	0.844**	0.983	0.730***
	No. of her own children under 5	1.047**	1.032	1.010
Con-text [a]	***Rural area (ref. Urban)***	0.875**	0.905*	0.869
	Constant	0.191***	0.0944***	0.0431***
	Observations	38,953	38,959	35,705

*** p < 0.01, ** p < 0.05, * p < 0.10

a) controlled for country dummies; b) controlled for covariates found not statistically significant: Respondent's education; respondent's age at first intercourse

4 Conclusions

Domestic violence against women is a worldwide issue. However, it is much more pervasive in some regions of the world, like Sub-Saharan Africa, where social and economic systems in place allow the persistence of discriminatory norms and inequalities. Our analysis offered a glance over factors associated with reported

violence, pointing out the relevance of the phenomenon in some African countries. Our findings indicated that women who have lived in a violent environment in the past tend to be more involved in the same kind of violence.

Noticeably, in contrast with our expectations, women's empowerment (in terms of education and employment) does not act as a protective factor. Indeed, having a paid job outside her house increases women's odds of being abused, and education is not statistically significant. This was observed in some other studies (see [7]) and should deserve further investigation. On the other side, richer families, with highly educated male partners, show less violence.

This study is not free from limitations, the main one likely being the reliability of collected data on violence, often biased by social desirability and the tendency of victims to deny experienced violence. However, we believe that the results offer an interesting contribution to the development of policies to fight violence against women. They suggest that the promotion of a model of women's emancipation in accordance with the cultural and social environment of the country could be successful in reducing violence only in association with more inclusive education and general better living conditions.

References

1. What Is Domestic Abuse? | United Nations. https://www.un.org/en/coronavirus/what-is-domestic-abuse. Published 2022. Cited 13 Feb 2022.
2. World Health Organisation. Violence Against Women Prevalence Estimates, 2018: Global, Regional and National Prevalence Estimates For Intimate Partner Violence Against Women And Global And Regional Prevalence Estimates For Non-Partner Sexual Violence Against Women. Executive Summary. Geneva: World Health Organisation; (2021)
3. Hindin M, Kishor S, Andara D. Intimate Partner Violence Among Couples In 10 DHS Countries: Predictors and Health Outcomes. DHS Analytical Studies No. 18. Calverton, Maryland, USA: Macro International Inc.; (2008)
4. Abrahams N, Jewkes R, Laubscher R, Hoffman M. Intimate Partner Violence: Prevalence and Risk Factors for Men in Cape Town, South Africa. Violence Vict. 21(2):247-264 (2006)
5. Kishor S, Johnson K. Reproductive health and domestic violence: Are the poorest women uniquely disadvantaged?. Demography. (2006) doi:10.1353/dem.2006.0014
6. Jeyaseelan L, Kumar S, Neelakantan N, Peedicayil A, Pillai R, Duvvury N. Physical spousal violence against women in India: some risk factors. J Biosoc Sci. 39(5):657-670 (2007)
7. Levinson D. Family Violence In Cross-Cultural Perspective. London: SAGE; (1989)
8. Gage A. Women's experience of intimate partner violence in Haiti. Soc Sci Med. 61(2):343-364 (2005)
9. Lawoko S. Factors Associated With Attitudes Toward Intimate Partner Violence: A Study of Women in Zambia. Violence Vict 21(5):645-656 (2006)
10. Kishor S, K J. Profiling Domestic Violence: A Multi-Country Study. Calverton, Maryland, USA: Macro International Inc.; (2004)
11. Kalmuss D. The Intergenerational Transmission of Marital Aggression. J Marriage Fam. (1984) doi:10.2307/351858
12. Klomegah R. Intimate Partner Violence (IPV) in Zambia: An Examination of Risk Factors and Gender Perceptions. J Comp Fam Stud. (2008) doi:10.3138/jcfs.39.4.557

Inequalities in undernutrition among Roma and non-Roma children in Western Balkans: an analysis of the determinants

Diseguaglianze nella malnutrizione infantile tra Rom e popolazione maggioritaria nei Balcani Occidentali: un'analisi delle determinanti

Annalisa Busetta[1], Valeria Cetorelli[2], Chiara Puglisi[3]

Abstract
Childhood undernutrition affects physical and cognitive growth and is strongly associated with morbidity and mortality. Roma children in Western Balkans display to date incomparably higher levels of undernutrition, compared to non-Roma children. Employing logistic regression models, we explore the determinants of such dramatic differences using Multiple Indicator Cluster Surveys carried out in Bosnia-Herzegovina, Kosovo, Macedonia, Montenegro, and Serbia in the period 2011-2013 and 2018-2020. Children's age, number of children in the household, type of place of residence and age of the mother at birth are found be especially important factors. In particular, we find substantial differences in the role played by child age on the risk of undernutrition among Roma and non-Roma children.

Abstract
La malnutrizione infantile influenza negativamente lo sviluppo fisico e cognitivo degli individui, ed è fortemente correlata con i livelli di morbilità e mortalità. I bambini Rom che vivono nei Balcani Occidentali mostrano ad oggi livelli di malnutrizione sproporzionatamente più alti rispetto al resto della popolazione. Attraverso modelli di regressione logistica, abbiamo esplorato le determinanti di tali disuguaglianze sulla base di Multiple Indicator Cluster Surveys condotte in Bosnia-Herzegovina, Kosovo, Macedonia, Montenegro e Serbia nei periodi 2011-2013 e 2018-2020. I risultati evidenziano l'importanza dell'età del bambino, del numero di bambini conviventi, del tipo di residenza e dell'età della madre alla nascita del figlio. In particolare, si rilevano sostanziali differenze nel rischio di malnutrizione per età del bambino tra la popolazione Rom e il resto della popolazione.

Key words: Child undernutrition, Roma, Western Balkans, Inequalities

[1] Annalisa Busetta, University of Palermo. E-mail: annalisa.busetta@unipa.it
[2] Valeria Cetorelli, United Nations. E-mail: valeriacetorelli@gmail.com
[3] Chiara Puglisi, European University Institute. E-mail: chiara.puglisi@eui.eu
The views expressed in this study are those of the authors and do not necessarily reflect the views of the United Nations.

1 Introduction

Despite some declines, undernutrition remains a serious problem in low- and middle-income countries, with lasting social, economic and health implications. Undernutrition hinders children from reaching their physical and cognitive potential and makes them much more vulnerable to disease and death [12;17].

The United Nations Sustainable Development Agenda aims at ending all forms of hunger by 2030, ensuring that all people –especially children– have sufficient and nutritious food. The pledge to 'leave no one behind' requires the prioritization of policies that tackle inequalities and reach the most marginalized communities [13]. Roma settlements in the Balkans are known to suffer from persistent social exclusion [11;18;7]. This paper focuses on prevalence of undernutrition among Roma and non-Roma children over the last decade, and sheds light on the main determinants of inequalities with the intent of informing policies.

Literature on the determinants of undernutrition is wide and has significantly increased in the last few decades. Recent studies highlight that economic growth is not always a sufficiently powerful element to reduce undernutrition, neither are interventions focused only on nutrition, health and water, sanitation, and hygiene [2]. Indeed, several studies emphasize the importance of a multifactorial framework for understanding child undernutrition [1] that includes more social and structural factors (wealth, socioeconomic status, and maternal education) associated with child nutritional outcomes.

Literature on undernutrition shows that several maternal characteristics affect the nutritional status of children, among which education, age at marriage and at childbirth. Maternal education has been identified in other populations as a strong protective factor for undernutrition [14]. Young maternal age at childbirth is also a risk factor for child undernutrition and is associated with an inverted U-shaped birth weight [15]. An increased risk of stunting is observed in event of young age at marriage/union [4].

As for child characteristics, sex has been showed to be an important determinant of undernutrition, with males consistently displaying higher risk of stunting [10], while the relationship between child age and undernutrition seems to follow an inverted U-shape pattern [3]. Household income, too, appears to have a substantial impact on children's nutrition, with children growing in households with higher socioeconomic status displaying a lower probability of stunting, wasting and underweight [6].

2 Data and Descriptive Statistics

The analysis is based on 18 Multiple Indicator Cluster Surveys (MICS) covering national and Roma settlements representative probabilistic samples in Bosnia-Herzegovina, Kosovo, Macedonia, Montenegro, and Serbia in the period 2011-2013

and 2018-2020. These surveys were designed by UNICEF to collect comparable data on key indicators on the well-being of women and children. The MICS questionnaire consisted of three core modules. A household module was administered to a responsible adult in the household, with the main purpose of gathering information regarding the age and sex of all household members. All women between ages 15 and 49 who were reported in the household module were eligible for individual interview. Mothers or caretakers of children aged less than 5 at the date of the survey were also asked to complete a child health module. The latter included an anthropometric section which involved recording height and weight of each child using standard measuring boards and electronic scales.

Following the WHO guidelines [16], we calculate three different measures of undernutrition. Children whose height-for-age was more than two standard deviations below the median height-for-age of the WHO Child Growth Standards were classified as stunted. Those whose weight-for-height was more than two standard deviations below the median weight-for-height of the WHO Child Growth Standards were classified as wasted. Those whose weight-for-age was more than two standard deviations below the median weight-for-age of the WHO Child Growth Standards were classified as underweight.

Overall, in 2011-2013 the weighted prevalence of stunting among children under 5 years was 10.8%, the prevalence of underweight 3.7% and the prevalence of wasting 3.4%, whereas in 2018-20 they were respectively 9.0%, 3.5% and 2.9%. The inequalities between Roma and non-Roma children are substantial in all the five Balkan countries under study (particularly for stunting and underweight) and persistent over time. The worst nutritional status is found in Bosnia-Herzegovina, Montenegro, and Serbia, with nearly one Roma child out of five being stunted.

3 Research Method and Preliminary Results

To investigate differences in undernutrition among Roma and non-Roma children, this study employs a logistic regression model with country and year fixed effects. The main independent variable is a binary variable, classifying children into non-Roma or Roma. Specifically, we look at the prevalence of underweight, wasting and stunting among Roma and non-Roma children, controlling for a number of important factors identified by the literature.

The stepwise inclusion of variables shows that the differences between Roma and non-Roma children persist even when controlling for the economic conditions of the household, as well as for selected mother- and child-level socio-demographic characteristics.

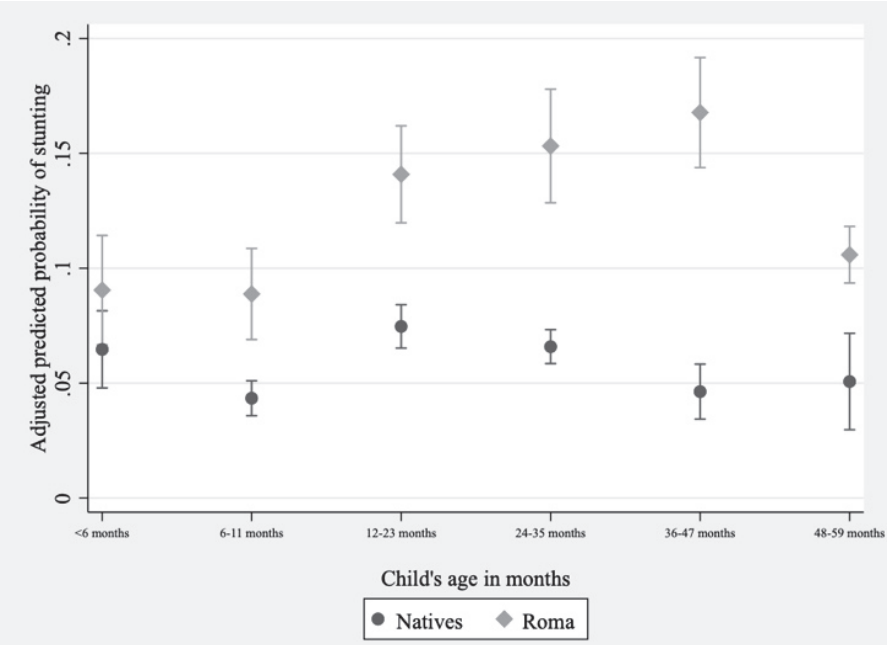
Despite most Roma families living in relative poverty, the wealth index shows a strong and significant effect. In particular, there is a clear risk gradient in the relationship between wealth index and stunting, with children in the top quintile

of the wealth index distribution displaying a lower risk of being stunted, compared with those at the bottom of the distribution,

As literature showed, undernutrition measures are strongly related with some socio-demographic characteristics of the mother, and with the quality of childcare. Having a higher-educated mother protects children from undernutrition both in the Roma and non-Roma community. Indeed, despite the low number of mothers who attended secondary school in the Roma samples, our models show that mother's education is a strong and significant protective factor for all the measures of child undernutrition. As expected, age at first union or marriage, age of woman at childbirth, and short birth interval all affect undernutrition (although some effects disappear once we include the interaction terms). The quality of childcare measures, too, displays a significant effect. The lack of breastfeeding and vaccination are associated with a higher risk of stunting, wasting and underweight. Similarly, being left alone or with other (older) children, as well as not having any children's books in the household, are both factors associated with one or more measures of undernutrition.

To understand whether the analyzed determinants play different roles among Roma and non-Roma children, we tested some interactions. The analysis reveals relevant differences in the role played by the age of the mother at birth and the type of place of residence among Roma and non-Roma children. Interestingly, we also find a significant, strong but opposite age gradient among Roma and non-Roma children (Figure 1). Specifically, while Roma children display a low probability of stunting during the first six months of life, comparable to the one displayed by non-Roma children, differences in the stunting risks among the two groups arise at age 6-11 months. Moreover, while non-Roma children's stunting risk starts decreasing during the second year of life, Roma children's risk of undernutrition steadily increases, and only slightly decreases between the fourth and the fifth birthday. It has been suggested that this result could be attributed to the lifestyle of nomads: in Roma settlements, children who are able to walk and thus exit the barracks are exposed to a greater risk of infections and consequent debilitation [5]. A further reason for these inequalities could be related to the different treatment that Roma mothers and their children are exposed to in health institutions, as well as to physical or linguistic barriers [8;9]: such factors could hinder Roma mothers from seeking medical help, making a delicate moment like the weaning phase an even more critical one.

Figure 1: *Adjusted predicted probabilities of stunting by age among Roma and non-Roma children (90% Confidence Interval)*



Source: Authors' elaborations on MICS data

References

1. Bhutta, Z.A. et al.: What works? Interventions for maternal and child undernutrition and survival. The Lancet. (2008) doi:10.1016/s0140-6736(07)61693-6
2. Bhutta, Z.A. et al.: Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost?. The Lancet. (2013) doi:10.1016/s0140-6736(13)60996-4
3. Boah, M. et al.: The epidemiology of undernutrition and its determinants in children under five years in Ghana. PLoS One. (2019) doi:10.1371/journal.pone.0219665
4. Efevbera, Y. et al.: Girl child marriage as a risk factor for early childhood development and stunting. Soc. Sci. and & Med. (2017) doi:10.1016/j.socscimed.2017.05.027
5. Geraci, S., Maisano, B., Motta, F.: Salute Zingara. Anterem, Roma (1998)
6. Harttgen, K., Klasen, S., Vollmer, S.: Economic Growth and Child Undernutrition in sub-Saharan Africa. Popul. Dev. Rev. (2013) doi:10.1111/j.1728-4457.2013.00609.x
7. Janevic, T., et al.: Risk factors for childhood malnutrition in Roma settlements in Serbia. BMC Public Health. (2010) doi:10.1186/1471-2458-10-509
8. Janevic, T., et al.: "There's no kind of respect here" A qualitative study of racism and access to maternal health care among Romani women in the Balkans. Int J Equity Health. (2011) doi:10.1186/1475-9276-10-53
9. Logar, M., Pavlič, D., Maksuti, A.: Standpoints of Roma women regarding reproductive health. BMC Womens Health. (2015) doi:10.1186/s12905-015-0195-0

10. Masibo, P.: Trends And Determinants Of Malnutrition Among Children Age 0-59 Months In Kenya (KDHS 1993, 1998, 2003 And 2008-09). ICF International. Calverton, Maryland, USA. (2013). https://dhsprogram.com/pubs/pdf/WP89/WP89.pdf. Cited 21 April 2022
11. UNICEF: Breaking The Cycle Of Exclusion: Roma Children In Southeast Europe. Belgrade. (2007)
12. UNICEF: Children, Food And Nutrition: Growing Well In A Changing World. New York. (2019)
13. United Nations: Transforming Our World: The 2030 Agenda for Sustainable Development. General Assembly Resolution. New York (2015)
14. Wachs, T.: Mechanisms linking parental education and stunting. The Lancet. (2008) doi:10.1016/s0140-6736(08)60144-0
15. Wang, S., et al. Changing trends of birth weight with maternal age: a cross-sectional study in Xi'an city of Northwestern China. BMC Pregnancy Childbirth. (2020) doi:10.1186/s12884-020-03445
16. WHO. World Health Organization Releases New Child Growth Standards. (2006) http://World Health Organization releases new Child Growth Standards. Cited 21April 2022
17. WHO. Fact sheets - Malnutrition. (2020) https://www.who.int/news-room/fact-sheets/detail/malnutrition. Accessed April 21, 2022
18. World Bank. Breaking The Cycle Of Roma Exclusion In The Western Balkans. World Bank, Washington DC (2019) https://documents1.worldbank.org/curated/en/642861552321695392/pdf/Breaking-the-Cycle-of-Roma-Exclusion-in-the-Western-Balkans.pdf. Cited 21 April 2022.

The manual, communicative and quantitative abilities of native and foreign workers according to their level of education in Italy

Le abilità manuali, comunicative e quantitative dei lavoratori nativi e stranieri in base al loro livello di istruzione in Italia

Camilla Pangallo[1], Oliviero Casacchia[2] e Corrado Polli[3]

Abstract Many studies address the labour market consequences of immigration flows. A branch of them argues that native workers respond to immigration by specializing in jobs where they have a comparative advantage, thus protecting themselves from labour market competition and possible wage losses. In order to examine this dynamic in the Italian labour market, we analysed the characteristics of jobs in Italy in terms of skills and tasks performed. The results show that foreigners hold more manual jobs, while natives hold jobs with more communicative and quantitative content.

Abstract Molti studi affrontano le conseguenze sul mercato del lavoro dei flussi migratori. Secondo alcuni i lavoratori nativi, per proteggersi dalla competizione sul mercato del lavoro e possibili perdite salariali, rispondono all'immigrazione specializzandosi in lavori dove hanno un vantaggio comparativo. Al fine di studiare questa dinamica sono state esaminate le caratteristiche dei lavori svolti in Italia in termini di abilità e compiti effettuati. Dai risultati emerge che gli stranieri svolgono professioni più manuali mentre i nativi svolgono professioni con maggiore contenuto comunicativo e quantitativo.

Key words: foreigner workers in Italy, specialization activities, level of education, abilities.

1 Camilla Pangallo, La Sapienza Rome University, Rome, Italy; camillapangallo@uniroma1.it

2 Oliviero Casacchia, La Sapienza Rome University, Rome, Italy; oliviero.casacchia@uniroma1.it

3 Corrado Polli, INAPP, Rome, Italy; c.polli@inapp.org

1 Background and Object

In recent decades, increasing migration flows have influenced labour market dynamics in most developed countries. Several studies have investigated the relationships between immigrant flows, workers' income and educational qualifications; Peri and Sparber (2011) argue that "…*when the proportion of foreign workers with high educational level increases in the U.S., native workers choose jobs with less analytical and more communicative content*". Zavodny (2014) emphasize that the jobs held by immigrants are more physically laborious than jobs held by U.S. natives. In other research, Peri and Sparber (2009) emphasize that "… *foreign-born workers specialize in occupations intensive in manual-physical labour skills while native pursue jobs more intensive in communication-language tasks*". Other authors [2] suggest that native workers respond to immigration by specialising in occupations requiring skills in which they have a comparative advantage. Few studies however have examined the differences between the foreign and native components in terms of requirements and characteristics of workers and work in Italy. In this paper, the Italian and foreign employed population in 2019 in Italy were analysed by detecting both specific information regarding the characteristics and requirements of the worker, their profession, and some socio-demographic data. The analyses refer to a period before the traumatic effect of the pandemic due to SARS-COV-2; for this reason, it does not reflect the impact of Covid-19 related changes on employment flows and outcomes.

2 Data and Methods

The data used for the analysis were those of the Survey on Italian Occupations (ISTAT and INAPP, 2013), provided by INAPP at a level of detail equal to 3 digits and the micro-data for public use of the Labour Force Survey–RCFL (ISTAT, 2019). The Survey on Italian Occupations (ICP) describes, with high analytical detail, all the professions existing in the Italian labour market, with particular reference to the content of the work performed and the organizational context in which the work takes place [3]. The ICP, borrowed from the Occupational Information Network (O*NET) survey, adopts the US conceptual model, the *content model*, which provided the basis for the creation of the contents of the questionnaire. The questionnaire is formulated by translating the one from the O*NET survey and, where necessary, adapting it to the Italian situation.

We have built an integrated database to simultaneously detect both specific information regarding the characteristics and requirements of the worker and the profession, and socio-demographic data of the workers, through statistical matching between the Sample Survey on Professions (INAPP, 2013) and the Labour Force Survey (ISTAT, 2019). We have also developed two methods for the construction of composite indicators, which allow us to analyze the world of professions in Italy in terms of workers' skills and work activities. The first methodology has referred to

the study by Peri and Sparber (2009) and the eight indicators (one Manual, two Communicative, three Quantitative, one Interactive and one Analytical) represent a weighted average value for each profession. The second method is the one proposed by Autor, Levy and Murnane (2003), and, in this case, the five indicators (Non-Routine Analytical Tasks, Non-Routine Interactive Tasks, Routine Cognitive Tasks, Routine Manual Tasks, Non-Routine Manual Tasks) were obtained using principal component analysis. The methodology developed by Autor, Levy and Murnane (2003) used the information present in the Dictionary of Occupational Titles (DOT), subsequently, Warman and Worswick (2015) incorporated the method proposed by the three authors to the data of the O*NET database as the latter has over time replaced the DOT; for this reason, in this work, the composite indicators were obtained with the variables suggested in Warman and Worswick's (2015) study.

Furthermore, in order to make the indicators constructed using the two methodologies described above comparable, composite indicators designed by Autor, Levy and Murnane (2003) were constructed, adopting the mathematical method used by Peri and Sparber (2009).

3 Results

The world of professions has been analyzed in terms of both the attitudes and skills of the worker and the activities and styles of work, according to sex, level of education, and citizenship.

This paragraph presents some of the main results relating to the indicators obtained by applying the methodology of Peri and Sparber (2009). In this case, the "Manual" and "Communicative" indicators are considered because they represent the most interesting results and for which the greatest differences have emerged between Italian and foreign workers. The comparison is between the Italian and foreign employed populations from developing countries; the employed of developed countries were excluded from the study as the results obtained are similar to those of Italian employed people. The analysis was carried out separately by sex and level of education: the workers without a school certificate or in possession of an elementary or middle school certificate are included in the low qualification group, while the workers with a university degree were in the high qualification group. Graduated workers have been excluded in this part of the analysis to highlight the differences between less-educated and highly educated workers.

As shown in Table 1 for employed men, the Italians with low educational qualifications engage in activities with a moderate manual content (the score is 27,71), which is higher than that observed for Chinese (22,81) and those employed in Latin America and Oceania (27,47), but lower than all the others.

Table 1: *Employed men: manual and communicative indicators by educational level and citizenship. Years 2019 (RCFL) and 2013 (ICP).*

Citizenship groups	Manual indicator		Communicative indicator		Relative difference between low and high levels of education[1]	
	Low level of education	High level of education	Low level of education	High level of education	Manual indicator	Communi-cative indicator
Italy	27,71	10,47	57,63	68,83	90,3%	17,7%
Developing countries:						
EU	33,46	23,11	45,04	53,59	36,6%	17,3%
Not- EU Eastern Europe	34,30	23,65	43,80	51,73	36,8%	16,6%
Asia (excluding China)	27,78	16,66	43,04	58,05	50,0%	29,7%
China	22,81	16,93	50,05	61,26	29,6%	20,1%
North Africa	31,65	21,50	44,63	51,16	38,2%	13,6%
Sub-Saharan Africa	28,95	17,56	44,65	56,30	49,0%	23,1%
Latin America and Oceania	27,47	20,84	43,94	54,52	27,5%	21,5%

Own elaboration on ISTAT-INAPP data.

Italian workers with high educational qualifications carry out activities - as expected - with a minimum manual content (10,47). Furthermore, it is observed that Italians with both low (57,63) and high (68,83) educational qualifications carry out more communicative occupations than all other groups of foreigners. The relative differences by educational qualification in exercising activities characterized by high manual skills are high in the case of Italians (90,3%), much more moderate for the other groups. Conversely, the differences in the degree of education when it comes to activities characterized by high levels of communicativeness are more limited for Italians (17,7%), and the same is true for the other groups of foreign workers, except Asians (almost 30%).

Regarding employed women, Italian women with low educational qualifications as well as foreign women work in occupations with a moderate manual content. The same result was found for men, although the differences by citizenship are even smaller in the case of women. Italian women workers with high qualifications also work in occupations with low manual content, showing lower values than all other groups of foreign workers; again, the result is very similar to that seen for men.

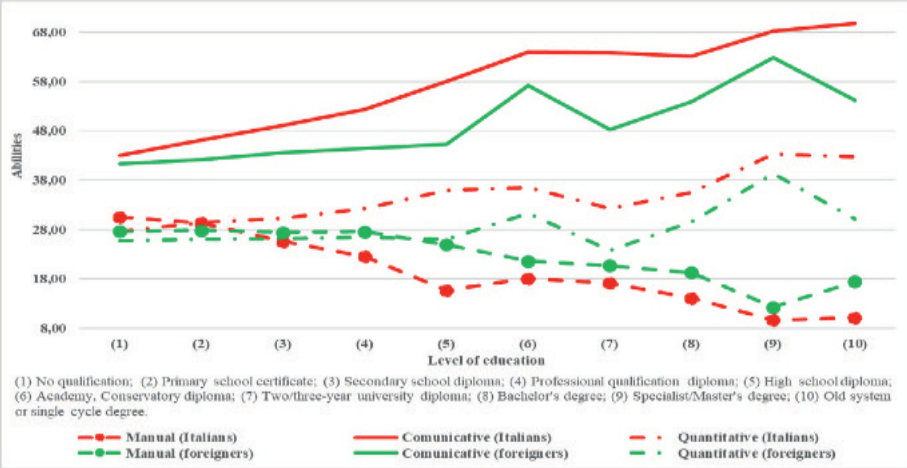
The second type of indicators (Autor, Levy and Murnane) also show greater differences between Italian and foreign workers in terms of content of manual and non-manual occupations, and the study was analysed further by adopting Peri and Sparber's Manual, Communicative and Quantitative indicators.

[1] Relative difference between low and high levels of education relative are calculated considering the average between the two values as denominator.

Data from the Survey on Italian Occupations (INAPP) confirms how the use of the attitudes and skills contained in the various occupations varies according to the level of education of employees; it is possible, using this database linked to the Labour Force Survey, to also test the hypothesis of a variation in this use according to citizenship.

Figure 1 shows the value of manual, communication, and quantitative skills by educational qualification for Italian and foreign workers. Communication and quantitative skills tend to increase with higher educational qualifications, while the use of manual skills decreases.

Figure 1: *Average abilities of Italian and foreign employed people in Italy by the level of education. Years 2019 (RCFL) and 2013 (ICP).*



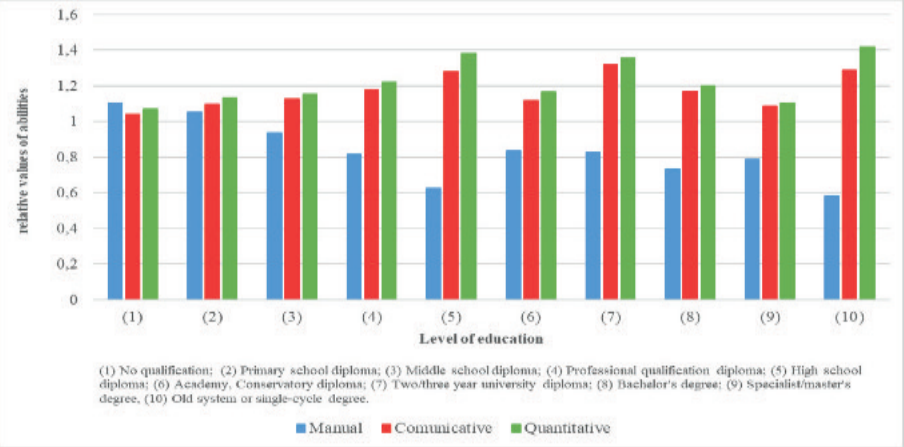
Own elaboration on ISTAT-INAPP data.

Employed people with high educational qualifications are unsurprisingly employed in jobs with high communication content. In some cases, however, there is also a high level of manual work: for example, workers in professions such as health technicians have a manual indicator value of 21,12 (6th decile) and a communication indicator value of 63,56 (7th decile) and doctors have a manual indicator value of 18,69 (5th decile) and a communication indicator value of 73,43 (10th decile). The professions with greater communication skills held by workers with university degrees are, for example, university lecturers (86,92) and researchers and technicians with university degrees (84,52). Educated workers in highly quantitative professions are Engineers and related professions (62,70) and quantitative, physical, and chemical science technicians (63,07). Native workers show an advantage in communication and quantitative skills, foreign workers an advantage in manual work.

Absolute skill levels have to be converted into relative values in order to obtain the comparative advantage in terms of skills of native workers over foreigners (Figure 2). The intensity of communicative and quantitative skills is higher for Italians at practically all levels of education: Italians show a systematic comparative advantage in communicative and quantitative skills compared to manual skills. Italians also show even higher levels of quantitative skills than communicative

skills, which is an unexpected result. Immigrants present a comparative advantage in manual work compared to other skills at all levels of education with the exception of the 'no qualification' category.

Figure 2: *Ratio between the average abilities of Italian and foreign workers. Years 2019 (RCFL) and 2013 (ICP).*



Own elaboration on ISTAT-INAPP data.

4 Conclusion

Thus, the results obtained in this work would seem to confirm part of the conclusions of Peri and Sparber's studies: in Italy, foreign workers would be complementary to native workers because they specialize in different tasks having different skills.

References

1. Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4), 1279-1333.
2. Gu, E., & Sparber, C. (2017). The native-born occupational skill response to immigration within education and experience cells. Eastern Economic Journal, 43(3), 426-450.
3. INAPP, Nota metodologica Indagine campionaria sulle professioni, 2013.
4. Peri, G., & Sparber, C. (2009). Task specialization, immigration, and wages. American Economic Journal: Applied Economics, 1(3), 135-69.
5. Peri, G., & Sparber, C. (2011). Highly educated immigrants and native occupational choice. Industrial Relations: a journal of economy and society, 50(3), 385-411.
6. Warman, C., & Worswick, C. (2015). Technological change, occupational tasks and declining immigrant outcomes: Implications for earnings and income inequality in Canada. Canadian Journal of Economics/Revue canadienne d'économique, 48(2), 736-772.
7. Zavodny, M., (2014). Do Immigrants Work in Worse Jobs than U.S. Natives? Evidence from California, IZA Discussion Papers, No. 8327, Institute for the Study of Labor (IZA), Bonn.

HIV Prevalence in some African Territories: Socio-Economic Drivers

La Prevalenza di HIV in alcuni Territori Africani: Fattori Socio-Economici

Micaela Arcaio,[1*] Daria Mendola,[1] Anna Maria Parroco[1]

Abstract In 2020, 35% of all HIV-positive people in the world lived in Eastern and Southern Africa. This work aims at assessing the relationship between socio-economic drivers and HIV prevalence at the sub-national level in these countries. The data used are drawn from the Demographic and Health Survey, in which a subset of respondents is tested for HIV. Using a fractional logistic regression model on clusters of individuals, middle-to-high wealth is positively associated with higher HIV prevalence, while a higher average number of children acts as a protective factor; moreover, higher proportions of people who have never been in sexual relationships lower their cluster's HIV prevalence but HIV-positive people are more likely to use condoms at last intercourse.

Abstract *Nel 2020 il 35% delle persone sieropositive nel mondo era concentrato in Africa del Sud e dell'Est. Questo studio mira a valutare la relazione tra fattori socio-economici e prevalenza di HIV a livello sub-nazionale in questi paesi. I dati derivano dalla Demographic and Health Survey, in cui un sotto-campione viene sottoposto al test dell'HIV. Utilizzando un modello di regressione logistica frazionaria su cluster di individui, il benessere medio-alto è associato positivamente a prevalenza di HIV più alta, mentre un numero medio di figli maggiore agisce da fattore protettivo; inoltre, maggiori proporzioni di persone che non hanno mai avuto relazioni sessuali abbassano la prevalenza del loro cluster ma persone sieropositive utilizzano con maggiore probabilità il profilattico.*

Key words: HIV, socio-economic status; fractional logistic regression; sexual behaviours.

[1] Micaela Arcaio, University of Palermo; email: micaela.arcaio@unipa.it;
Daria Mendola, University of Palermo; email: daria.mendola@unipa.it;
Anna Maria Parroco, University of Palermo; email: annamaria.parroco@unipa.it
*Corresponding author

1 Introduction

In 2020 an estimated 38 million people were living with HIV in the world. Among these people, about 50% are adult women (age 15+) and 5% are children aged 0-14 [1]. When it comes to HIV prevalence, it is estimated to be globally at around 0.7% among adults aged 15-49, with varying prevalence between both regions and countries. Indeed, the regional prevalence ranges between less than 0.1% in North Africa and the Middle East and 6.5% in Eastern and Southern Africa.

As for socio-economic drivers of the HIV epidemic, there is no evidence for an unequivocal relationship between some socio-economic characteristics of a certain population and HIV infection. Different authors have found both positive and negative association with education and wealth, with employment status and age, highlighting the heterogeneity in the relationship [3]. Indeed, some studies have found that higher standards of living, as well as higher household wealth, are positively associated with higher odds of HIV infection [4, 5, 6], while others have found that poorer people incur to income-generating strategies that put them at risk of seropositivity [7]. Education has also been shown to reduce vulnerability to HIV [6].

Focusing on the sexual transmission of HIV, this work aims to assess the relationship between socio-economic drivers and HIV prevalence at the sub-national level.

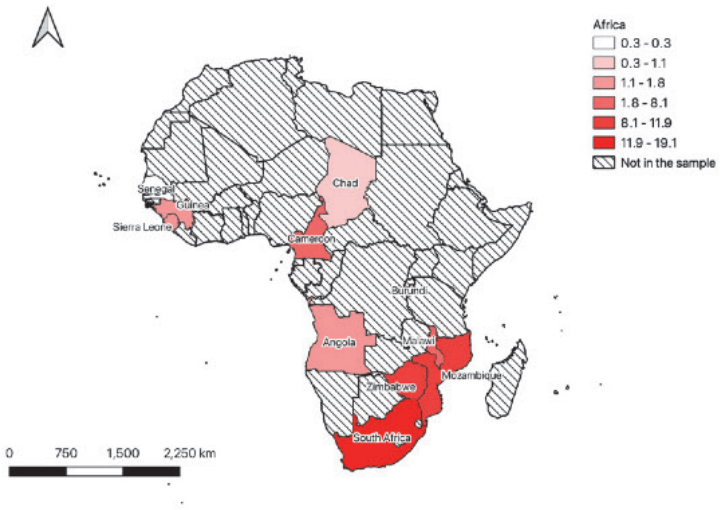
2 Data and Methods

The data used in this analysis are drawn from the Demographic and Health Survey (DHS). It is a nationally representative household survey that has covered over 90 countries in more than 350 surveys in the last forty years [8]. Standardized questionnaires are employed to improve comparisons in time and space on fertility, mortality, family planning, reproductive health, and HIV/AIDS – sub-samples are subjected to ELISA (enzyme-linked immunosorbent assay) tests to determine respondents' serostatus.

In this study, data are from 11 African countries: Angola, Burundi, Cameroon, Chad, Guinea, Malawi, Mozambique, Senegal, Sierra Leone, South Africa, and Zimbabwe. These were selected due to the availability of surveys in a period ranging from 2015 to 2019, choosing one survey (the most recent) for each country.

Indeed, despite accounting only for 3.21% of the global population in 2020 [2], these countries reached a total of 13.5 million cases in that year, accounting for almost 36% of all global HIV-positive patients [1]. The HIV prevalence in these countries ranges from 1% in Burundi to 19.1% in South Africa (see Figure 1).

Figure 1: *HIV Prevalence in the chosen countries*



Our sample is made of almost 150,000 people, aged 15-64, belonging to 11 countries. In the original dataset, households that participated in the survey are grouped in the so-called "clusters"; they are 5,960 with size from 2 to 113 individuals. The original DHS clusters were combined into purposive territorial areas with a greater demographic size to keep the analysis at a territorial level. A spatially-aware k-means clustering method was employed to group the original clusters. Here, the mean point is represented by the barycenter of the clustered geographical coordinates. The procedure [9] was iterated on QGIS for all countries in the dataset, so that the number of macro-clusters would amount to about 50 per country, with a mean population of 270 individuals. The final sample is made of 549 macro-clusters. All variables are aggregated at the macro-cluster level to check for macro socio-economic drivers of HIV.

HIV prevalence at the macro-cluster level is the response variable of the model in this study. It ranges from 0-3.73% in Burundi to 0.19-35.43% in Mozambique, with a total average 5.41% (SD = 7.10).

All variables in the model – including the response – are either presented as means (for continuous variables) or as proportions (for dichotomous variables). Given the nature of the phenomenon, and the way the response variable was built, a fractional logistic regression model was implemented to evaluate the effect of the selected covariates [10]. The log-likelihood function for this model is:

$$\ln L = \sum_{j=1}^{N} y_j \ln\left(\frac{\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}\right) + (1-y_j)\ln\left(1-\frac{\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}\right) \quad (1)$$

in which N stands for the number of territorial units, y_j is the local HIV prevalence, and $\mathbf{x}_j$ is the set of regressors for each territorial unit j.

The option "clustered standard errors" was used so to account for macro-clusters being nested in countries.

Regressors were classified into three different domains: 1) *Socio-demographic characteristics of the macro-cluster:* average age; average no. of years of education; proportion of middle-to-high wealth people; proportion of workers; 2) *Sexual behaviours:* proportion of people who have never had sexual relationships; proportion of married people; proportion of people who used a condom during their last intercourse; 3) *Context variables:* proportion of respondents in a rural setting; country dummies. Country dummies also account for regional specificity.

3 Results

Table 1 shows odds ratio estimates for the model in equation (1). Among socio-demographic characteristics of the macro-clusters, the average number of children is a particularly interesting characteristic, with more children on average lowering the odds for respondents' to be HIV positive – probably due to women having access to prevention education during antenatal care. While the model does not show any significant association between HIV and age at the territorial level, it shows that more educated areas are less likely to be affected by HIV. However, as expected, wealthy areas are twice as likely to have a higher prevalence of HIV.

As for sexual behaviours, a higher proportion in the area of people who have never been in sexual relationships lowers their cluster's HIV prevalence (83% less likely), and a higher prevalence of HIV-positive people is associated with higher proportions of people using condoms at last intercourse. On the other hand, being married is not statistically significant at the territorial level in explaining HIV prevalence.

Rural areas are not statistically different from urban ones in terms of HIV prevalence, while differences among countries emerge with, other things being equal, Burundi, Senegal, and Sierra Leone showing lower levels of HIV prevalence concerning Angola (assumed as a reference country).

Table 1: Fractional logit model for HIV prevalence (odds ratio estimates)

Domains	*Covariates*	*OR estimates*
Socio-demographic characteristics	Age (avg.)	1.001
	No. of years of education (avg.)	0.884***
	No. of children (avg.)	0.786***
	Middle-to-high wealth (prop.)	1.872**
	Workers (prop.)	0.507
Sexual behav-iours	Never in a sexual relationship (prop.)	0.172***
	Married people (prop.)	0.216
	Used condom at last intercourse (prop.)	3.665**
Context	Rural area	0.898
	Constant	0.289*
	Observations	549

*** p < 0.01, ** p < 0.05, * p < 0.10 *Note: controlled for country dummies*

4 Conclusions

This study is focused on eleven African countries, which, despite only accounting for 3% of the global population, represented more than 35% of all global HIV cases. This analysis aimed to assess the relationship between socio-economic drivers and HIV prevalence at the sub-national level, focusing exclusively on sexual transmission and the socio-economic background of the territorial areas.

Literature on determinants of HIV prevalence has found discordant results. Using a fractional logit model on aggregated data, our findings showed that while age and employment rates do not seem to have an effect on a macro level, a highly educated population is less likely to be HIV positive, but richer people are more likely to drive forward this epidemic.

This mechanism may be due to social networking: richer people have higher chances of having more encounters, which in turn increases their odds of being HIV positive. Social relationships and conventions may also explain why higher proportions of people using condoms during their last intercourse are highly associated with people being HIV positive: people aware of their serostatus are more conscious and, thus, try to avoid infecting their partners by using correct preventative methods.

There is no doubt that these effects need to be further investigated. The erratic way the socio-economic status affects HIV is a sign that the relationship between these two dimensions depends not only on the territories where these data are collected but also on how each population changes over time.

This study is not free from limitations, among these: the arbitrariness of aggregation of territorial clusters; a poor structure of context-level variables. Further steps include robustness tests to different spatial aggregation strategies and a deep reflection on such methodological and substantive aspects.

References

1. AIDSinfo | UNAIDS. Aidsinfo.unaids.org. https://aidsinfo.unaids.org. Published 2022. Cited 27 Feb 2022.
2. World Population Prospects - Population Division - United Nations. Population.un.org. https://population.un.org/wpp/. Published 2022. Cited 27 Feb 2022.
3. Fortson J. The gradient in sub-Saharan Africa: Socioeconomic status and HIV/AIDS. Demography. (2008) doi:10.1353/dem.0.0006
4. Msisha W, Kapiga S, Earls F, Subramanian S. Socioeconomic status and HIV seroprevalence in Tanzania: a counterintuitive relationship. Int J Epidemiol. (2008) doi:10.1093/ije/dyn186
5. Igulot P, Magadi M. Socioeconomic Status and Vulnerability to HIV Infection in Uganda: Evidence from Multilevel Modelling of AIDS Indicator Survey Data. AIDS Res Treat. (2018) doi:10.1155/2018/7812146
6. Seeley J, Malamba S, Nunn A, Mulder D, Kengeya-Kayondo J, Barton T. Socioeconomic Status, Gender, and Risk of HIV-1 Infection in a Rural Community in South West Uganda. Med Anthropol Q. (1994) doi:10.1525/maq.1994.8.1.02a00060
7. Bunyasi E, Coetzee D. Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa. BMJ Open. (2017) doi:10.1136/bmjopen-2017-016232
8. Croft T, M.J. Marshall A, Allen C. Guide to DHS Statistics. Calverton, Maryland, USA: Macro International Inc.; (2018)
9. K-means clustering. docs.qgis.org. https://docs.qgis.org/3.16/en/docs/user_manual/processing_algs/qgis/vectoranalysis.html#k-means-clustering. Published 2022. Cited 28 Feb 2022.
10. Fractional regression. Stata.com. https://www.stata.com/manuals/rfracreg.pdf. Published 2022. Cited 28 Feb 2022.

A longitudinal cross country comparison of migrant integration policies via Mixture of Matrix-Normals

Un confronto longitudinale tra paesi sulle politiche di integrazione degli immigrati attraverso Mixture of Matrix-Normals

Leonardo Salvatore Alaimo, Francesco Amato and Emiliano Seri

Abstract In recent decades, there has been a growing research interest in comparative studies of migrant integration, assimilation and the evaluation of policies implemented for these purposes. With this aim, the Migrant Integration Policy Index (MIPEX) measures policies to integrate migrants in 52 countries all over the world, over time. However, the comparison of very different countries on complex and multidimensional phenomena can lead to misleading interpretations and evaluations of the results. In this paper, we improve and facilitate the comparison between the treated countries on 7 MIPEX dimensions, applying a Mixture of Matrix-Normals classification model for longitudinal data. Trough the analysis, 5 clusters of countries have been discovered, allowing us to add new levels of interpretation of the data.

Abstract *Negli ultimi decenni, c'è stato un crescente interesse di ricerca negli studi comparativi sull'integrazione dei migranti, l'assimilazione e la valutazione delle politiche attuate per questi scopi. Con questo obiettivo, il Migrant Integration Policy Index (MIPEX) misura le politiche di integrazione dei migranti in 52 paesi di tutto il mondo nel tempo. Tuttavia, il confronto tra paesi molto diversi su fenomeni complessi e multidimensionali può portare a interpretazioni e valutazioni fuorvianti dei risultati. In questo articolo miglioriamo e facilitiamo il confronto tra i paesi trattati su 7 dimensioni MIPEX, applicando un modello di classificazione Mixture of Matrix-Normals per dati longitudinali. Attraverso l'analisi sono stati individuati 5 cluster di paesi, permettendoci di aggiungere nuovi livelli di interpretazione dei dati.*

Leonardo Salvatore Alaimo
Department of Social Sciences and Economics, Sapienza University of Rome, e-mail: leonardo.alaimo@uniroma1.it

Francesco Amato
ERIC EA 3083, Université de Lyon, Lyon 2, e-mail: francesco.amato@univ-lyon2.fr

Emiliano Seri
Department of Statistic, Sapienza University of Rome, e-mail: emiliano.seri@uniroma1.it

Key words: Mixture of matrix-normals, MIPEX, Model-based classification, Migration policies

1 Introduction

Immigration regulation and immigrant assimilation have been a salient political issue in all industrialised countries for many decades [1]. The growing interest in comparative analyses of immigration has led to a variety of attempts to quantify immigration policies, i.e. to assess and put into numerical form what countries are doing to foster the integration and assimilation of immigrants. However, the study of these phenomena from a quantitative point of view is rather recent, due to the previous lack of data. Moreover, quantifying migrant integration is a difficult challenge, due to its complex nature and lack of uniformity in migration policies of many countries, which is based on multiple criteria. In the present work, we focus on the Migrant Integration Policy Index (MIPEX) [1]. The project informs and engages key policy actors about how to use indicators to improve integration governance and policy effectiveness. Its aim is to measure policies that promote integration in both social and civic terms, evaluating trough a survey the migration policies of each considered country, to construct a multi indicator system, first aggregated in 8 dimensions, each aggregated in one single composite indicator. The aim of this paper is to add new perspectives on the MIPEX data while respecting the complexity of the phenomenon under consideration by discovering structures and patterns in the behaviour of the considered countries. The research question from which this paper starts is:

- *Given the complexity of the phenomenon under consideration, in order to improve the comparison between the surveyed countries, is it possible to identify homogeneous groups over time among them, i.e. groups of countries which behave similarly across and within time?*

To answer this question, a *Finite Mixture of Matrix-Normals model* has been applied to cluster the units, taking into account the time dimension. The MIPEX includes 52 countries and collects data from 2007 to 2019, in order to provide a view of integration policies across a broad range of differing environments. It considers a system of 58 indicators (for more information, please consult [1]) covering 8 policy areas that have been designed to benchmark current laws and policies against the highest standards through consultations with top scholars and institutions[1]. The policy areas of integration covered by the MIPEX are the following:

[1] The highest standards are drawn from Council of Europe Conventions, European Union Directives and international conventions (for more information see: http://mipex.eu/methodology)

- Labour Market Mobility
- Family Reunion
- Education
- Political Participation
- Long-term Residence
- Access to Nationality
- Anti-discrimination
- Health[2]

Each dimensional synthetic indicator is bounded between $[0,100]$, in which the maximum of 100 is awarded when policies meet the highest standards for equal treatment. These values are chosen by experts from each country, by means of a questionnaire. The analysis carried out in the present work uses the listed above dimensions[3] excluding health.

2 Mixture of Matrix-Normals

Finite Mixture of Matrix-Normals (MNN), as introduced in [2], can be a useful tool to cluster time-dependent data. Let $\mathbf{Y} = \{Y_i\}_{i=1}^{N}$ be a sample of $J \times T$-variate matrix observations (i.e. $Y_i \in R^{J\times T}$), arose from studies with J-variate vector observations measured repeatedly over T time points, as in a longitudinal study case[4]. Assume that each Y_i follows a matrix-normal distribution, $Y_i \sim \mathscr{MN}_{(J\times T)}(M,\Phi,\Omega)$,where $M \in R^{J\times T}$ is the matrix of means, $\Phi \in R^{T\times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Omega \in R^{J\times J}$ is the covariance matrix containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is given by

$$f(Y \mid M,\Phi,\Omega) = (2\pi)^{-\frac{TJ}{2}} \mid \Phi \mid^{-\frac{J}{2}} \mid \Omega \mid^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}[\Omega^{-1}(Y-M)\Phi^{-1}(Y-M)]\right\} \tag{1}$$

The matrix-normal distribution is a natural extension of the multivariate normal distribution, since if $Y \sim \mathscr{MN}_{(J\times T)}(M,\Phi,\Omega)$, then $\mathrm{vec}(Y) \sim \mathscr{MVN}_{TJ}(\mathrm{vec}(M),\Phi\otimes\Omega)$, where $\mathrm{vec}(.)$ is the vectorization operator and $\otimes$ denotes the Kronecker product. Being a special case of the multivariate normal distribution, the matrix-normal distribution shares the same various properties, like, for instance, closure under marginalization, conditioning and linear transformations [3]. The pdf of the MMN model is given by

$$f(Y \mid \pi,\Theta) = \sum_{k=1}^{K} \pi_k \phi^{(J\times T)}(Y \mid M_k,\Phi_k,\Omega_k) \tag{2}$$

[2] Health data are only available for years 2014 and 2019, therefore this dimension could not be used in the analysis

[3] An extensive explanation of the MIPEX dimensions is given in [1]

[4] The three-way data time arrays analysed are represented as: $\mathbf{Y} \equiv \{y_{ijt} : i = 1,\ldots,N; j = 1,\ldots,J; t = 1,\ldots,T\}$, where $i = 1,2,\ldots,52$ indicates the generic country, $j = 1,2,\ldots,7$ the generic MIPEX dimension and $t = 2014,2015,\ldots,2019$ the generic year.

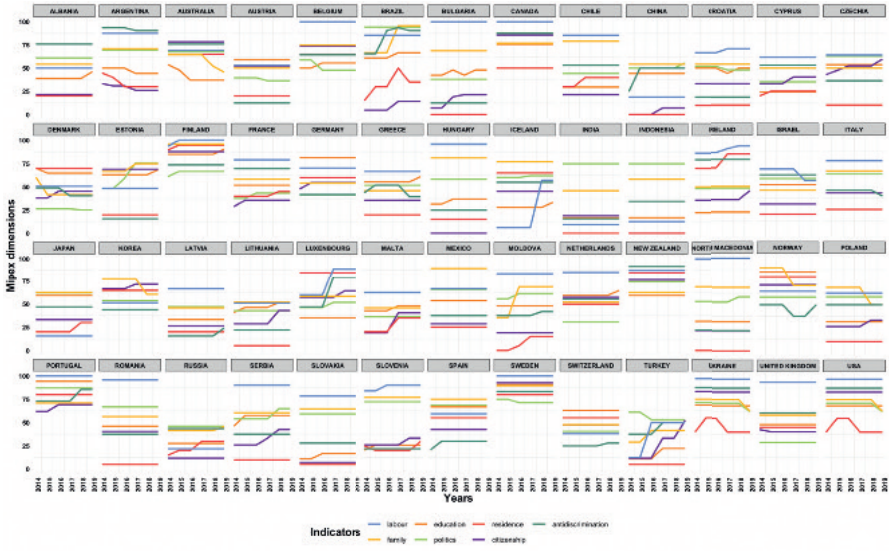
where K is the number of mixture components, $\pi = \{\pi_k\}_{k=1}^{K}$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^{K} \pi_k = 1$ and $\Theta = \{\Theta_k\}_{k=1}^{K}$ is the set of component-specific parameters with $\Theta_k = \{M_k, \Phi_k, \Omega_k\}$.

In [4], carrying forward from the previous papers, the over-parametrization issue is addressed. For identifiability issues of the model, the determinant of the time-covariance matrix must be restricted to be $\mid \Phi_k \mid = 1$.

3 Results

Figure 1 outline that most of the countries does not change much the values of their indicators through time. Following, the MNN will be used to model together the changes between and within time, grouping together the units which behave similarly across and within time[5].

Fig. 1 Country trajectories of the 7 MIPEX dimensions over time

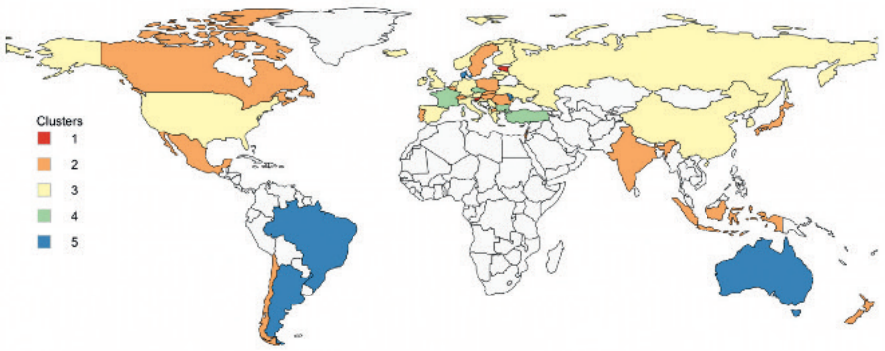


Since our dataset is composed by 52 units, we carried out the MMN model for K ranging from 1 to 8, and choose the best number of clusters by means of the BIC. The selected K is 5. According to BIC, the selected parametrization of the model is A-VEV-VV, which means that the means are better parsimoniously parametrized in additive way, Ω_k with varying volume, equal shape and varying orientation and

[5] The analyzes have been carried out using the package of the softaware R, MatTransMix [4]

Φ_k has both varying shape and orientation. Because of the matrices Φ_k Ω_k, each MMN component models not only the conditional means, but also covariances of the response variables and the covariances among times. In this way, each cluster provides a broad profile of units belonging to it. We report graphically the countries that belongs to each cluster in Figure 2 and the interpretation of the results is as follows:

Fig. 2 MIPEX dimensional indices: MMN clusters' composition of MIPEX countries. Years 2014 – 2019



- **Cluster 1:** *Estonia and Slovenia*. This cluster presents lower correlations in time between the first three years (2014-2016) and the second ones (2017-2019). Moreover, it has negative correlation between labour and the other dimensions, except for family reunification policies. Countries in this cluster have the lowest score for the access to citizenship and rank low for political participation as well, while ranking average for labour mobility, educational policies and high for family reunification, long-term residence and anti-discrimination legislation.
- **Cluster 2:** *Belgium, Canada, Chile, Hungary, India, Indonesia, Israel, Japan, Mexico, New Zealand, North Macedonia, Poland, Portugal, Romania, Slovakia, Sweden, Switzerland*. During the study period, countries belonging to this cluster did not change much their policies, and countries that rank high in some areas tend to rank high in the others as well. The countries of this group tend to have good policies for long-term residency, family reunification and anti-discrimination, but rank low for education and political participation.
- **Cluster 3:** *Albania, Austria, China, Croatia, Cyprus, Finland, Germany, Greece, Iceland, Ireland, Italy, Korea, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Russia, Serbia, Spain, Ukraine, UK, USA*. Countries in this cluster represents the group that reformed less their immigration legislation during the study period. They tend to rank average in most of the policies areas, with the exception of residence and anti-discrimination laws, where they tend to rank

higher. However, low correlation among variables signals that countries do not move homogeneously among the policies areas.

- **Cluster 4:** *Bulgaria, Czech Republic, France, Turkey*. Despite ranking generally high for anti-discrimination policies, countries within this cluster tend to rank low for policies in education, access to citizenship and labour market mobility, while scoring average for long-residence legislation. Yet, low correlation among variables indicates that countries do not move homogeneously among the dimensions, with the exception of policies regarding access to long-term residence and anti-discrimination, that have high positive correlation. These countries have seen their score moderately changing in time, indicating that some changes in the legislation have happened.
- **Cluster 5:** *Argentina, Australia, Brazil, Denmark, Moldova*. Countries belonging to this cluster have high-correlation in time, but they tend to decrease faster with time, meaning that some changes in the policies have been made especially in the last years. They generally rank low in policies related to educational support for foreign pupils and political participation, average for legislation related to access to labour market and high in family reunion, residence, access to citizenship and anti-discrimination. The policies' dimensions have low correlation, meaning that the countries tend not to move homogeneously among them.

4 Conclusions

This paper has explored immigrant integration policies, analyzing 7 dimensions of the MIPEX from the year 2014 to 2019, to identify groups of units with similar behaviour, to improve the ease of reading of the phenomenon. We addressed this issue trough the application of an unsupervised approach to clustering for longitudinal data namely Mixture of Matrix-Normals model, that accounts simultaneously for the within and between time dependency structures. The analysis, allowed the comparison of clusters with each other and of the countries within each cluster. Also, the correlations in time shown the general trend of each indicator over time in each cluster, and the correlations between variables purified from time effect shown the behaviour of each indicator in relation to the others within each cluster.

References

1. Giacomo Solano and Thomas Huddleston. Migrant integration policy index 2020. *Barcelona Center for International Affairs (CIDOB)*, 2020.
2. Cinzia Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4):511–522, 2011.
3. A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.
4. Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling. *Journal of Classification*, pages 1–24, 2021.

Education and job placement

Measuring happiness at work with categorical Principal Component Analysis

Misurazione della felicità aziendale attraverso l'Analisi delle Componenti Principali categorica

Kocollari, U., Cavicchioli, M., Demaria, F.

Abstract Despite the large number of contributions on happiness at work, accurate measures are still missing. In this realm, our study investigates the influence of happiness drivers in the workplace, considering the core elements related to both hedonia and eudaimonia, on the most commonly used outcome variable in the management literature that analyses happiness, i.e., the sum of positive feelings. We first explored the drivers of happiness using categorical PCA to find the latent dimensions able to summarize original variability. Then, we analyzed their configuration within a binary regression framework to identify their relative contribution in predicting the probability of employee happiness.

Abstract *Nonostante l'ammontare di contributi sulla felicità aziendale, mancano ancora misure accurate. In questo ambito, il nostro studio indaga l'influenza dei driver di felicità sul posto di lavoro, considerando gli elementi fondamentali relativi sia all'edonia che all'eudaimonia, sulla variabile più comunemente usata nella letteratura manageriale che analizza la felicità, cioè la somma dei sentimenti positivi. Abbiamo prima esplorato i driver della felicità utilizzando la PCA categorica per trovare le dimensioni latenti in grado di riassumere la variabilità originale. Poi, abbiamo analizzato la loro configurazione in un framework di regressione binaria per identificare il loro contributo relativo nel predire la probabilità di felicità dei dipendenti.*

Key words: happiness at work, categorical PCA, optimal scaling, employee well-being.

[1] Kocollari, U., University of Modena and Reggio Emilia & Softech-ICT, Department of Economics "Marco Biagi"; email: ulpiana.kocollari@unimore.it

[2] Cavicchioli, M., University of Modena and Reggio Emilia & ReCent, Department of Economics "Marco Biagi"; email: maddalena.cavicchioli@unimore.it

[3] Demaria, F., "Marco Biagi" Foundation, University of Modena and Reggio Emilia; email: fabio.demaria@unimore.it

1 Introduction

Even though happiness in the workplace has been considered critical in management and organizational research, accurate measures for happiness at work are still missing (Fisher, 2010, Salas-Vallina et al., 2020).

In general terms happiness is a subjective experience, closely related to psychological well-being (Grant et al., 2007), which consists of the eudaimonic aspect of happiness (Ryff, 1989). Such well-being depends upon both person-level aspects (e.g. values, attitude, goals) and organization-level characteristics like work environment and culture (De Sousa & Porto, 2015). Moreover, Fisher (2010) identifies three main measures for employee happiness: engagement, job satisfaction and affective organizational commitment. Thus, the happiness construct does not have a single agreed-upon definition, requiring an exploratory and multidimensional approach able to capture the nuances of the term.

1.1 Literature review

The relevance of happiness at the workplace has been examined through different concepts. Research has conceptualized happiness as a multidimensional construct (Ryan & Deci, 2001), considering two main dimensions: hedonic happiness (hedonia) and eudaimonic happiness (eudaimonia). Many authors focus on the hedonic experiences of pleasure and liking, and/or positive beliefs about a certain object i.e. affective commitment, the experience of positive emotions while working (Gaston-Breton et al., 2021). Other investigations include both hedonic and eudaimonic content, the latter involving learning/development, growth, autonomy, and self-actualization (Salas-Vallina et al., 2020).

Fisher (2010) reviewed the concept of 'happiness at work', including both hedonic and eudaimonic orientations. In our study we incorporate both hedonia and eudaimonia, seeing happiness as the sum of positive feelings (Kahneman et al, 1999; Wright & Cropanzano, 2004). We first investigated the drivers of happiness in organizations using categorical PCA to explore the latent dimensions of happiness. Then, we analyzed their configuration within a Logistic regression model that answers to our research question: which are the drivers of happiness in organizations that can positively influence the level of employee happiness?

2 Methodology

Principal Component Analysis, also known by the acronym PCA, is a dimension reduction technique which aims to reduce the number of variables 'm' to a number of 'p' uncorrelated principal components. This technique, however, assumes only linear relationships between variables, requiring a numeric level of measurement (interval or ratio) for all the variables analyzed. In nonlinear PCA all variables are handled as categorical and we perform optimal scaling to assign a numeric value to each of the category label, in a way that maximizes the variance in the quantified variables (Linting & van der Kooij, 2012). In the following we will provide the fundamental aspects of nonlinear PCA. For an extensive review of the procedure and its historical developments we refer to Gifi (1990).

Let **H** be the *n*x*m* observed data matrix, where 'm' is the number of ordinal variables measured on 'n' individuals (objects). Let $\mathbf{h}_j$ be a categorical variable of the matrix **H**, with $j = 1,\ldots, m$, and **G** be an indicator matrix of order nxk_j, where $\mathbf{k}_j$ represents the number of possible categories of variable j. Categorical variables require nonlinear (optimal) scaling, which assigns category quantifications to $\mathbf{h}_j$ through the nonlinear function $\mathbf{q}_j = \varphi_j(\mathbf{h}_j)$. In the following we assume a weight of 1 for all the variables. $\mathbf{Y}_j$ is the category quantification matrix (k_jxp) of variable j, $\mathbf{y}_j$ denotes the vector (of $\mathbf{k}_j$ order) of category quantifications, and $\mathbf{q}_j$ is the associated transformed variable. Thus, the transformed variable $\mathbf{q}_j$ can be denoted by $\mathbf{G}_j\mathbf{y}_j$. Let **A** be the *m*x*p* component loading matrix, where $\mathbf{a}_j$ is the vector of coordinates (component loadings) to represent the j^{th} variable. Let **X** denote the *n*x*p* object scores matrix, containing the coordinates to represent *n* objects in a p-dimensional space. Nonlinear PCA solution is derived by minimizing the least-square loss function σ (X;Y;A), which ultimately minimizes the difference between object scores and original data. To obtain the solution, original data matrix **H** is replaced by the *n*x*m* matrix **Q**, which contains the set of optimally transformed variables $\mathbf{q}_j$. Hence, the loss function can be derived as follows:

$$\sigma(\mathbf{X};\mathbf{Y};\mathbf{A}) = \frac{1}{\mathrm{m}} \sum_{j=1}^{\mathrm{m}} tr\,(\mathbf{X} - \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^{\mathrm{T}})^{\mathrm{T}}(\mathbf{X} - \mathbf{G}_j\mathbf{y}_j\mathbf{a}_j^{\mathrm{T}}) \qquad (1)$$

where *tr* is the trace function and the product $G_j\, y_j\, a_j^T$ provides the coordinates to represent the j^{th} variable in a p-dimensional space. The minimization of such loss function is performed by an Alternating Least Squares algorithm under normalization conditions and restrictions (Gifi, 1990). Imposing $\mathbf{q}_j^T\,\mathbf{q}_j = n$ we derive that transformed variables are standard scores, and component loadings ($\mathbf{a}_j$) constitute the Pearson correlation between principal components and transformed variables.

2.1 Data collection and variables

Because of the subjective nature of happiness, we opted for a data collection based on self-reported measures. In particular, we viewed happiness as the outcome of five interacting dimensions — economic aspect, interpersonal relationships, personal and professional growth, shared values, and innovative behavior. We conceptually defined these dimensions as potential contributors to employee happiness, and we grounded the questionnaire items design on principles and theories drawn from literature (Warr, 2007). These items were included in an-online questionnaire that was distributed to employees by email. Employees participated on a voluntary basis and confidentiality of responses was assured.

The questionnaire consisted of two sections: the first one gathered demographics and occupational data whereas the second one contained the main variables of the study, measuring the individual work-related experience. All the items were measured on a five-point Likert scale, ranging from total disagreement (1) to total agreement (5) with the question statement.

In addition, we included four multiple-choice questions with the aim to evaluate employees working experience and collect personal thoughts on the organization. One of these questions were used to build a proxy variable for happiness at work. In particular, we defined the proxy variable for happiness at work the multiple-choice question *"Describe your experience in the workplace"*, in which respondents could tick three out of nine possible statements describing their experience.

To build our proxy variable we coded as 1 all the answers including the statement *"I am happy to work for my company"*, and as 0 otherwise. Thus, we dummy-coded happy workers as 1 and non-happy workers as 0.

2.2 Respondents and procedure

The study's sample comprises a total of 201 workers across 13 companies of different industries (e.g. steel industry, industrial machineries, insurance, research and development, etc.) in the Northern Italy. Sex is equally represented as half of respondents are male and half female. Most of them (91) are between 36 and 50 years old and have secondary (47.8%) or higher education (44.3%). The average organizational tenure is 9.35 years.

For the purpose of our study we considered 39 variables, arranged on a five-point Likert scale (i.e. 1-5). The first goal is to find latent dimensions among variables which represent the drivers of happiness at work. In order to check for multicollinearity between variables and justify the adoption of Categorical PCA, association between variables was checked by using Spearman's Rank Correlation All the Spearman's Correlation coefficients were statistically significant at 5%. Correlation between groups of variables was checked through Kendall's Tau-b, which was performed for pairs of variables between groups. All the correlations are within the range 0.204 to 0.356, suggesting that between different groups correlations are from low to moderate. Hence, we can assume that more than one component is needed to summarize the information in the data.

All the 39 Likert-scale variables were scaled at ordinal level, so that φ_j is a monotonic function and transformed categories in $\mathbf{q}_j$ respect the rank order of $\mathbf{h}_j$. Transformation plots showed monotonic and non-decreasing curves, so ordinal treatment is appropriate. The whole analysis was run considering a weight = 1 for all the variables. To maximize Variance Accounted For (VAF) across Principal Components while keeping the orthogonal constraint, *Varimax* rotation (with Kaiser Normalization) was chosen.

3 Results

As can be seen from Table 1, the total variance explained resulting from CATPCA varies from 49.05% with two Principal Components to 65.18% with six latent factors.

Table 1: *VAF comparison*

	Eigenvalue						
DIMENSIONS	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	**% VARIANCE**
2	14.045	5.085					**49.050**
3	9.658	7.791	3.477				**53.655**
4	8.445	6.765	4.691	2.736			**58.040**
5	7.321	5.963	4.393	3.741	2.731		**61.922**
6	7.373	6.121	4.288	2.755	2.459	2.426	**65.183**

The choice of the number of components was based on VAF, which should always be considered in the light of interpretability and compared across different dimensions (Linting et al., 2007). In our case study we selected four principal

components as the leap in VAF between three and four dimension was the greatest, whereas after four components the marginal increment was not substantial. Overall, the four dimensions explain 58% of the observed variance. We then compared results assuming the least (i.e. nominal) and the most (i.e. numeric) restrictive analysis level. The difference between the variance in the nominal (49.057) and numeric (46.986) analysis is 2.07%, suggesting that when nonlinearity is assumed, the proportion of explained variance increases. Moreover, the graphical inspection of the object plot highlighted 5 outlier observations, which were removed from the analysis.

Table 2: *Model Summary of CATPCA*

	Variance Accounted For		
Dimension	**Cronbach's Alpha**	**Total (Eigenvalue)**	**% of Variance**
1	.955	8.445	21.653
2	.950	6.765	17.346
3	.936	4.691	12.027
4	.842	2.736	7.014
Total	.981[b]	22.636	58.040

a. Rotation Method: Varimax with Kaiser Normalization
b. Total Cronbach's Alpha is based on the total Eigenvalue

The model summary indicates that the four components extracted account for the 58.04% of the quantified variables. Cronbach's Alpha scores confirm the good internal consistency between items, with values greater than 0.842. We only retained in the analysis variables with a VAF greater than 0.25, that is, at least 25% of the variance in a quantified variable is explained across the principal components

Based on the common traits of the variables selected for each dimension, we labeled them as follows. Dimension 1, which explain 21.65% of variance, is related to self-realization through work and skills and is labelled 'Envisioning'. Dimension 2 captures 17.35% of variance and was labelled 'Equity' as it refers to financial reward and career advancement, traditional HR management, and work-life balance. The third dimension, 'Empowerment', covers 12% of variance and includes non-ordinary matters management, critical issues, and teamwork aspects. Finally, the fourth dimension captures 7% of observed variance, and represents interpersonal relationships aspects ('Empathy').

4 Conclusions

Considering the positive effects associated with happiness at work for both employees and organizations (Fisher, 2010; Wright & Cropanzano, 2004), it is worth investigating its drivers and how to improve it.

The first goal was to find the latent dimensions underlying the original data that could be used to build a data-driven happiness construct. Data were gathered through a questionnaire in which the (39) variables measuring the happiness construct were measured by close-ended questions based on a five-point Likert scale. Such questionnaire was administered by email to 201 employees of 13 different small and medium companies in the North of Italy. The data from the survey of self-reported happiness were submitted to a factor analysis using nonlinear principal components extraction. The resulting four-factor solution was subject to the orthogonal Varimax rotation, since the four dimensions were envisaged to be conceptually distinct. The four components extracted capture almost 60% of original variability (58.04%) and the Cronbach's Alpha score indicate a good internal

consistency between items. The variables included in each factor guided their interpretation. PC1 refers to self-realization through work and was labelled 'Envisioning'. PC2 was named 'Equity' as reflects aspects like financial reward and career growth. The third component, 'Empowerment', includes critical matters management and teamwork. Finally, the last dimension ('Empathy') represents interpersonal relationship with peers. Then, we used the extracted components as explanatory variables to predict the probability of being happy by means of a logistic regression model. The response variable of our model was a proxy variable for happiness at work, dummy-coded 1 for happy workers and 0 otherwise.

To conclude, the exploratory and data-driven approach of the present work produced a multifaceted construct of happiness, which is comprised of four main drivers with different contributions on the probability of happiness at work.

5 References

1. De Sousa, J., & Porto, J. (2015). Happiness at Work: Organizational Values and Person-Organization Fit Impact. *Paidéia (Ribeirão Preto)*, *25*(61), 211-220. https://doi.org/10.1590/1982-43272561201509
2. Fisher, C. (2010). Happiness at Work. *International Journal Of Management Reviews*, *12*(4), 384-412. https://doi.org/10.1111/j.1468-2370.2009.00270.x
3. Gaston-Breton, C., Lemoine, J., Voyer, B., & Kastanakis, M. (2021). Pleasure, meaning or spirituality: Cross-cultural differences in orientations to happiness across 12 countries. *Journal Of Business Research*, *134*, 1-12. https://doi.org/10.1016/j.jbusres.2021.05.013
4. Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, West Sussex: Wiley.
5. Grant, A., Christianson, M., & Price, R. (2007). Happiness, Health, or Relationships? Managerial Practices and Employee Well-Being Tradeoffs. *Academy Of Management Perspectives*, *21*(3), 51-63. https://doi.org/10.5465/amp.2007.26421238
6. Kahneman, D., Diener, E., & Schwarz, N. (1999). *Well- being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.
7. Linting, M., Meulman, J., Groenen, P., & van der Koojj, A. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, *12*(3), 336-358. https://doi.org/10.1037/1082-989x.12.3.336
8. Linting, M., & van der Kooij, A. (2012). Nonlinear Principal Components Analysis With CATPCA: A Tutorial. *Journal Of Personality Assessment*, *94*(1), 12-25. https://doi.org/10.1080/00223891.2011.627965
9. Ryan, R., & Deci, E. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52*, 141–166. https://doi.org/10.1146/annurev.psych.52.1.141
10. Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology, 57(6), 1069–1081.
11. Salas-Vallina, A., Simone, C., & Fernández-Guerrero, R. (2020). The human side of leadership: Inspirational leadership effects on follower characteristics and happiness at work (HAW). *Journal Of Business Research*, *107*, 162-171. https://doi.org/10.1016/j.jbusres.2018.10.044
12. Warr, P. (2007). *Work, happiness, and unhappiness*. Mahwah: Laurence Erlbaum Associates, Inc.
13. Wright, T., & Cropanzano, R. (2004). The Role of Psychological Well-Being in Job Performance. A Fresh Look at an Age-Old Quest. *Organizational Dynamics*, *33*(4), 338-351. https://doi.org/10.1016/j.orgdyn.2004.09.002

References

1. G. Garofalo: Il Progetto ARCHIMEDE obiettivi e risultati sperimentali, Istat Working Paper n.9, (2014)
2. Istat: Le differenze territoriali di benessere - una lettura a livello provinciale, (2019)
3. https://data.education.it/opendata/opendata/
4. https://data.education.it/opendata/opendata/
5. https://data.education.it/opendata/opendata/
6. Istat: Bes 2015 - Il Benessere Equo e Sostenibile in Italia, (2015)

Early and accurate: a Machine Learning approach to predict students' final outcome with registry data

In tempo e accurato: un approccio di Machine Learning per prevedere i risultati finali degli studenti utilizzando i dati del registro elettronico

L. Rossi, M. Cannistrà, T. Agasisti

Abstract
Early prediction of students' outcomes represents a challenge for scholars and practitioners. Predicting the final performance of students represents a valuable support for a more informed (and in time) intervention from schools. This paper aims at predicting student performance for a high school in Italy.
Random Forest algorithm is adopted to find the optimal moment to have accurate predictions of students' final grades, as earlier as possible during the academic year. Models adopt data from electronic registers, updated weekly. Results show that predictive models after four months of schooling, on the middle of the academic year, carry out correctly for more than 75% of students based on the final grades average and on the final grade in three subjects (Maths, Italian and English).

Abstract *La previsione corretta e puntuale dei risultati degli studenti rappresenta una sfida per studiosi e professionisti in ambito scolastico. La predizione del rendimento finale degli studenti è un valido supporto per un intervento più informato (e tempestivo) delle scuole. Questo articolo mira a prevedere il rendimento degli studenti di una scuola superiore in Italia.*
L'algoritmo di Random Forest viene adottato per trovare il momento ottimale per avere previsioni dei voti finali degli studenti in modo accurato e tempestivo. I modelli adottano dati provenienti da registri elettronici, aggiornati settimanalmente. I risultati mostrano che i modelli predittivi dopo quattro mesi di scuola, a metà anno accademico, funzionano

Lidia Rossi, Dipartimento di Ingegneria Gestionale Politecnico di Milano, via Lambruschini 4/b, Milano, Italia ; email: lidia.rossi@polimi.it
Marta Cannistrà, Dipartimento di Ingegneria Gestionale Politecnico di Milano, via Lambruschini 4/b, Milano, Italia ; email: marta.cannistra@polimi.it
Tommaso Agasisti, Dipartimento di Ingegneria Gestionale Politecnico di Milano, via Lambruschini 4/b, Milano, Italia ; email: tommaso.agasisti@polimi.it

correttamente per oltre il 75% degli studenti sulla base della media dei voti finali e del voto finale in tre materie (Matematica, italiano e inglese).

Key words: Learning Analytics; High school Students' performance prediction; Machine Learning

1 Introduction, motivation and research questions

Predicting student academic performance is an important research topic in Learning Analytics (LA) field [1]. Accurate prediction of students' school performance can be useful to evaluate a student's academic performance in advance, which allows to plan and apply personalized and on-time learning intervention. These plans are useful for improving student performance for both those who already have a high academic achievement and for students considered to be at risk of failure [7]. The increasing academic interest over these topics, favours the application and the study of Early Warning Systems (EWS) in the educational domain. The idea of EWS is to use the available data to predict students' school performance early on time and to propose personalized interventions aimed either at their improvement or at the prevention of failure [5] [3] [9]. However, measuring students' academic performance is challenging due to the various factors influencing it: they can be psychological and personal, demographics or academic, as well as concerning family, economic context, or educational setting [10] and they are often correlated in a complex and non-linear way. The complexity of the factors influencing students' achievements and the increasing available information led to the development of new technologies and information systems to study and predict students' career. Machine learning approaches can be adopted in this context to formulate accurate predictions and deal with highly complex data [2] [8]. In this work, an in-depth study using data from the electronic register of an Italian high school is carried out. The objective is the classification of students based on the final average or on the final evaluation in three specific subjects (Mathematics, Italian and English). The main research questions this work aspires to answer are the following:

How **accurately** and how **early** Machine Learning models are able to predict students' final grades in the academic year?

2 Data

The data used for the present research refer to the students enrolled in a private high school in Milan in the academic year 2018/2019. In particular, data are about students' characteristics and come from administrative database and electronic register's data. The variables used for the study can be classified in three categories:

- Static data. This group includes all information known before the beginning of the academic year regarding the personal characteristics of students, information about classes and students' academic performance of the previous year. This group of variables contains demographic information such as gender, citizenship, age (compared to that of peers) and PEI and PDP that indicates if the student follows some personalized plans due to minor or major certified disabilities. The total number of teachers in the class and how many of them are tenured is also known. This set of variables includes prior academic results. Among these, there is a variable indicating whether the student was promoted in June the previous year or he/she had some specific restorative exams in September for being promoted or he/she was rejected or transferred. Also, the average of marks (on a scale between 0 and 10) and the final grade obtained by the students in Mathematics, Italian and English. Lastly, data indicates the school track in which the student is enrolled (the high school system in Italy includes various types of schools: scientific, linguistic, etc...).
- Time-variant information. Information deriving from the electronic registry during each school week. These variables concern the grades' average in Mathematics, Italian and English, the number of delays (i.e. entry at school when the lesson already started), of absences, of merit, diligence and disciplinary notes.
- Output variables. This set of variables includes students' performance, which in the Italian Educational system has a 10-level scale. The variables included in the dataset are the final grade in Mathematics (range of values: 3-10), Italian (range of values: 5-10) and English (range of values: 5-10) and the mean final mark, ranging from 6 to 10.

Table 1 describes variables distributions.

Table 1: Descriptive statistics

Name	***Mean or proportion***	***Standard deviation or range***
Age (-1)	13%	
Age (0)	81,5%	
Age (+1)	5,3%	
Gender (1=Male)	51,1%	
Citizenship (1=Italian)	97,7%	
PEI (1=Yes)	0,7%	
PDP (1=Yes)	16%	
Sector (L. Scientifico Paritario)	25,3%	
Sector (L. Classico Paritario)	15,8%	
Sector (L. Scienze Umane Paritario)	24,9%	
Teachers	18,00	3,42
Tenured teachers	14,96	2,46
Final status 2017/2018 (1=Promoted in June)	82,2%	
Marks average 2017/2018	7,546	0,837
Mathematics final grade 2017/2018	6,68	1,45
Italian final grade 2017/2018	7,14	1,06
English final grade 2017/2018	7,41	1,05
Absences	5,627	0-59
Merit notes	0,563	0-6
Diligence notes	0,847	0-19
Disciplinary notes	0,146	0-7
Delays	2,960	0-31

Mathematics grades	6,504	1.5-10
Italian grades	7,033	3-10
English grades	7,255	3-10
Final grades average (<=5 / at risk)	0	
Final grades average (>=8 /high score)	60,5%	
Mathematics final grade (<=5 / at risk)	9,8%	
Mathematics final grade (>=8 / high score)	36,1%	
Italian final grade (<=5 / at risk)	1,4%	
Italian final grade (>=8 / high score)	42,5%	
English final grade (<=5 / at risk)	0,9%	
English final grade (>=8 / high score)	57,5%	

3 Methodology

In this paper a Machine Learning approach (Random Forest algorithm) is used for the empirical analysis. This study is part of the supervised learning field, it has predictive goals and concern classification. The paper also provides robust checks with alternative algorithms: Classification Trees and Neural Networks, the results are consistent and similar to the ones obtained using RF. To evaluate the results of the study, two important indices will be considered about the performance of the prediction models: accuracy and AUC [4] [6]. A Monte Carlo simulation is also done to test results with a more numerous dataset.

4 Results

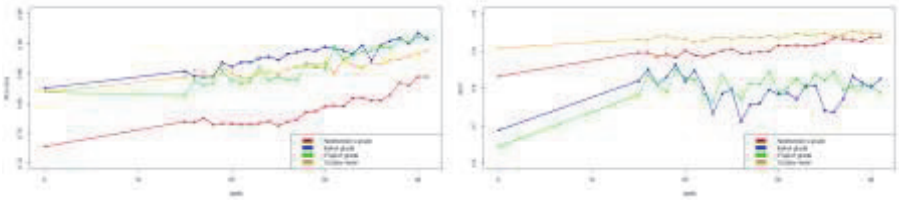
At first the evolution during the year of accuracy and AUC value are evaluated on the original dataset then on the simulated one. Finally, to investigate determinants, the variable importance is considered.

4.1 The performance of the algorithm for predicting students' results

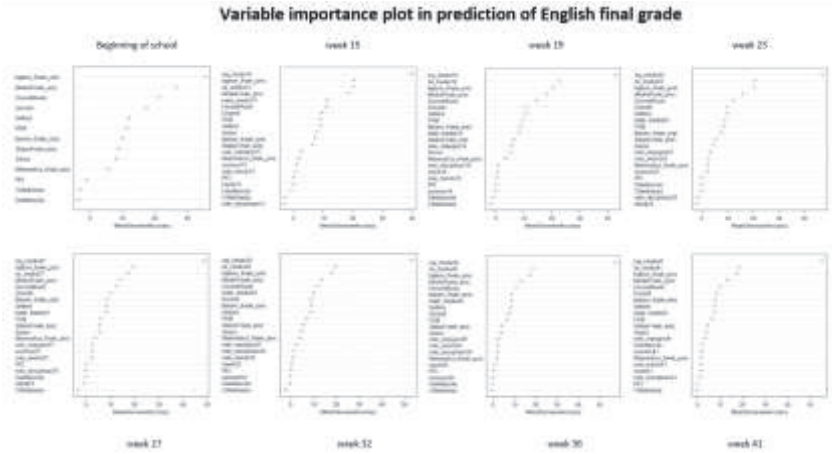
When considering the classification of students based on grades average, initially there is an accuracy of about 82% and an AUC of 0.908, at the end of the year an accuracy of 88.8% is reached and an AUC of 0.946. Already in the 15th week of school out of 41, i.e. the week of December 1st, the accuracy increase about of 3% with the inclusion of registry's data in the analysis.
As for the prediction of the final grade in mathematics, it is initially predicted with lower performance indices than the previous case (73% of accuracy and AUC value equal to 0.833); at the end of the academic year the accuracy is equal to 84.4% and the AUC to 0.939. Already with the addition of the December data 4% in accuracy

is added, exceeding the 75% of accuracy; 80% is reached after 33 weeks. Data of the register increase the accuracy more than 10%.
The final grade of Italian is already correctly predicted at the beginning of the year for about 83% of students, at this time of the year the AUC is equal to 0.689. These values reach respectively the values 90.7% and 0.825 at the end of the academic year. In December the threshold of 85% in accuracy is exceeded with the respective AUC value of 0.802, while after 36 weeks 90% accuracy is reached with AUC value 0.736. After 24 weeks from the beginning of the academic year more than 5% in accuracy is added by using information of the electronic register for the predictions.
Finally, regarding the classification of students based on their final grade in English, at the beginning of the academic year about 82\% of students are correctly classified, with an AUC value of 0.645. The prediction improves more and more until it reaches an accuracy of 91.2% and an AUC value of 0.788; at this time data of the register increase the accuracy of about 10% while of 5% from the 31th week.
Evolution of these indices are reported in Figure 1.

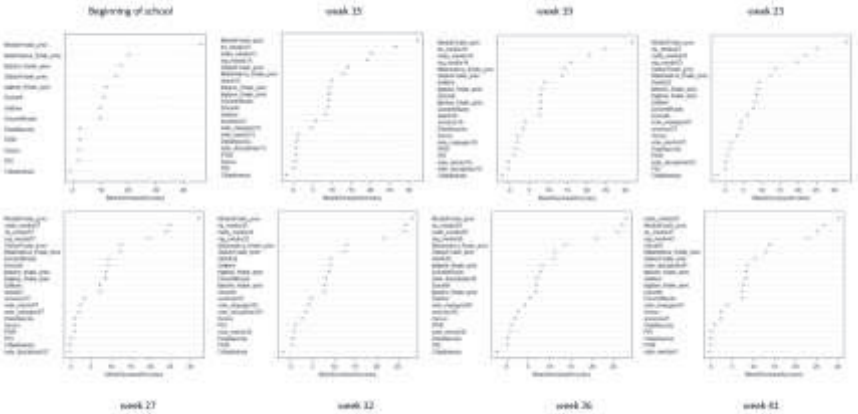
Figure 1: Evolution over the year of Accuracy and AUC value



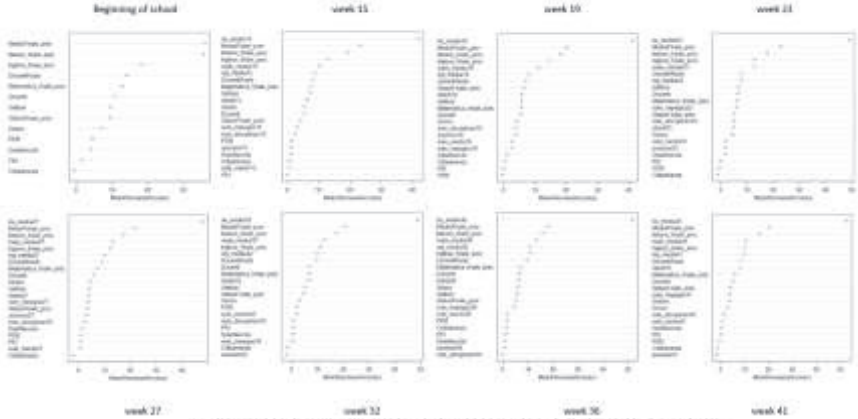
Using variable importance plot of Random Forest on the original dataset the evolution over time of variable importance in the prediction of each output variable is evaluated at the beginning and at the end of the academic year and every first of the month starting from December until June. Results show that grade of the current year and performances of the previous year are always determinants.



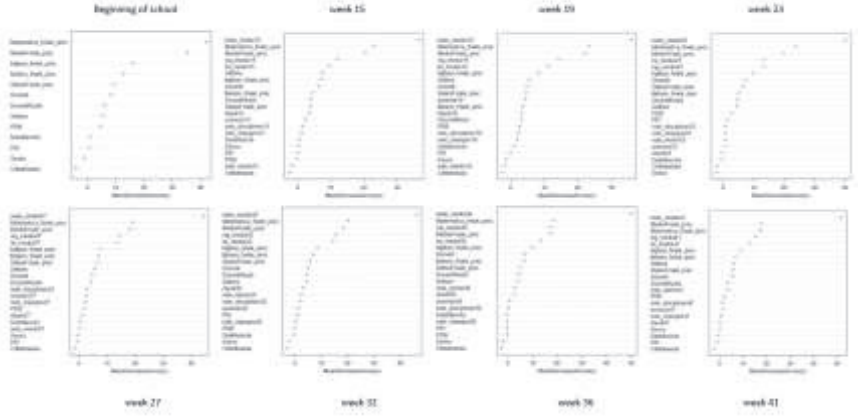
Variable importance plot in prediction of final grade mean



Variable importance plot in prediction of Italian final grade



Variable importance plot in prediction of Mathematics final grade



4.2 A Monte Carlo Simulation

One limitation of this study comes from the low number of observations (as data is available for about 440 students). To test the results obtained with a larger dataset, a data simulation is implemented starting from the original dataset. With the sim_df command on R software data are reproduced with the same distributions and correlations as the original dataset. The simulated dataset consists of 2,000 observations and has the same distributions as the original dataset. The obtained results show indices that are lower than the one obtained using the original dataset but consistent with them and a greater variation in terms of accuracy added by registry's data.

5 Concluding remarks and implications

This paper employs a machine learning approach to realize an EWS finalized at predicting students' academic results. The main innovation proposed here is the use of weekly data extracted from the registry, merging this information with administrative data. The immediate potential use of the EWS proposed here is to realize informative dashboards which can be at disposal of teachers and principals. This information can help teachers to reflect critically on interventions for the students of their classes, also with a specific attention to personalized interventions, that might be personalized over time. A dashboard built at school level can instead be at disposal of the school principal, with the aim of an overview of current academic performance of students as projected at the end of the school year.
A potential limitation of the present study is that the output variable predicted are somehow predicted by the intermediate grades assigned by the teachers - so creating a potentially endogenous circle. In this vein, it would be very useful to check the validity of EWSs in contexts in which the output can be measured by means of external evaluations, as for example standardized test scores.

References

1. Shahiri, A. M.; Husain, W.; Rashid, N. A. (2015) A Review on Predicting Student's Performance Using Data Mining Techniques; Procedia Comput. Sci., Volume 72, Pages 414-422, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2015.12.157.
2. Aydoğdu, Ş. (2020) Predicting student final performance using artificial neural networks in online learning environments; Educ Inf Technol 25, 1913-1927. https://doi.org/10.1007/s10639-019-10053-x
3. Carl, B.; Richardson, J. T.; Cheng, E.; Kim, H.; Meyer, R. H. (2013) Theory and Application of Early Warning Systems for High School and Beyond; J. Educ. Stud. Placed Risk (JESPAR), 18:1, 29-49. DOI:10.1080/10824669.2013.745374
4. Efron, B.; Hastie, T. (2016) Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (1st. ed.). Cambridge University Press, USA.
5. Henry, K.L.; Knight K.E.; Thornberry, T.P. (2012) School disengagement as a predictor of dropout, delinquency, and problem substance use during adolescence and early adulthood. J Youth Adolesc. 41(2):156-66. doi: 10.1007/s10964-011-9665-3.
6. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.). Springer.

Students' experience with distance learning during Covid 19 pandemic in Southern Italy

Una esperienza di didattica a distanza in alcuni studenti del Sud Italia durante la pandemia Covid 19

Angela Maria D'Uggento and Nunziata Ribecco

Abstract Approximately two years after the Covid19 pandemic, researchers are interested in assessing its impact on society, mental health, and education. To ensure continuity of instruction, students participated in classes from home. While the technical problems were somehow manageable, the deterioration of mental health was undoubtedly more difficult to cope with and, likely, had traumatic effects on many young people. With a survey of 1,887 high school students in the city of Bari (Italy) we aimed to investigate relevant aspects of social well-being, mental health, and relationships between distance learning and student performance. The results showed that students maintained their performance in distance learning despite suffering from psychological weakness, restlessness and anxiety. This could have interesting implications for school administrators and students.

Abstract *A due anni dalla pandemia, i ricercatori stanno valutandone l'impatto sulla società, in particolare sulla salute mentale e sull'istruzione. Per garantire continuità nella formazione, le scuole hanno adottato la didattica a distanza e, se i problemi tecnici sono risultati in qualche modo gestibili, il peggioramento dello stato psicologico degli studenti è stato senza dubbio più difficile da affrontare. Attraverso un'indagine su 1.887 studenti delle scuole superiori di Bari, una città del Sud Italia, si è cercato di approfondire i temi del benessere sociale e mentale e delle relazioni tra didattica a distanza (DL) e performance scolastica. I risultati hanno rivelato prestazioni invariate, pur in uno stato di disagio psicologico, ansia e paura. Lo studio può risultare di interesse per dirigenti scolastici e studenti.*

Key words: distance learning, high school performance, Covid19 pandemic, mental health, classification trees.

[1]Corresponding author: Angela Maria D'Uggento, Department of Economics and Finance, University of Bari; angelamaria.duggento@uniba.it.

Nunziata Ribecco, Department of Economics and Finance, University of Bari; nunziata.ribecco@uniba.it.

1 Introduction

At the end of 2019, an unknown pandemic is spreading around the world, causing many deaths and changing people's lives by drastically limiting their social activities. Movie theatres, restaurants, sports centres, and schools were closed to avoid physical contact between people that could promote the spread of the Covid 19 virus. Education was no exception, and classes were taught by distance learning (DL), although this was unusual for Italian schools, which normally teach face-to-face. This alternative approach was introduced abruptly and without prior preparation for students and teachers and caused no fewer difficulties. It was to be expected that students' academic performance would deteriorate in such a context (Engzell et al, 2021). One of the main critical issues, that was certainly unavoidable in the initial phase of the lockdown, concerned how DL could be implemented.

The Italian Statistical Office's report on the availability of learning tools and IT devices at home shows that 12.3% of school-age children do not have a computer at home, a figure that is close to 20% in the South, i.e., about 850,000 young people between 6 and 17 years old (ISTAT, 2020). Another 57.0% have a PC at home but have to share it with other family members. Only 6.1% of children aged 6-17 live in families where there is at least one computer per component. In addition, about 41.9% of minors live in crowded housing.

To find out if this context affects students' learning, the results of the National Organization for the Evaluation of Educational Achievement (INVALSI) are useful. They report that the learning skills of high school students in Italian and Mathematics have decreased in the pandemic period compared to 2019, and the existing territorial differences between Italian regions have even widened. Most importantly, the INVALSI report confirms a stable correlation between student performance and socioeconomic environment, with students in unfavourable socioeconomic circumstances losing almost twice as many points as students in more favourable circumstances (Mascheroni et al., 2021). The territorial distribution of the INVALSI analysis shows that the educational gap between the richer regions in central and northern Italy and those in the south is widening.

This paper focuses on the DL experiences of 1,887 high school students from a large city in southern Italy, one of the regions that, according to INVALSI results, experienced increased learning difficulties during the pandemic. The aim of the study is to understand the students' opinions on DL and to find out if their emotional situation during the pandemic affected their performance.

The paper is organized as follows. After a brief introduction to the framework, Section 2 describes the survey and statistical methods. Section 3 presents the main results, which contribute to a better understanding of the phenomenon. Finally, brief conclusions are drawn in Section 4.

2 Data and methods

A web questionnaire was sent to high school students in the city of Bari. 1,887 questionnaires were collected while the students participated in the Statistics Project within the National Plan for Scientific Degrees (PLS), supported by the Ministry of Education and Research for the academic year 2020-2021. Their high schools, as well as other Italian high schools, participated in the PLS program, which is conducted every year by some Italian universities to promote the acquisition of scientific and statistical skills through laboratory activities.

The questionnaire was divided into six sections dealing with different aspects: 1. Sociodemographic information 2. Sources of information 3. Social well-being and mental health 4. Opinions on lockdown rules 5. DL and school performance 6. Future perspective. Of all the aspects investigated, this paper deals with those related to DL and school performance, also in relation to students' main concerns. The research hypotheses were based on evaluating possible differences in performance between the two periods and whether the socioeconomic context might have influenced them in some way. Since household income was not available, the educational level and employment status of the household head were used as indicators of economic conditions. Specifically, the study examined students' opinions of DL, the factors that might lead them to prefer face-to-face education to DL, and their performance compared to the previous year. In addition, the possible relationship between psychological condition and the possibility of making extremely dangerous decisions, such as dropping out of school due to the threat of a pandemic, was examined using the chi-square test and the Classification tree method.

3 Main results

Students expressed a preference for face-to-face classes, citing as the main reason the higher level of attention despite the various sources of distraction at home and the opportunity to interact with the teachers and the class, followed by the better comprehensibility and clarity of the teachers and the lower learning effort. The only reason for preferring distance learning was anxiety about oral exams, which is lower when activities occur through a filter such as the screen (see Table 1). However, when asked for a general preference among the three options (DL or face-to-face instruction or no preference), the gap between distance education (39.5%) and face-to-face instruction (36.6%) narrowed significantly in favour of the former. These data suggest that there are some aspects that were not directly considered in the questionnaire, such as the distance to the school attended, the use of transportation to reach the school, or other factors that made students slightly prefer DL. The fact that more than one-third of respondents preferred the traditional approach indicates a desire for greater social and educational involvement that only the school can provide. In fact, 72.9% of them said they greatly missed social activities with

classmates and 48.6% preferred face-to-face classes (p-value=0.000). We also tried to find out if students' performance deteriorated during the pandemic. 74.9% of the students surveyed reported that their performance had remained unchanged (see Table 2). Relationships with classmates had also not changed for 62.2% of respondents and had actually improved for another 22.4%. When asked about relationships with teachers, the percentages were almost the same, 67.1% and 12.5% respectively, although 55.1% of students defined their teachers more strictly in their evaluation than before. Overall, DL received an average score of 6.7 (SD =1.91; range 1-10) from respondents.

Table 1: Motivations and preferences for distance learning or face to face learning

Motivations	***DL***	***Face to Face***	***No difference***
Better understanding of lessons	11.3	47.0	41.7
Attending lessons more carefully	10.7	58.3	31.1
Ease of participation	20.3	47.8	31.9
Heavier study load	54.1	14.6	31.3
Greater anxiety for oral tests	12.7	53.4	33.9

Table 2: Comparison between the school performance before and during pandemic

Performance before pandemic	***Performance during pandemic***				
	Scarce	Sufficient	Good	Excellent	Total
Scarce (D)	5	15	8	0	28
Sufficient (C)	29	183	146	1	359
Good (B)	15	123	954	63	1,155
Excellent (A)	1	11	60	273	345
Total	50	332	1,168	337	1,887

Following these results, we sought to understand whether they remained unchanged when the influence of the socioeconomic context or the respondents' opinions about DL problems, was considered. As shown in Figure 1, there are no major differences in the proportions of respondents' preferences depending on the father's employment status (χ^2 p-value=0.133). This result is consistent with the previous ones. More likely, all students had either a device (smartphone/tablet/PC) or an Internet connection at home that allowed them to participate in lessons from DL without much effort; in particularly critical cases, schools loaned their students the devices. A similar result emerges by the question *"Have you ever thought about dropping out of school because of distance learning?",* as responses are not affected by father's employment status (χ^2p-value=0.053, combining "not employed" and "retired" categories to meet testing assumptions). Overall, 45.7% of respondents felt that DL was not a reason for dropping out of school, although this response depended on the possibility of returning to "normal life" at the end of the year (χ^2p-value=0.005).

Figure 1: Respondents' preferences and father's employment status

The importance of the experience of DL as a predictor of the possible decision to drop out and, at the same time, the null influence of the respondent's economic conditions are confirmed by the Classification tree (see Fig.2). It allows us to draw two paths: on the right side are the students who had a negative DL experience, so that 33.4% of them sometimes/often thought about dropping out of school. They were also characterized by very low hope (node 6) and serenity (node 10). In contrast, respondents who scored a serenity level of 3 to 4 showed undifferentiated ratings of DL (node 14). On the left are the more resilient students who were able to view the experience of DL in a positive light. Among them, those who felt strongly lonely accounted for the highest percentage of dropout thoughts (28.1%, node 5). In contrast, those who felt less lonely and never experienced psychological fragility had the lowest percentage (5.5%, node 8) of this thought. Finally, the percentage of students who did not feel sustained by hope, although they tended to be satisfied with DL, increased to 22.3%, as shown in terminal node 12. Respondents who experienced positive feelings such as little loneliness and sometimes mental fragility, but saw the future with hope and serenity, repressed negative decisions as they never thought of dropping out of school (92.4%, as can be seen in terminal node 16).

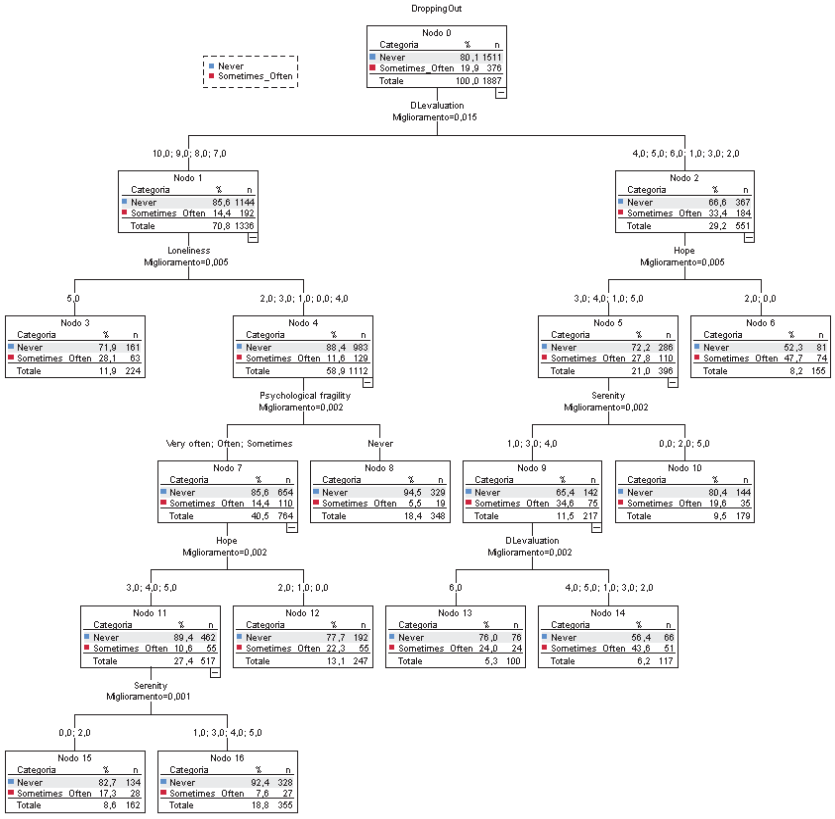


Figure 2: Classification tree for decision of dropping out of school because of distance learning

4 Final remarks

Despite all the concerns mainly expressed during the pandemic period of learning from home, the results of our survey show an overall positive balance. About 39% of respondents preferred it to face-to-face instruction. Logistical and organizational issues probably played some role in the overall positive assessment. Undoubtedly, students also faced some difficulties. It emerged that most students had difficulties in understanding the lessons and interacting with the teachers. In terms of teaching effectiveness, then, the traditional face-to-face form of teaching offers undeniable advantages, which were also acknowledged by the students themselves.

It is also reassuring that about 80% of students have never thought about dropping out of school. Among those who "sometimes" or "often" thought about it, we noted the influence of the emotional state of psychological fragility caused by the inconvenience and difficulties of parents during the pandemic: their proportion decreases when students report being more hopeful and relaxed about the temporary situation and it is high among those who suffered more.

Italian high schools were not really prepared for this emergency in terms of adequate IT equipment and staff preparation. This is certainly one of the factors that could have had the greatest impact on the achievement of the educational objectives offered. Be that as it may, our survey showed that the performance of the students did not deteriorate but remained quite stable during DL.

Undoubtedly, this study deals with a convenience sample, composed of students living in a major city, but it provides evidence that students' online learning experiences can vary widely. This may have policy implications that can help policymakers and schools improve online learning. The pandemic has plunged Italy into crisis, but there have also been many economic and consequently digital inequalities in the past. School attendance represents a fundamental moment of growth in a person's life, and education must be guaranteed without economic or social inequalities.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont, CA (1984).
2. Engzell, P., Frey, A., Verhagen, M.D.: Learning loss due to school closures during the COVID-19 pandemic. PNAS, Apr 2021, 118 (17) e2022376118; DOI: 10.1073/pnas.2022376118.
3. ISTAT: La situazione del paese. Spazi in casa e disponibilità di computer per bambini e ragazzi (2020).
4. INVALSI https://www.invalsiopen.it/risultati/risultati-prove-invalsi-2021.
5. Mascheroni, G., Saeed, M., Valenza, M., Cino, D., et al: La didattica a distanza durante l'emergenza COVID-19: l'esperienza italiana. Centro di Ricerca Innocenti dell'UNICEF, Firenze (2021).
6. PLS program available at: https://www.pianolaureescientifiche.it/.

Time series methods and Applications

Trend and cycle decomposition in nonlinear time series

Decomposizione in trend e ciclo di serie storiche nonlineari

Maddalena Cavicchioli

Abstract In this work we derive the Beveridge-Nelson decomposition and the state space representation for multivariate (co)integrated time series subject to Markov switching in regime. Then we provide explicit matrix expressions for the trend and cyclical components which improve computational performance since they are readily programmable. Further we develop impulse-response function analysis. Applications illustrate the feasibility of the proposed approach.

Abstract *In questo lavoro si ottengono una decomposizione di tipo Beveridge-Nelson e una rappresentazione state-space di serie storiche multivariate (co)integrate soggette a cambiamenti di regime. Inoltre si sviluppano espressioni matriciali esplicite della decomposizione in trend e del ciclo che migliorano le performance computazionali essendo facilmente programmabili. Infine si analizzano le funzioni impulso-risposta. La fattibilitá dell'approccio proposto é illustrata tramite applicazioni.*

Key words: Beveridge-Nelson decomposition, trend and cyclical component, Markov switching (co)integrated processes, Impulse-response function.

1 Introduction

Since the influential works [9] and [10] , Markov switching (MS) models have attracted considerable interest among econometricians to model various nonlinear time series and to capture the business cycle asymmetries (see, for example, [5] and [6]). Moreover, the problem of decomposing a time series into unobservable components having meaningful interpretation has always been very important in econometrics. For example, macroeconomists are interested in modeling the behaviour of aggregate output to analyze the characteristics of trends and cycles in

Maddalena Cavicchioli, Department of Economics Marco Biagi, University of Modena and Reggio Emilia, Viale Berengario 51, 41121 Modena (Italy); e-mail: maddalena.cavicchioli@unimore.it

the economy (see [4], [11], [8] and [2]). Thus, linear models have been augmented with MS processes, so as to incorporate asymmetries associated with business cycle or other macroeconomic nonlinearities. [7] modified the BN decomposition to the univariate MS AR model proposed by [9], and showed that this modified decomposition can be carried out without the necessity of truncating an infinite sum. [3] incorporated MS into a BN decomposition formulated in a single source of error state space form, allowing regime shifts in the long-run multiplier as well as in the short-run parameters. [13] specified a time series with MS to model business cycle phases in permanent and transitory components of post-war US real GNP and consumption. [17] introduced a new approach to trend/cycle decomposition of MS time series, which they called the regime-dependent steady-state (RDSS) decomposition. This procedure extends the BN decomposition to the setting where the reduced form dynamics of a given series can be captured by a MS forecasting model. [12] provided exact calculation of the BN trend and cyclical components for certain classes of univariate and multivariate MS autoregressive sequences. The key to exact BN decomposition is to recognize that the latent first-order MS process in the model has an AR(1) representation, and that the model can be cast into a state space form. From this, he showed that impulse-response function analysis can be processed with respect to a symmetric continuous shock (from the given process) and an asymmetric discrete shock (to the latent MS variable in the model).

We contribute to the existing literature in threefolds. First, we provide a unified framework for the BN decomposition and the state space representation for multivariate (co)integrated MS time series, which include those considered in [12], [15], [16] and [1]. Then we give a state space representation of the process, and derive explicit expressions for the BN trend and cyclical components in terms of the matrices involved in such a representation. Our matrix expressions in closed-form improve computational performance since they are readily programmable and greatly reduce the computational cost. Second, impulse-response function analysis can be developed within our multivariate context by using the proposed matrix calculus. Third, the theoretical results are illustrated via applications.

2 Integrated processes with Markov switching

Let $\mathbf{y}_t$ be a K-dimensional random vector with values in $\mathbb{R}^K$ with $\mathbf{y}_t \sim I(1)$. Suppose that the first difference of $\mathbf{y}_t$ is driven by the following MSI(M) VAR(p) model

$$\Phi(L)\,\Delta\mathbf{y}_t = \nu_{s_t} + \mathbf{u}_t$$

where $\nu_{s_t} \in \mathbb{R}^{K\times 1}$, $\Phi(L) = \mathbf{I}_K - \Phi_1 L - \cdots - \Phi_p L^p$ is a $K \times K$ matrix polynomial in the lag operator L and $\mathbf{u}_t \sim \text{i.i.d.}\, N(0,\Sigma_{\mathbf{u}})$ being $\Sigma_{\mathbf{u}} \in \mathbb{R}^{K\times K}$ positive definite.
Such MS models serve well to depict the business cycle of various industrialized economies as shown, for example, in [14]. We introduce the following assumptions:

Assumption 1. The roots of the equation $\det \Phi(z) = 0$ lie outside the unit circle.

Assumption 2. The process (s_t) is an irreducible, aperiodic and ergodic Markov chain with values in the set $\{1,\dots,M\}$, stationary transition probabilities $p_{ij} = Pr(s_t = j|s_{t-1} = i)$ for $i, j = 1,\dots,M$, and unconditional (or steady state) probabilities $\pi_i = Pr(s_t = i)$ for $i = 1,\dots,M$.

Collect p_{ij} and π_i into a $M \times M$ matrix $\mathbf{P} = (p_{ij})$ and a $M \times 1$ vector $\pi = (\pi_1 \cdots \pi_M)'$, called the *transition probability matrix* and the *stationary vector* of the chain, respectively.

Assumption 3. The shock $(\mathbf{u}_t)$ is independent of the Markov chain (s_t).

The MSI(M) VAR(p) model has the following state space representation

$$\begin{cases} \Phi(L)\,\Delta\mathbf{y}_t^* = \widetilde{\Lambda}\,\delta_t + \mathbf{u}_t \\ \delta_t = \mathbf{F}\,\delta_{t-1} + \mathbf{w}_t \end{cases}$$

where:

$$\begin{aligned} \Delta\mathbf{y}_t^* &= \Delta\mathbf{y}_t - E[\Delta\mathbf{y}_t] \\ E[\Delta\mathbf{y}_t] &= \Phi(1)^{-1}\,\Lambda\,\pi \\ \Lambda &= (\nu_1 \ \cdots \ \nu_M) \in \mathbb{R}^{K\times(M)} \\ \widetilde{\Lambda} &= (\nu_1 - \nu_M \cdots \nu_{M-1} - \nu_M) \in \mathbb{R}^{K\times(M-1)} \end{aligned}$$

and $\mathbf{w}_t$ is the asymmetric discrete shock of the latent state variable $\delta_t \in \mathbb{R}^{M-1}$. Finally, $\mathbf{F}$ is the $(M-1)\times(M-1)$ transition matrix associated with δ_t:

$$\mathbf{F} = \begin{pmatrix} p_{11} - p_{M1} & p_{21} - p_{M1} & \cdots & p_{M-1,1} - p_{M1} \\ p_{12} - p_{M2} & p_{22} - p_{M2} & \cdots & p_{M-1,2} - p_{M2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{1,M-1} - p_{M,M-1} & p_{2,M-1} - p_{M,M-1} & \cdots & p_{M-1,M-1} - p_{M,M-1} \end{pmatrix}.$$

3 Trend and cycle decomposition

As in the case of a linear VAR process, the BN trend component (i.e., the value of the series would take if it were on its long-run path) is given by

$$\mathbf{y}_t^{TR} = \mathbf{y}_t + \sum_{j=1}^{\infty} E(\Delta\mathbf{y}_{t+j}^*|I_t)$$

where I_t is the information set up to time t.

The second summand (whose opposite is the cyclical component $\mathbf{y}_t^c$) can be explicitly calculated from the following state space representation of the basic MSI(M) VAR(p) model (with $n = pK + M - 1$):

$$\widetilde{\mathbf{X}}_t = \widetilde{\boldsymbol{\Phi}}\,\widetilde{\mathbf{X}}_{t-1} + \widetilde{\mathbf{e}}_t$$

where

$$\widetilde{\mathbf{X}}_t = (\Delta\mathbf{y}_t^{*'} \quad \Delta\mathbf{y}_{t-1}^{*'} \cdots \Delta\mathbf{y}_{t-p+1}^{*'} \quad \boldsymbol{\delta}_{t+1}^{'})^{'} \in \mathbb{R}^n$$
$$\widetilde{\mathbf{e}}_t = (\mathbf{u}_t^{'} \quad 0^{'} \cdots 0^{'} \quad \mathbf{w}_{t+1}^{'})^{'} \in \mathbb{R}^n$$

and

$$\widetilde{\boldsymbol{\Phi}} = \begin{pmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p & \widetilde{\Lambda} \\ \mathbf{I}_K & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mathbf{I}_K & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_K & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \mathbf{F} \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Proposition 1. *Under the above Assumptions, the above state space representation of the first difference stationary process* $(\mathbf{y}_t)$*, following the basic MSI*(M) *VAR*(p)*, is stable.*

This follows from the Assumptions and $\rho(\mathbf{F}) < 1$, which imply $\rho(\widetilde{\boldsymbol{\Phi}}) < 1$.

Starting from the latter state space representation of the basic MSI(M) VAR(p) model, we derive an explicit expression for the BN trend $\mathbf{y}_t^{TR}$ (and hence for the cyclical component $\mathbf{y}_t^c$) which involves the matrix $\widetilde{\boldsymbol{\Phi}}$ and the vector $\boldsymbol{\delta}_{t|t}$.

Proposition 2. *Let* $(\mathbf{y}_t)$ *be the first difference stationary process, driven by the basic MSI*(M) *VAR*(p) *model. Under the above Assumptions, the permanent or trend component of the BN decomposition of* $(\mathbf{y}_t)$ *is given by*

$$\mathbf{y}_t^{TR} = \mathbf{y}_t + \mathbf{L}\,\widetilde{\boldsymbol{\Phi}}\,(\mathbf{I}_n - \widetilde{\boldsymbol{\Phi}})^{-1}\,\widetilde{\mathbf{X}}_{t|t}$$

where

$$\mathbf{L} = (\mathbf{I}_k \quad 0 \cdots 0) \in \mathbb{R}^{K\times n}$$
$$\widetilde{\mathbf{X}}_{t|t} = E(\widetilde{\mathbf{X}}_t|I_t) = (\Delta\mathbf{y}_t^{*'} \quad \Delta\mathbf{y}_{t-1}^{*'} \cdots \Delta\mathbf{y}_{t-p+1}^{*'} \quad \boldsymbol{\delta}_{t+1|t}^{'})^{'} \in \mathbb{R}^n$$
$$\boldsymbol{\delta}_{t+1|t} = E(\boldsymbol{\delta}_{t+1}|I_t) = \mathbf{F}\,\boldsymbol{\delta}_{t|t}$$

and $n = pK + M - 1$.

4 Impulse-Response Analysis

In the considered MS model there are two independent (under Assumption 2) shocks: a symmetric continuous shock $\mathbf{u}_t$ and an asymmetric discrete shock $\mathbf{w}_t$. Given the latter state space form, we find the impulse-response with respect to these shocks:

Proposition 3. *Let* $(\mathbf{y}_t)$ *be the first difference stationary process drive by the basic* MSI(M) VAR(p) *model. Under the above Assumptions, the impulse-response function analysis for* $(\mathbf{y}_t)$ *with respect the two independent shocks* $\mathbf{u}_t$ *and* $\mathbf{w}_t$ *is given by*

$$\frac{\partial \mathbf{y}_{t+j}}{\partial \mathbf{u}_t'} = \mathbf{L}\,(\mathbf{I}_n - \widetilde{\boldsymbol{\Phi}}^{j+1})\,(\mathbf{I}_n - \widetilde{\boldsymbol{\Phi}})^{-1}\,\mathbf{L}' \in \mathbb{R}^{K\times K}$$

$$\frac{\partial \mathbf{y}_{t+j}}{\partial \mathbf{w}_t'} = \mathbf{L}\,\widetilde{\boldsymbol{\Phi}}\,(\mathbf{I}_n - \widetilde{\boldsymbol{\Phi}}^{j+1})\,(\mathbf{I}_n - \widetilde{\boldsymbol{\Phi}})^{-1}\,\mathbf{R}' \in \mathbb{R}^{K\times(M-1)}$$

where $\mathbf{L} = (\mathbf{I}_k \quad 0\cdots 0) \in \mathbb{R}^{K\times n}$ *and* $\mathbf{R} = (0\cdots 0 \quad \mathbf{I}_{M-1}) \in \mathbb{R}^{(M-1)\times n}$ *(with* $n = pK + M - 1$*).*

5 A numerical example

We simulate a MS VAR process such that $M = K = 2$ and $p = 1$, that is, a 2–state bivariate AR(1) model:

$$\mathbf{y}_t - \boldsymbol{\Phi}\,\mathbf{y}_{t-1} = \nu_{s_t} + \mathbf{u}_t \qquad\qquad \mathbf{u}_t \sim \text{NID}(0, \boldsymbol{\Omega})$$

where

$$\nu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \qquad \nu_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix} \qquad \boldsymbol{\Phi} = \begin{pmatrix} 0 & 0.4 \\ -0.3 & 0.8 \end{pmatrix} \qquad \boldsymbol{\Omega} = \begin{pmatrix} 1.36 & 1.2 \\ 1.2 & 1.36 \end{pmatrix}$$

where $s_t \in \{1,2\}$. The transition probability matrix is $\mathbf{P} = \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix}$ where $p_{11} = 0.8$ and $p_{22} = 0.1$. The unconditional probabilities are $\pi_1 = 0.8182$ and $\pi_2 = 0.1818$. We generate $T = 100$ observations of $\mathbf{u}_t$ by using Gaussian deviates with zero mean and variance $\boldsymbol{\Omega}$. Then the bivariate process $\mathbf{y}_t$ is obtained from the above model by using true parameters. Trend and cycle components of the two processes included in $\mathbf{y}_t$ are reported in Figure 1: the trend of the first process is more pronunced than the second (first row) and cycles are reported in the second row.

References

1. Alvarez, R., Camacho, M., Ruiz, M.: Inference on filtered and smoothed probabilities in Markov-switching autoregressive models. J. Bus. Econ. Stat. **37**(3), 484–495 (2019)
2. Anatolyev, S., Gospodinov, N.: Modeling financial return dynamics via decomposition. J. Bus. Econ. Stat. **28**, 232–245 (2010)
3. Anderson, H.M., Low, C.N., Snyder, R.D.: Single source of error state-space approach to the Beveridge-Nelson decomposition. Econ. Lett. **91**(1), 104–109 (2006)

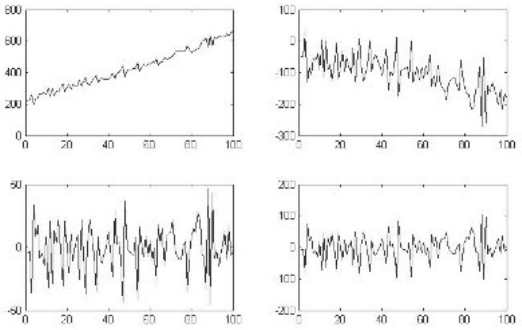


Fig. 1 First row: plots of trend components of the first and second process (from right to left). Second row: plots of cyclical components of the first and second process (from right to left)

4. Beveridge, S., Nelson, C.R.: A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. J. Monet. Econ. **7**, 151–174 (1981)
5. Cavicchioli, M.: Statistical analysis of mixture vector autoregressive models. Scand. J. Stat. **43**(4), 1192–1213 (2016)
6. Cavicchioli, M.: Higher order moments of Markov switching VARMA models. Econ. Theory **33**(6), 1502–1515 (2017)
7. Chen, C.C., Tsay, W.J.: The Beveridge-Nelson decomposition of Markov-switching processes. Econ. Lett. **91**(4), 83–89 (2006)
8. Gomez, V.: The use of Butterworth filters for trend and cycle estimation in economic time series. J. Bus. Econ. Stat. **19**(3), 365–373 (2001)
9. Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica **57**, 357–384 (1989)
10. Hamilton, J.D.: Analysis of time series subject to changes in regime. J. Econom. **45**, 39–70 (1990)
11. Harvey, A.C.: Trends and cycles in macroeconomic time series. J. Bus. Econ. Stat. **3**(3), 216–227 (1985)
12. Kim, C.J.: Markov–switching and the Beveridge–Nelson decomposition: Has US output persistence changed since 1984? J. Econom. **146**, 227–240 (2008)
13. Kim, C.J., Piger, J.M., Startz, R.: The dynamic relationship between permanent and transitory components of US business cycles. J. Money Credit. Bank. **39**(1), 187–204 (2007)
14. Knüppel, M.: Testing business cycle asymmetries based on autoregressions with a Markov-switching intercept. J. Bus. Econ. Stat. **27**(4), 544–552 (2009)
15. Krolzig, H.M.: Markov- Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin-Heidelberg-New York (1997)
16. Krolzig, H.M., Marcellino, M., Mizon, G.E.: A Markov-switching vector equilibrium correction model of the UK labour market. In: Advances in Markov-Switching Models (Hamilton, J.D., and Raj, B., eds.), Springer Verlag, Berlin-Heidelberg-New York (2002)
17. Morley, J.C., Piger, J.: Trend/cycle decomposition of regime-switching processes. J. Econom. **146**(2), 220–226 (2008)

Asymptotic properties of the SETAR parameters: a new approach

Un nuovo approccio per derivare le proprietà asintotiche dei parametri dei modelli SETAR

Marcella Niglio and Guy Mélard

Abstract A new method for obtaining the asymptotic properties of self-exciting threshold autoregressive models estimators is introduced. Like for most non-linear models, the usual method for obtaining the asymptotic properties for such models is based on the stationary and ergodicity of the model. A new method for obtaining these asymptotic properties was recently proposed by the authors for threshold models with exogenous threshold variable independent of the innovations. This method is based on the asymptotic theory for scalar or vector time-dependent ARMA (*td*ARMA) models, where the coefficients are deterministic functions of time and of a small number of parameters. These results cannot be directly extended to the self-exciting threshold models, where the coefficients are random, so we propose a different approach that can be adapted to this particular class of models.

Abstract *In questo contributo si presenta un nuovo approccio per esaminare le proprietà asintotiche del modello autoregressivo a soglia con variabile soglia endogena (SETAR). A differenza dei metodi usualmente impiegati in questo contesto, qui non si farà uso della stazionarietà ed ergodicità del processo. Il metodo è basato sull'uso di alcuni risultati recentemente presentati dagli autori per il modello a soglia con variabile soglia esogena ed è fondato sulla teoria asintotica sviluppata per i modelli ARMA con parametri dipendenti in modo deterministico dal tempo. La struttura del modello SETAR non consente il diretto uso di questi risultati che devono essere adattati al modello soglia in esame.*

Key words: Threshold models, time dependent coefficients, asymptotic properties.

Marcella Niglio
University of Salerno, DISES, Via Giovanni Paolo II, 132 - 84048 Fisciano (SA), Italy, e-mail: mniglio@unisa.it

Guy Mélard
Université Libre de Bruxelles, SBS-EM and ECARES, 50 avenue F.D. Roosevelt - B-1050 Bruxelles, Belgique, e-mail: guy.melard@ulb.be

1 Introduction

In this paper is proposed a new method to obtain the asymptotic properties of the autoregressive coefficients of the non-linear self-exciting threshold autoregressive (SETAR) model (among others, see [8], [9], [3]).

This model belongs to the wide class of threshold models that are non-linear structures with k regimes for a time series $x_t, t = 1, ..., n$, where the switching among them depends on the value assumed by an exogenous or endogenous (as the self-exciting case) threshold variable. If each regime follows a first-order autoregressive model, we deal with the so-called TAR(2; 1) model whose simplest form is:

$$x_t = \begin{cases} \phi^{(1)} x_{t-1} + \varepsilon_t, \text{ if } y_{t-d} \leq r, \\ \phi^{(2)} x_{t-1} + \varepsilon_t, \text{ if } y_{t-d} > r, \end{cases} \tag{1}$$

where $\phi^{(1)}$ and $\phi^{(2)}$ are the autoregressive coefficients, $\{\varepsilon_t\}$ is a sequence of independent random variables with mean zero and variance σ^2, d is the threshold delay, y_{t-d} is the threshold variable and r is the threshold value. When $y_{t-d} \equiv x_{t-d}$, the endogeneity of the threshold variable gives rise to the SETAR model.

In non-linear time series analysis the usual method for obtaining the asymptotic properties of such models consists in exhibiting a stationary and ergodic solution of the model and using ergodicity to prove the consistency and the asymptotic normality of the estimators. In this domain, [2] shows the consistency and the asymptotic distribution of the least squares estimators of the SETAR autoregressive coefficients and further investigates the consistency and limiting distribution of the estimator for the threshold r. A new approach has been recently proposed by [5, 6] that establish the asymptotic properties of the autoregressive coefficients of a TAR model with an exogenous threshold variable, y_{t-d}, independent of the innovations ε_t. These results are mainly based on the asymptotic theory developed for the ARMA models with time-dependent coefficients, also called tdARMA (see [1, 4]), where the autoregressive coefficients are deterministic functions of time and a few parameters. [5] give evidence that these results can be extended not only to TAR models but even to threshold models with ARMA regimes, shortly called TARMA models in [8], and that these extensions can be considered both in the univariate and multivariate domain (where each regime follows a Vector ARMA structure), provided that the threshold variable remains exogenous.

When the threshold variable is endogenous, $y_{t-d} \equiv x_{t-d}$, the autoregressive coefficients become random variables and then the results of [5] cannot be further used. Nevertheless we can adapt the results of [1] to the SETAR case and even to its SETARMA generalization, both in the scalar and in the vectorial form. There remains that the new approach requires assuming or proving the existence and the non-singularity of the information matrix.

To ease the presentation, we treat here only the SETAR case, leaving the SETARMA and SETVARMA (SETAR model with Vector ARMA regimes) cases for further paper.

2 TAR, SETAR and tdAR models

It will be shown that there is a strict relation among TAR, SETAR and tdAR models and that they can be connected to the same model structure.

Let $x_t \sim \text{TAR}(k;p)$ be a threshold model:

$$x_t = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \phi_j^{(i)} x_{t-j} \right] \mathbb{I}_{\{y_{t-d} \in \mathbb{R}_i\}} + \varepsilon_t, \tag{2}$$

where the autoregressive parameters $\phi_j^{(i)}$ are the object of our proposal, the innovations ε_t's have been defined before, $\mathbb{R}_i$ defines a partition of the real line $\mathbb{R}$, such that $\bigcup_{i=1}^{k} \mathbb{R}_i = \mathbb{R}$ and $\mathbb{R}_i \cap \mathbb{R}_s = \emptyset$, for $i \neq s$. When $y_{t-d} \equiv x_{t-d}$, we have a SETAR$(k;p)$ model.

The autoregressive model with time-dependent coefficients of order p, tdAR(p), is defined as:

$$x_t = \sum_{j=1}^{p} \phi_{tj} x_{t-j} + \varepsilon_t, \tag{3}$$

where the coefficients ϕ_{tj} are deterministic functions of time and of a vector of parameters β, and the innovations ε_t are like before.

By comparing the models (2) and (3), it can be noted that model (2) can be written in terms of (3) with $\phi_{tj}(\beta) = \sum_{i=1}^{k} \phi_j^{(i)} \mathbb{I}_{\{y_{t-d} \in \mathbb{R}_i\}}$, with β the vector of autoregressive parameters $\beta = (\phi_1^{(1)} \ldots \phi_p^{(1)} \phi_1^{(2)} \ldots \phi_p^{(k)})'$, where $\mathbf{A}'$ is the transpose of $\mathbf{A}$.

Further, let m be the number of the parameters included in β, $e_t(\beta)$ the residuals of the model and β_0 the true value of β, such that $e_t(\beta_0) = \varepsilon_t$, for all t.

Following the results of [5, 6], the SETAR model can be written as a tdVAR(1) model:

$$\mathbf{X}_t(\beta) = \mathbf{A}_t(\beta)\mathbf{X}_{t-1}(\beta) + \mathbf{E}_t(\beta), \tag{4}$$

with

$$\mathbf{A}_t(\beta) = \begin{bmatrix} \Phi_t(\beta) & \underset{p\times 1}{\mathbf{0}} \\ \underset{1\times p}{\mathbf{0}} & 0 \end{bmatrix}, \quad \Phi_t(\beta) = \begin{bmatrix} \phi_{t1}(\beta) & \ldots & \phi_{tp}(\beta) \\ & \underset{(p-1)}{\mathbf{I}} & \underset{(p-1)\times 1}{\mathbf{0}} \end{bmatrix},$$

and

$$\mathbf{X}_t(\beta) = \begin{bmatrix} X_t \\ \cdots \\ X_{t-p+1} \\ e_t(\beta) \end{bmatrix}, \quad \mathbf{E}_t(\beta) = \begin{bmatrix} e_t(\beta) \\ \underset{(p-1)}{\mathbf{0}} \\ e_t(\beta) \end{bmatrix},$$

then the moving average representation becomes (see [4]) :

$$\mathbf{X}_t(\beta) = \sum_{\ell=0}^{t-1} \Psi_{t\ell}(\beta)\mathbf{E}_{t-\ell}(\beta_0), \tag{5}$$

where $\Psi_{t\ell}(\beta) = \sum_{s=0}^{\ell} \left(\prod_{j=0}^{\ell-s-1} \mathbf{J}\mathbf{A}_{t-j}(\beta)\right) \mathbf{K} \left(\prod_{j=0}^{s-1} \mathbf{A}_{t-\ell+s-j}(\beta_0)\right)$, $\mathbf{J}$ and $\mathbf{K}$ are square matrices of dimension $(p+1)$, $\mathbf{J}$ is an identity matrix with $J_{1,1} = 0$ and $J_{(p+1),1} = -1$ and $\mathbf{K}$ is a null matrix with $K_{1,1} = K_{(p+1),1} = 1$.

Noting that $e_t(\beta)$ is the $(p+1)$-th element of $\mathbf{X}_t(\beta)$:

$$e_t(\beta) = \sum_{\ell=1}^{t-1} \psi_{t\ell}(\beta)\varepsilon_{t-\ell}, \tag{6}$$

with $\psi_{t\ell}(\beta) = \mathbf{U}'_{p+1}\Psi_{t\ell}(\beta)\mathbf{U}_1$, where $\mathbf{U}_1$ and $\mathbf{U}_{p+1}$ are two $[(p+1)\times 1]$ vectors of zeroes where the first and the $(p+1)$-th elements are replaced with 1, respectively.

Given $e_t(\beta)$ in (6), its first derivative with respect to β_i becomes:

$$\frac{\partial e_t(\beta)}{\partial \beta_i} = \sum_{\ell=1}^{t-1} \psi_{t\ell i}(\beta)\varepsilon_{t-\ell}, \tag{7}$$

where $\psi_{t\ell i} = \partial \psi_{t\ell}(\beta)/\partial \beta_i$, $i = 1, ..., m$.

Finally note that the theory requires also the second and the third derivatives of $e_t(\beta)$ that can be obtained likewise.

3 Parameters estimation

Assume that the time series x_t, for $t = 1, \ldots, n$, is a realization of the SETAR process (2), so with $y_t \equiv x_t$, and $\beta = \beta_0$. We estimate β using quasi maximum likelihood estimators that are obtained assuming that the innovations ε_t have a Gaussian distribution with $E[\varepsilon_t] = 0$ and $E[\varepsilon_t^2] = \sigma^2$.

Let $\widehat{\beta}_n$ be the estimator for β . We show that, under well-defined conditions, $\widehat{\beta}_n \to \beta_0$ almost surely for $n \to \infty$ and that $\sqrt{n}(\widehat{\beta}_n - \beta_0) \xrightarrow{L} N(\mathbf{0}, V^{-1})$, as $n \to \infty$.

The novelty of these results, if compared to those given in [2] where conditional least squares estimators are obtained for the parameters of the SETAR model, is that x_t is not assumed to be stationary and ergodic. This makes the use of the proposed inferential results more general.

4 Contribution

The results in [5, 6] deal with the estimation of the autoregressive coefficients of a threshold model with an exogenous threshold variable and then the theory given by [1] can be directly applied. On the contrary, in the SETAR case the coefficients $\psi_{t\ell}(\beta)$ of model (6) and consequently the coefficients $\psi_{ti\ell}(\beta)$ in (7) are random variables not independent from ε_t. This makes the SETAR model not only different from the tdAR model but even from the Random Coefficient Autoregressive (RCA)

model where the autoregressive coefficients are random variables also independent from the innovations (see [7]).

It implies that the assumptions considered for the tdAR model need to be changed to cope with the SETAR case. We here discuss only the assumptions that are critical for the SETAR model whereas for the other assumptions we refer to [4] and [6].

First, we need to keep two of these assumptions: the existence of the fourth moment of ε_t and the existence and the invertibility of the Fisher information matrix V, with elements V_{ij} given by $\lim_{n\to\infty} \frac{1}{n\sigma^2} \sum_{t=1}^{n} E_{\beta_0} \left(\frac{\partial \varepsilon_t(\beta)}{\partial \beta_i} \frac{\partial \varepsilon_t(\beta)}{\partial \beta_j} \right)$.

Further, it appears that we need to obtain an upper bound for quadratic forms like

$$E_{\beta_0}\{(\partial e_t(\beta)/\partial \beta_i)(\partial e_t(\beta)/\partial \beta_j)\}$$

(and even quartic forms, that will not be faced here). The main idea was to replace the derivatives making use of (7) and to transform the expression

$$\left| E\left(\sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} \varepsilon_{t-\ell_1} \psi_{ti\ell_1} \psi_{tj\ell_2} \varepsilon_{t-\ell_2} \right) \right|, \tag{8}$$

with $\psi_{ti\ell} = \psi_{ti\ell}(\beta_0)$, such that it is finite and fulfills the conditions given in [1, Technical Appendix Lemma 4.7].

To do this, we considered the existence of constants N and Φ, $0 < \Phi < 1$ such that $\sum_{k=\nu}^{t-1} \psi_{ti\ell}^2 < N\Phi^{\nu-1}$, for $\nu = 1,...,t-1$ and $i = 1,...,m$.

As said before, the random nature of the coefficients $\psi_{ti\ell}(\beta)$ in the SETAR case does not allow the use of the same arguments, so we need another way to bound (8). Denote $C_{tij\ell_1\ell_2} = \psi_{ti\ell_1}\psi_{tj\ell_2}$ and $D_{t\ell_1\ell_2} = \varepsilon_{t-\ell_2}\varepsilon_{t-\ell_1}$. Then:

$$E\left(\left| \sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} C_{tij\ell_1\ell_2} D_{t\ell_1\ell_2} \right| \right) \leq \sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} \{E(|C_{tij\ell_1\ell_2}|^2)\}^{1/2} \{E(|D_{t\ell_1\ell_2}|^2)\}^{1/2},$$

but $\{E(|D_{t\ell_1\ell_2}|^2)\}^{1/2} \leq \sigma^2$ and $|C_{tij\ell_1\ell_2}|^2 = \psi_{ti\ell_1}^2 \psi_{tj\ell_2}^2$.

We further assume that all the roots of all the autoregressive polynomials of the SETAR regimes $(1 - \phi_1^{(i)} z - ... - \phi_p^{(i)} z^p)$, $i = 1,...,k$, are in absolute value greater than 1 (this condition should not be confused with the stationarity of the SETAR model whose conditions are more restrictive).

If Φ exists, with $0 < \Phi < 1$, and if the matrix $\Phi_t(\beta)$ has all its eigenvalues less than Φ, it follows that the Frobenius norm of the coefficients $\Psi_{t\ell}$ included in (5) is almost surely bounded by Φ^{ℓ} up to a constant factor.

We can to verify that a similar bound exists for the coefficients $\psi_{ti\ell}$, $i = 1,...,m$, included in (7), implying the existence of an upper bound for $E(|C_{tij\ell_1\ell_2}|^2)$. Similar arguments can be used for the quadratic forms obtained from the other derivatives of $e_t(\beta)$ and for the other conditions given in [1].

For ARMA regimes, so for a univariate SETARMA model, a more general tdVAR(1) representation than in (4) is required. Also, it may not be obvious from the development given above, but it is possible to handle a vector of $e_t(\beta)$ and a matrix of coefficients $\psi_{ti\ell}$, making it possible to extend our results to vector self-exciting models, without any assumption of stationarity and ergodicity.

References

1. Alj, A., Azrak, R., Ley, C., Mélard, G.: Asymptotic properties of QML estimators for VARMA models with time-dependent coefficients. Scand. J Statist. **44**, 617–635 (2017)
2. Chan, K.S.: Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model, Ann. Stat., **21**, 520–533 (1993)
3. De Gooijer, J. G.: Elements of Nonlinear Time Series Analysis and Forecasting. Springer-Verlag, New York (2017)
4. Mélard G.: An indirect proof for the asymptotic properties of VARMA model estimators, Econometrics Statist. **21**, 96–111 (2022)
5. Mélard G., Niglio, M.: Link between threshold ARMA and tdARMA models. In: Perna C., Salvati N., Schirripa Spagnolo F. (eds) Book of Short Papers of SIS2021, pp. 720–725. Pearson (2021)
6. Mélard G., Niglio, M.: Another approach for asymptotic properties of threshold models. Manuscript in preparation (2022)
7. Nicholls, D.F., Quinn, B.G.:Random Coefficient Autoregressive Models. An introduction. Lecture Notes in Statistics. Springer-Verlag, New York (1982)
8. Tong, H.: Threshold Models in Non-linear Time Series Analysis, Springer-Verlag, New York (1983)
9. Tong, H.: Non-linear Time Series: A Dynamical System Approach, Oxford University Press, Oxford (1990)

Food prices forecast using post-sampled crowdsourced data with Reg-ARMA model: the case of Nigeria

Previsione dei prezzi del cibo usando dati post-sampled crowdsourced con un modello Reg-ARMA: il caso della Nigeria

Ilaria Lucrezia Amerise, Gloria Solano Hermosilla, Vincenzo Nardelli and Giuseppe Arbia

Abstract The current Sars-Cov-2 epidemic has once more highlighted the need of accurate and timely food price information for decision makers, citizen and food supply chain stakeholders, especially in developing countries and in particular in Africa. The conventional methods adopted by official statistical offices do not allow the dissemination of such information in real time. Therefore, we used data collected by the JRC project Food Price Crowdsourcing Africa in Nigeria. In the first stage, we pre-processed data using the Spatial Post Sampling approach, and we then estimated a Reg-ARMA model to provide an accurate forecast of price trends.

Abstract *L'attuale epidemia da Sars-Cov-2 ha evidenziato ancora una volta la necessità di informazioni accurate e tempestive sui prezzi degli alimenti sia per i decisori ed i cittadini sia per le parti interessate della filiera alimentare, in particolare nei paesi in via di sviluppo e segnatamente in Africa. I metodi convenzionali adottati dagli uffici statistici ufficiali non consentono la diffusione di tali informazioni in tempo reale. Abbiamo utilizzato, dunque, i dati raccolti dal progetto Food Price Crowdsourcing Africa del JRC in Nigeria. Nella prima fase, abbiamo pre-elaborato i dati utilizzando l'approccio del Spatial Post Sampling, quindi abbiamo stimato un modello Reg-ARMA per fornire una previsione accurata dell'andamento dei prezzi.*

Key words: Crowdsourcing, Food price Forecasting, Time-series, Reg-ARMA

Ilaria Lucrezia Amerise
University of Calabria - Rende (CS) - Italy e-mail: ilaria.amerise@unical.it

Gloria Solano Hermosilla
European Commission Joint Research Centre (JRC) - Seville - Spain e-mail: Gloria.SOLANO-HERMOSILLA@ec.europa.eu

Vincenzo Nardelli
Università degli Studi di Milano - Bicocca - Milano - Italy e-mail: v.nardelli2@campus.unimib.it

Giuseppe Arbia
Università Cattolica del Sacro Cuore - Roma - Italy e-mail: giuseppe.arbia@unicatt.it

1 Introduction

Unexpected food price fluctuations can affect the poor and vulnerable in Sub-Saharan Africa, where incomes are low and poverty rates are high. The COVID-19 pandemic has once again underlined the need to predict changes for policymakers planning to adjust food security initiatives, and for food supply chain players [1]. Notably, decisions are made according to price expectations. However, because of delays in the publication of official statistics, the full situation will not be available until some time after the period's end. Nowadays, mobile phone technologies and alternative data collection methods such as crowdsourcing, offer great potential for timely data collection, even in remote areas, to complement official statistics [2] and improve forecasts [3]. As such, it involves a large group of non-professionals volunteers (the crowd) at a lower cost or to a greater extent than traditional, professional data collection methods [4].Two key challenges arise: (i) involving a crowd providing a consistent data flow [5] and (ii) the quality of their inputs, which might be lower than that of professionals [4]. In terms of quality, it is essential that the crowdsourcing tool incorporates automatic quality checks for which the temporal/spatial dimension of the data has proved to be helpful [6]. In this context, the aim of this paper is twofold: firstly, to assess if the use of crowdsourced prices can help producing accurate forecasts, and secondly, to assess if the spatial post-sampled data (SPS prices) may lead to a more accurate prediction than the average of raw data (raw prices). In the remainders of this paper, Section 2 describes the data, Section 3 introduces the time-series forecasting methods, Section 4 includes the empirical application and Section 5 presents the discussion and conclusions.

2 Data

We use data collected from April to December 2021 from Nigeria's European Commission Joint Research Centre FPCA platform. A crowd of 1300 volunteers using a mobile app submitted in Kano, Katsina, Kaduna, and Lagos states daily market prices of several food commodities for a reward given to a maximum of 30 individuals per day. Volunteers owned a smartphone with GPS and the system was built based on Open Data Kit and deployed on a compatible server. Automatic routines (implemented in an R) were in place to process, quality check and validate data, which fed into a daily updated web dashboard [7]. In a first step, based on the auto-recorded time and geo-location, prices were compared to others submitted during the same week within a 12 km radius (cluster). Multiple close price points (in time and space) are expected to be correlated so this allow to filter out spurious data points and confining price values to reasonable ranges. Isolated points (prices that did not find a price in the given neighbourhood) were relocated to a neighbouring cluster, if available, or discarded. In a second step, the data were reweighted to resemble a formal spatial sampling design with the spatial post-sampling procedure [6], attenuating potential effects of sampling bias inherent to the voluntary nature of

crowdsourcing and the non-random distribution of smartphone owners and digital competence. Prices collected in Kano from April 19th to October 3rd act as training period and prices from October 4th to December 5th as validation set. The crowd submitted 10058 garri (derived from cassava) price observations during the period, translating into 6184 valid prices and aggregated in 33 weekly observations.

3 Methods

3.1 Assessing point forecast accuracy

The main problem with assessing the reliability of forecasts is that the magnitude of forecast errors cannot be evaluated until the actual values have been observed. To simulate such a situation, we split the time series into two parts: the "training" period, which ignores a number of the most recent time points and the "validation" period, which comprises only the ignored time points. Eight weeks of observations ($H = 8$) are set-aside to serve as a benchmark for forecasting purposes. There are a number of indices that assess predictive accuracy. For our current purposes, we choose the relative absolute error of forecast (RAEF), which is an index that varies in a fixed interval and makes good use of the observed residuals. Its formal expression is the following:

$$RAEF = 100\left[1 - H^{-1}\sum_{h=1}^{H} A_{n+h} - \widehat{A}_{n+h}A_{n+h} + \widehat{A}_{n+h} + \varepsilon\right] \quad h = 1,2,\cdots,H \quad t = 1,2,\cdots,n \tag{1}$$

where ε is a small positive number (e.g. $\varepsilon = 0.00001$) which constitutes a safeguard against division by zero. Coefficient (1) is independent on the scale of the data and, due to the triangolar inequality, ranges between zero and 100. The maximum is achieved in the case of perfect forecasts: $L_{n+h} = \widehat{L}_{n+h}$ for each h. The lower the $RAEF$ is, the less accurate the model is. The minimum stands for situations of inadequate forecasting such as $\widehat{H}_{n+h} = 0$ or $\widehat{L}_{n+h} = -L_{n+h}$ for all h.

3.2 Simultaneous prediction intervals

When forecasting pricing for marketing, planning, development, and policy-making, it is critical to not only create point predictions for several steps ahead, but also to provide an evaluation of the uncertainty associated with projections. Thus, the challenge is to combine point forecasts with prediction intervals (PIs) which apply simultaneously to all possible future values of the predictors. There is an extensive literature on simultaneous prediction intervals for autoregressive time series (see, for example, [8] and reference therein).

A reasonable strategy can be as follows. Given the availability of H future values, we can construct two bands such that, under the condition of independent Gaussian distributed random residuals, the probability that consecutive future prices $A_{n+h}, h = 1,2,\cdots,H$ lie simultaneously within their respective range is at least is γ

$$P\left[\bigcap_{h=1}^{H}\left(A_{1,h,\gamma} \leq A_{n+h} \leq A_{2,h,\gamma}\right)\right] \geq \gamma \tag{2}$$

where

$$\left\{A_{1,h,\gamma} = \widehat{A}_{n,h} - \theta_{H,\nu,\gamma}\widehat{\sigma}_h A_{2,h,\gamma} = \widehat{A}_{n+h} + \theta_{H,\nu,\gamma}\widehat{\sigma}_h.\right.$$

(3)

The multiplier $\theta_{H,\nu,\gamma}$ is the γ-th quantile of the of the maximum absolute value t of the H-variate Student t probability density function with ν degrees of freedom. See [9]. In short, $\theta_{H,\nu,\gamma}$ is the solution of

$$\int_{-\theta}^{\theta}\int_{-\theta}^{\theta}\cdots\int_{-\theta}^{\theta} f(t_1,t_2,\cdots,t_h;\nu)\,dt_1 dt_2\cdots dt_H = \gamma \tag{4}$$

The critical values can be found solving iteratively $u_{H,\nu,\gamma}$ using the command *pmvt* of the *R* package *mvtnorm*, which provides the multivariate t probability. See [10].

The most important characteristic of PIs is their actual coverage probability (PIAC). We measure PIAC by the proportion of true prices of the validation period enclosed in the bounds

$$PIAC_\gamma = 100H^{-1}\sum_{h=1}^{H} c_{h,\gamma} \quad \text{where } c_{h,\gamma} = \left\{1 \quad \text{if } A_{n+h} \in \left[A_{1,h,\gamma}, \quad A_{2,h,\gamma}\right] 0 \quad \text{otherwise}\right. \tag{5}$$

If $PIAC_\gamma \geq \gamma$ then prices tend to be covered by the constructed bounds, but this may also imply that the estimates of the variances in the forecast errors are positively biased. A $PIAC_\gamma < \gamma$ indicates under-dispersed forecast errors with overly narrow prediction intervals and unsatisfactory coverage behavior.

Ceteris paribus, narrow PIs are desirable as they reduce the uncertainty associated with forecasts. However, high accuracy can be easily obtained by widening PIs. A complementary measure that quantifies the sharpness of PIs might be useful in this context. Here, we use the score function.

$$R_{h,\gamma} = (\gamma 2)\left(C^2_{h,\gamma} - C^1_{h,\gamma}\right)A_{n+h}, \quad h = 1,2,\cdots,H. \tag{6}$$

This expression reflects a penalty proportional to the narrowness of the intervals that encompass the true values at the nominal rate. The penalty increases as γ decreases, to compensate for the tendency of prediction bands to be broader as the confidence level increases. Of course, the lower $R_{h,\gamma}$ is, the more accurate PI will

be. The average value of the score width across time points

$$ASW_{\gamma} = 1H \sum_{h=1}^{H} R_{h,\gamma} \tag{7}$$

can provide general indications of PIs performance.

4 Empirical analysis

Table 1 shows the results obtained with the Reg-ARMA method applied to the data of the training period for the prices in the validation period.

The Reg-ARMA approach offers several improvements. The temporal autocorrelation is almost inexistent because the p-value of the LB statistics is now larger that 99%. Also, the quality of the fitting is increased as proven by a lower AICc and a higher $\bar{R}^2$.

Table 1 Reg-ARMA estimation and forecasting.

		β_0	t	$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{1,2}$	$\gamma_{2,3}$	$\gamma_{1,5}$	RAEF	PIAC	ASW
Raw prices											
	Reg-A	419.9	374.6	-135.3	-175.4	278.8	219.8	157.0			
	$\bar{R}^2$:	0.879	AICc:	207.1	LB:	7.7			72.1%	65.5%	11.6
SPS prices											
	Reg-A	415.2	390	-169.9		284.7	-287.1				
	$\bar{R}^2$:	0.981	AICc:	195.4	LB:	6.7			89.7%	99.0%	3.1

In both cases the improvement is quite substantial. In addition, as can be readily seen from Figure 1, the coverage rate is now significantly higher compared to raw prices. The cost of these enhancements is a larger width of the simultaneous forecast intervals, which, as it is well known, determine broader brackets. Furthermore, the stability of the RAEF index across the estimation methods of the two time series, is a demonstration that the predictive accuracy does not deteriorate when Reg-ARMA method is used. Nonetheless, we must remark that, the Reg-ARMA method, yields prediction intervals whose actual coverage rate resulted to be less than the nominal level (95%). We explain this result as due to a change in the evolutionary trajectory of prices, which in the last year, has not been following the trends of previous years. This is an obvious consequence of the need of bracketing future values within the same scheme used for past observations. Thus a constraint is imposed on the forecasting tool: strong local fluctuations or outliers cannot appear in the set of future values even if we know that they are there; thus, some failures are inevitable.

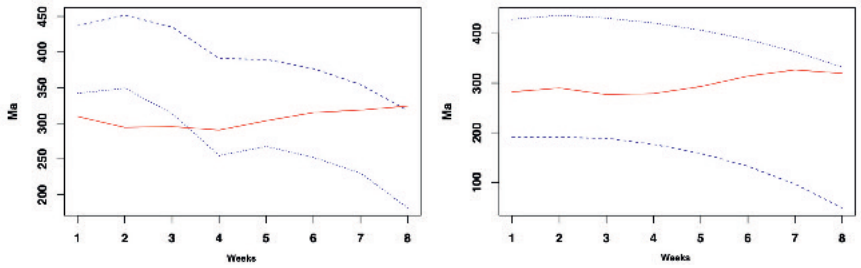


Fig. 1 Prediction intervals (PIs) of garri raw prices (left) and SPS prices (right) in Kano state (Nigeria)

5 Conclusion

Our work is an example of using crowdsourcing and quality control methods for accurate food price predictions to help governments and other actors in decision-making processes. The results show that the raw data are not sufficient to correctly forecast price trends. A two-stage procedure (spatial pre-processing and temporal modelling) solves most of these problems, ensuring greater reliability both in the data collected and in the predictions. Further research could exploit the combination of high-frequency crowdsourced prices with official prices released at lower frequencies also focusing on spatio-temporal modelling.

References

1. J. B. Adewopo, G. Solano-Hermosilla, L. Colen, and F. Micale, "Using crowd-sourced data for real-time monitoring of food prices during the covid-19 pandemic: Insights from a pilot project in northern nigeria," *Global Food Security*, vol. 29, p. 100523, 2021.
2. DGINS, "Scheveningen memorandum - big data and official statistics," 2013.
3. V. Yadav and A. Das, "Nowcasting inflation in india with daily crowd-sourced prices using dynamic factors and mixed frequency models," *Applied Economics Letters*, pp. 1–11, 2021.
4. J. Minet, Y. Curnel, and j. v. p. y. p. Gobin, Anne et al, "Crowdsourcing for agricultural applications: A review of uses and opportunities for a farmsourcing approach,"
5. G. Solano-Hermosilla, J. Barreiro-Hurle, J. B. Adewopo, and C. Gorrín-González, "Increasing engagement in price crowdsourcing initiatives: Using nudges in nigeria," *World Development*, vol. 152, p. 105818, 2022.
6. G. Arbia, G. Solano-Hermosilla, F. Micale, V. Nardelli, and G. Genovese, "Post-sampling crowdsourced data to allow reliable statistical inference: the case of food price indices in nigeria," *arXiv preprint arXiv:2003.12542*, 2020.
7. JRC, "Food price crowdsourcing africa platform - datam," 2018.
8. N. Ravishanker, L. S.-y. Wu, and J. Glaz, "Multiple prediction intervals for time series: comparison of simultaneous and marginal intervals," *Journal of Forecasting*, vol. 10, no. 5, pp. 445–463, 1991.
9. G. J. Hahn, "Simultaneous prediction intervals for a regression model," *Technometrics*, vol. 14, no. 1, pp. 203–214, 1972.
10. A. Genz, "Multivariate normal and t distribution. r package version 1.0-11," 2019.

Universal change point testing for dependent data

Test d'ipotesi universale per punti di cambio in dati dipendenti

Federica Spoto, Alessia Caponera and Pierpaolo Brutti

Abstract One of the main interests in time series analysis is the detection of the so called change-points, defined as timestamps where the model parameters experience a substantial shift in value. Once a candidate change-point is identified, we may want to test whether there is a significant difference in distribution before and after the structural break. In this work we approach the problem from a split-sample perspective and we implement and test on both simulated and real data a two-sample test for time dependent streams that we call *universal change-point testing.*

Abstract *Uno dei problemi principali nell'analisi di serie storiche è l'individuazione dei cosiddetti punti di cambio (change-points), definiti come quegli istanti temporali in cui i parametri del modello vanno incontro ad una sostanziale variazione. Più specificatamente, una volta individuato un valore candidato per il punto di cambio, ci possiamo chiedere se c'è sufficiente evidenza sperimentale a favore dell'esistenza di una differenza significativa nella distribuzione dei dati prima e dopo tale valore. In questo lavoro affrontiamo il problema con delle opportune tecniche di suddivisione del campione, implementando e testando su dati, sia simulati che reali, un test a due campioni per serie dipendenti che chiameremo "test universale per il punto di cambio".*

Key words: Universal Inference, Change-point detection, Likelihood ratio test, Time series model

Federica Spoto
Sapienza University of Rome, e-mail: federica.spoto@uniroma1.it

Alessia Caponera
École Polytechnique Fédérale de Lausanne, e-mail: alessia.caponera@epfl.ch

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

In temporal processes analysis, one of the main interests is the detection and the analysis of the so called change-points, defined as timestamps where the model parameters have a change in value.

Once a change-point is given as an endogenous information, one may want to test whether there is a significant difference in distribution before and after the structural break. To do so, in this work we approach the problem from the split-sample perspective discussed in [5] (mostly for i.i.d samples) and we implement a two-sample test for time dependent streams of data. We refer to such procedure as *universal change-point testing* and the key steps are formally described in Section 2 for the case of an autoregressive change-point model.

This basic procedure can also be seen as the stepping stone for building a more structured *sequential change-point detection*, which finds important applications for instance in industrial quality control, environment surveillance, computer network security (see among others [4]).

2 Materials and Methods

In this section, we introduce the autoregressive change-point model and the main steps to build the universal change-point testing procedure in this context. For the sake of simplicity, our arguments are presented for a AR(1) model, allowing a single change-point; however, the analysis can be generalized to higher autoregressive orders and multiple change-points. In this setting, the model is written as the composition of two stationary AR segments and takes the form

$$X_t = \begin{cases} \phi_1 X_{t-1} + \varepsilon_{1t} & t \leq \tau \\ \phi_2 X_{t-1} + \varepsilon_{2t} & t > \tau \end{cases}, \tag{1}$$

that, given τ, are assumed to be independent. We consider centered X_t's, but clearly one can allows for changes also in the mean, thus including two different intercepts in the model. We also assume that the errors ε_{jt} are Gaussian with mean zero variance σ_j^2, $j = 1,2$.

We assume to be able to observe a finite stretch $\{X_1,\dots,X_N\}$ and to know that a change-point occurred at $1 < \tau < N$. Our goal is to test whether there is a significative difference among the set of parameters before and after the structural break. Hence, formally, define $\theta_1 = (\phi_1, \sigma_1^2), \theta_2 = (\phi_2, \sigma_2^2) \in (-1,1) \times (0,\infty)$ and consider the following set of hypotheses:

$$\begin{cases} H_0 : \theta_1 = \theta_2 \\ H_1 : \theta_1 \neq \theta_2 \end{cases}.$$

Then, split $X_1,\dots,X_\tau$ into two segments A_0, A_1, and $X_{\tau+1},\dots,X_N$ into other two segments B_0, B_1, that is,

$$A_1 = (X_1,\dots,X_\gamma), \qquad A_0 = (X_{\gamma+1},\dots,X_\tau),$$
$$B_1 = (X_{\tau+1},\dots,X_{\tau+\gamma}), \qquad B_0 = (X_{\tau+\gamma+1},\dots,X_N),$$

for some integer $1 < \gamma < \tau$. We can hence define $D_0(\tau) = A_0 \cup B_0$, $D_1(\tau) = A_1 \cup B_1$ and the *conditional likelihood*

$$L_{0|1}(\theta_1,\theta_2) = p_{\theta_1,\theta_2}(D_0|D1) = p_{\theta_1,\theta_2}(B_0|B_1)p_{\theta_1,\theta_2}(A_0|A_1).$$

In particular for the autoregressive model (1), we have

$$p_{\theta_1}(A_0|A_1) = p_{\theta_1}(X_{\gamma+1},\dots,X_\tau|X_1,\dots,X_\gamma) = \prod_{t=\gamma+1}^{\tau} p_{\theta_1}(X_t|X_{t-1})$$
$$= \prod_{t=\gamma+1}^{\tau} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{1}{2}\frac{(X_t-\phi_1 X_{t-1})^2}{\sigma_1^2}\right\},$$

and similarly,

$$p_{\theta_2}(B_0|B_1) = p_{\theta_2}(X_{\tau+\gamma+1},\dots,X_N|X_{\tau+1},\dots,X_{\tau+\gamma}) = \prod_{t=\tau+\gamma+1}^{N} p_{\theta_2}(X_t|X_{t-1}),$$
$$= \prod_{t=\tau+\gamma+1}^{N} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{1}{2}\frac{(X_t-\phi_2 X_{t-1})^2}{\sigma_2^2}\right\}.$$

Under H_0, $\theta_1 = \theta_2 = \theta$ and we can compute the MLE of θ based on D_0 as

$$\widehat{\theta} = \arg\max_{\theta} L_{0|1}(\theta,\theta).$$

Let then $\widehat{\theta}_1, \widehat{\theta}_2$ be any estimators based on D_1 (under the alternative hypothesis).

$$\text{Reject } H_0 \text{ if:} \quad \frac{L_{0|1}(\widehat{\theta}_1,\widehat{\theta}_2)}{L_{0|1}(\widehat{\theta},\widehat{\theta})} > \frac{1}{\alpha}.$$

This is essentially a modified version of the usual likelihood ratio statistic with an *out-of-sample* estimator in the numerator and a "universal" threshold, in the sense that it does not rely on approximations based on limiting distributions, but ensures finite sample guarantees (without additional regularity conditions). Indeed, it is possible to show that the Type-I error can be (conservatively) controlled at level α.

We stress that this method is general and can be applied to different contexts, as long as the conditional likelihood can be computed. The idea is then to extend the result to more general dependency structures, and also to implement a multiple

testing procedure that jointly test different values of τ, to then obtain a complete detection procedure.

The present framework could be also generalized to the so-called *spherical functional autoregressive model*, defined for collections of time dependent random fields on a spherical domain (see for instance [1]). In this setting, an alternative change-point analysis is discussed in [3]. All these topics are the object of current ongoing research.

3 Results

We tested the performance of our proposal on synthetic time series and then on a real data-set collecting the *wave heights* in the East Scotian Slope, Canada.

The simulation study has been carried out on two different scenarios to explore *size* and *power* of the universal LRT. In the first scenario we work under the null with $\theta_1 = \theta_2 = (\phi = 0.4, \sigma^2 = 0.3)$, and the observations are generated from the following AR(1) model:

$$X_t = 0.4 \cdot X_{t-1} + \varepsilon_t \qquad 1 \leq t \leq N,$$

where ε_t are Gaussian with zero mean and variance $\sigma^2 = 0.3$. We evaluated the Type I error probability based on $M = 10{,}000$ runs with increasing sample size N and $\alpha = 5\%$. In this scenario $\tau = \frac{N}{2}$ and $\gamma = \frac{\tau}{2}$. The results are shown in Table 1.

N	Test Size
100	0.0009
500	0.0004
1000	0.0001

Table 1: Type I error probability at different sample sizes.

The second scenario follows the alternative hypothesis $\theta_1 \neq \theta_2$. We considered three different settings, increasing the distance between the parameters vectors $\theta_1 = (\phi_1, \sigma_1^2)$ and $\theta_2 = (\phi_2, \sigma_2^2)$. More specifically, we kept the model parameters of the first segment fixed at $\theta_1 = (\phi_1 = 0.4, \sigma_2^2 = 0.2)$ while varying the parameters of the second segment from $\theta_2 = (\phi_2 = 0.6, \sigma_2^2 = 0.3)$ in *Setting 1*, to $\theta_2 = (\phi_2 = 0.8, \sigma_2^2 = 0.4)$ in *Setting 2*, and finally $\theta_2 = (\phi_2 = -0.8, \sigma_1^2 = 0.5)$ in *Setting 3*. For each setting we generated $N \in \{100, 500, 1000\}$ observations with $\tau = \frac{N}{2}$, and $\gamma = \frac{\tau}{2}$. Each setting has been simulated $M = 10{,}000$ times and we evaluated the power of the test with $\alpha = 5\%$. Table 2 contains the results.

As anticipated, we also applied our universal change-point test on the publicly available time series of wave heights collected by Fisheries and Oceans

	N	Power
Setting 1	100	0.010
	500	0.297
	1000	0.696
Setting 2	100	0.117
	500	0.907
	1000	0.996
Setting 3	100	0.781
	500	0.999
	1000	1.000

Table 2: Power of the test for different settings.

Canada (more specifically by the East Scotian Slope buoy), and contained in the `changepoint R` package (see the Github repository). The observations were taken at hourly intervals from January 2005 until September 2012. Here we focus only on the period January - September 2005. The detection of change-points in this time series is helpful to get a better understanding of the variability of the ocean in a certain period of the year, a crucial information for planning operations on offshore infrastructures whose risk of failure strongly increase in the presence of larger wave heights. Quite understandingly, we expect a transition point when moving fro winter to summer, but its exact timing is unknown.

This dataset has been previously used by [2] to test the performance of their algorithm. The analysis there focused on the first order difference of the original data and, between January and September 2005, they detected a change-point at the beginning of April 2005 (dashed blue line in Fig. 1), a result that is consistent with the seasonal behaviour of wave heights. As a confirmatory step, we applied our technique to test the presence of a significant change in the data by setting $\tau = \{\text{April } 1^{\text{st}}\}$. On the log-scale, the rejection rule is then:

$$\text{Reject } H_0 \text{ if:} \quad \log\left(L_{0|1}(\widehat{\theta}_1, \widehat{\theta}_2)\right) - \log\left(L_{0|1}(\widehat{\theta}, \widehat{\theta})\right) > -\log(\alpha).$$

The test confirmed the detection in early April, rejecting the null hypothesis and estimating the model parameters as $\widehat{\theta}_1 = (\widehat{\phi}_1 = 0.012, \widehat{\sigma}_1^2 = 0.048)$, and $\widehat{\theta}_2 = (\widehat{\phi}_2 = 0.029, \widehat{\sigma}_2^2 = 0.019)$

References

1. Caponera, A., Marinucci, D.: Asymptotics for spherical functional autoregressions. Annals of Statistics **49**(1), 346–369 (2021)
2. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association **107**(500), 1590–1598 (2012). DOI 10.1080/01621459.2012.737745. URL `http://dx.doi.org/10.1080/01621459.2012.737745`

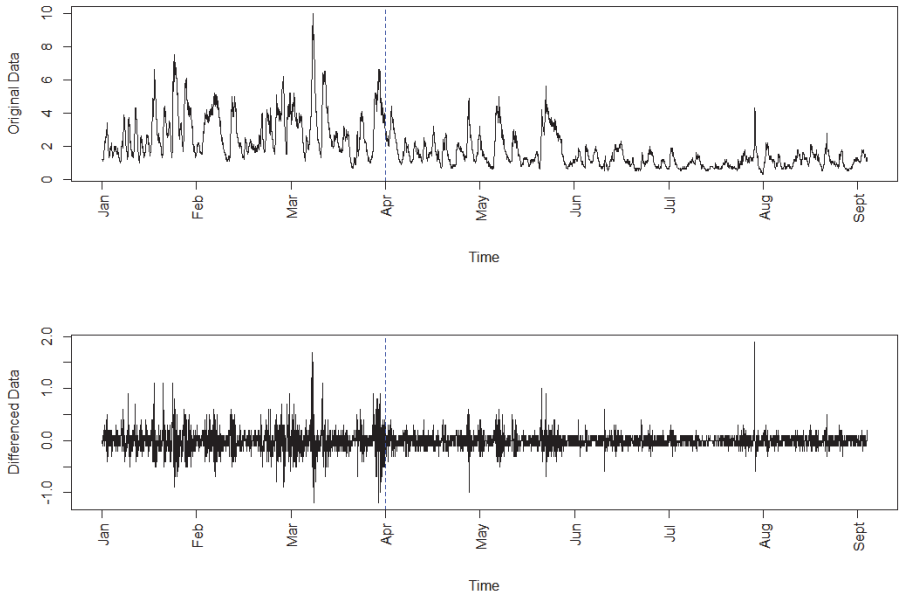


Fig. 1: Top: Original North Atlantic Wave Heights. Bottom: Differenced North Atlantic Wave Heights.

3. Spoto, F., Caponera, A., Brutti, P.: Spherical autoregressive change-point detection with applications. In: Book of Short Papers SIS 2021 (2021). ISBN: 9788891927361
4. Tartakovsky, A.: Sequential Change Detection and Hypothesis Testing: General Non-i.i.d. Stochastic Models and Asymptotically Optimal Rules (1st ed.). Chapman and Hall/CRC (2019)
5. Wasserman, L., Ramdas, A., Balakrishnan, S.: Universal inference. Proceedings of the National Academy of Sciences **117**(29), 16,880–16,890 (2020). DOI 10.1073/pnas.1922664117. URL https://www.pnas.org/content/117/29/16880

Change point detection in fruit bioimpedance using a three-way panel model

Rilevamento del punto di cambiamento nella bioimpedenza della frutta utilizzando un modello panel a tre vie

F. Marta L. Di Lascio and Selene Perazzini

Abstract We propose a three-way dynamic threshold panel model defined without time-varying covariates. The model is thought to detect changes in fruit ripeness observed at several frequencies of electrical impedance spectroscopy. The transition variable is the lagged dependent variable, which is fruit bioimpedance, and the threshold parameter varies by the level of the third way, which is the electrical frequency of the spectroscope. Consistently with the applied aim, we assume cross-independence in the third way. We apply the model to an innovative dataset containing bioimpedance measurements on on-plant strawberries. Our findings suggest that there is a physio-chemical change in the observed strawberries during days 6-7 of the maturation process irrespective of the electrical frequency.

Abstract *Proponiamo un modello panel a soglia dinamica a tre vie definito senza covariate variabili nel tempo. Il modello è pensato per rilevare i cambiamenti di maturità della frutta osservati a diverse frequenze della spettroscopia di impedenza elettrica. La variabile di transizione è la variabile dipendente ritardata, che è la bioimpedenza della frutta, e il parametro di soglia varia con il valore della terza via, che è la frequenza elettrica dello spettroscopio. Coerentemente con l'obiettivo applicato, assumiamo l'indipendenza trasversale nella terza via. Applichiamo il modello a un set di dati innovativo contenente misure di bioimpedenza relativa a fragole su pianta. I nostri risultati suggeriscono che c'è un cambiamento fisico-chimico nelle fragole osservate durante i giorni 6-7 del processo di maturazione, indipendentemente dalla frequenza elettrica.*

Key words: Bioimpedance, Change point detection, IV-GMM estimator, Three-way panel model, Threshold model

F. Marta L. Di Lascio
Faculty of Economics and Management, Free University of Bozen-Bolzano, e-mail: marta.dilascio@unibz.it

Selene Perazzini
Faculty of Economics and Management, Free University of Bozen-Bolzano, e-mail: selene.perazzini@unibz.it

1 Introduction

The reliable assessment of fruit quality is extremely important in order to meet production and consumer demands, and at the same time drastically reduce food waste. Bioimpedance analysis for fruit quality control is currently severely limited by the lack of an accurate method for modeling fruit ripeness. For these reasons, innovative and reliable statistical methods able to assess fruit physio-chemical changes from the bioimpedance response are needed.

Change point detection methods [1] for panel data are suitable for this kind of issues as the bioimpedance of the fruits is detected over time using a bench-top impedance analyser set at different frequencies. Hence, [2] proposed a three-way dynamic threshold panel model as change detection method by extending the work by [5] to the three-way framework and defining a criterion to detect the change point. Here we consider a particular version of the model proposed by [2] and we show its performance in assessing fruit ripeness.

In summary, in Sect. 2 we describe the fruit bioimpedance data to which we applied the three-way dynamic threshold panel model described in Sect. 3. Empirical findings are shown in Sect. 4, while conclusions are presented in Sect. 5.

2 Fruit bioimpedance data

Bioimpedance typically refers to the measurement of the electrical impedance of a biological tissue that is a physical quantity describing the ability of an object to oppose, or impede, a flow of electrical current caused by an external voltage [4]. Biological tissues are composed of complex liquids containing structures separated by diverse membranes, through which the electric current can travel depending on the frequency of the applied signal. Therefore, bioimpedance measurements are dependent on the electrical frequency. For this reason, in order to assess the quality of a fruit, multiple frequencies should be considered.

The innovative dataset used for the proposed analysis has been collected during the development of an interdisciplinary research project between the Faculty of Economics and Management and the Faculty of Science and Technology of the Free University of Bozen-Bolzano by the Sensing Technologies Laboratory. The original dataset contains bioimpedance measurements on 99 on-plant strawberries observed twice a day at frequencies ranging from 20 to 13668764.956 Hz. Data have been collected from the fruiting of the plants until fruit deterioration. Since strawberries ripened at different times, the dataset constitutes an unbalanced panel where the three ways are given by the strawberry $i = 1,\dots,n$, the electrical frequency $j = 1,\dots,J$, the time of observation $t = 1,\dots,T$.

Data have been pre-processed in order to increase the data quality. First of all, a huge variability in the bioimpedance of the strawberries has been observed during the first 4 times of observation. This may trace back to data collection: bioimpedance is measured by means of electrodes placed on the surface of the fruit, and the small-

est volumes might have affected the measurements. Therefore, for all fruits, the first 4 time observations, i.e. the first two days of observation, have been discarded. Then, we selected the subset of strawberries with similar time to deterioration. We thus restricted the analysis to 61 strawberries, for which at least 16 and at most 30 times of observation were available. Finally, no missing data or outliers were detected.

Preliminary analyses showed very similar bioimpedance patterns at close frequencies. Therefore, sufficiently different frequencies should be considered. Furthermore, bioimpedance time series at low frequencies appear highly random, and tend to flatten at higher frequencies. Neither one of the two extreme cases is suitable to change point detection. As far as the observed strawberries concern, we identified the following 9 electrical frequencies that well capture regime switches in the fruit maturation process: 8690.783, 23635.120, 74152.869, 201663.287, 548435.709, 1043389.435, 1720663.222, 2290015.458, 3047761.311 Hz, and restricted the analysis to them. The final dataset we analyse is therefore a balanced panel with $i = 1,\dots,61$, $t = 5,\dots,16$, and $j = 1,\dots,9$.

3 Three-way dynamic panel threshold model

Di Lascio and Perazzini [2] developed a three-way dynamic panel threshold model with the aim of detecting change points. Here we present their proposal in a simplified version that do not use any time-varying covariate. Hence, we work with the following three-way dynamic panel threshold model

$$y_{ijt} = \phi_{1j}\mathbb{1}\left\{y_{ij(t-1)} \leq \gamma_j\right\} + \phi_{2j}\mathbb{1}\left\{y_{ij(t-1)} > \gamma_j\right\} + \varepsilon_{ijt} \tag{1}$$

where y_{ijt} is the value of variable Y observed on the i-th statistical unit at time t for the level j of the considered factor, and ϕ_{1j} and ϕ_{2j} are the two slope parameters associated, respectively, to the lower and upper regimes of the j-th category. Moreover, $\mathbb{1}$ is the indicator function that captures the change in regime and is defined by the threshold parameter γ_j, which depends on the j-th category, the lagged dependent variable $y_{ij(t-1)}$ is the transition variable, and ε_{ijt} is the error term. The latter is defined as the sum of three components

$$\varepsilon_{ijt} = \mu_i + \lambda_t + \nu_{ijt} \tag{2}$$

where μ_i is the unobserved individual fixed effect, λ_t is the unobserved time fixed effect, and ν_{ijt} is a zero mean random error.

We assume that the J levels of the third way are independent to each others and they represent the categories of a variable. This is due to the nature of data analysed in Sect. 4 where observations on the same statistical unit by varying the third level, i.e. the frequencies of electrical impedance spectroscopy, do not have any effect each other, i.e. are independent.

The model in Eq. (1) is estimated through the generalized method of moments [3] based on the first-difference transformation and on the use of instrumental variables

(IV-GMM). This estimation method has been developed by [2] extending the one in [5] to the three-way model case. The extensive Monte Carlo study in [2] proved the ability of the three-way dynamic threshold panel model to both detect the time and the extent of the regime switch and provide reliable estimates. These results support the application presented here.

4 Findings on fruit bioimpedance

We apply the three-way dynamic panel threshold model in Eq. (1) to the strawberries bioimpedance data presented in Sect. 2. Observations at $t = 5, \dots, 7$ have been selected as instrumental variables and have therefore been excluded from the estimation of the thresholds. The IV-GMM estimates of the threshold parameters and of the difference between the slope parameters are shown in Tab. 1. Coherently with the spectroscopy, the estimated values of both the parameters of interest decrease as j increases. Fig. 1 shows the observed time series for each considered frequency j

Table 1: Estimation results: j indicates the used electrical frequency, $\hat{\gamma}_j$ is the IV-GMM estimate of the threshold parameter, $\hat{\delta}_j = \hat{\phi}_{2j} - \hat{\phi}_{1j}$ is the IV-GMM estimate of the difference between slope parameters associated with the two different regimes.

j	$\hat{\gamma}_j$	$\hat{\delta}_j$
8690.78	6263.79	1320.50
23635.12	3871.95	719.61
74152.87	2091.80	344.89
201663.29	1493.37	587.34
548435.71	1086.95	569.00
1043389.44	918.67	460.52
1720663.22	910.33	433.78
2290015.46	876.74	418.97
3047761.31	798.19	359.79

together with the corresponding threshold value estimated. As one can notice, the strawberries time series are highly heterogenous in terms of bioimpedance irrespective of the frequency j, but they show more clearly a negative trend in ripeness if lower frequencies are used. Nevertheless, the method appears to be able to identify successfully the thresholds as shown by the change point detection presented in Fig. 2.

In order to detect a change point common to all the observed series, they have been averaged over i once fixed the value of j. Hence, the change point has been computed as the minimum value of t such that $\frac{\sum_{i=1}^{n} Y_{ijt}}{n} > \hat{\gamma}_j$, for $j = 1, \dots, 9$. The three-dimensional plot in Fig. 2 shows that, on average, the regime switch occurs at time $t = 12, 13, 14$ by varying j. Therefore, these results suggest that there is a

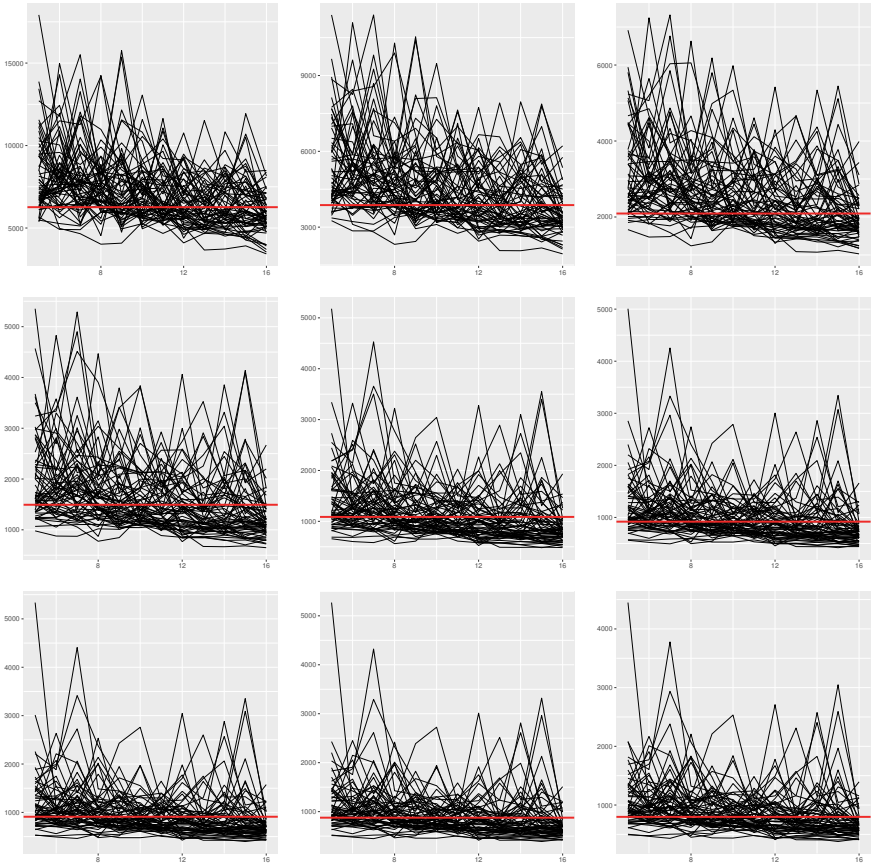


Fig. 1: Strawberries bioimpedance (y-axis) observed over time $t = 5,\ldots,16$ (x-axis) and estimated thresholds $\hat{\gamma}_j$ (red lines). Each plot refers to a different electrical frequency: from left, $j = 8690.783, 23635.120, 74152.869$ (top panel); $j = 201663.287, 548435.709, 1043389.435$ (middle panel); $j = 1720663.222, 2290015.458, 3047761.311$ (bottom panel). Note that the y-axis range varies among the plots to better show the time series behaviour.

physio-chemical change in the observed strawberries during days 6-7 of the on-plant maturation process irrespective of the electrical impedance spectroscopy frequency.

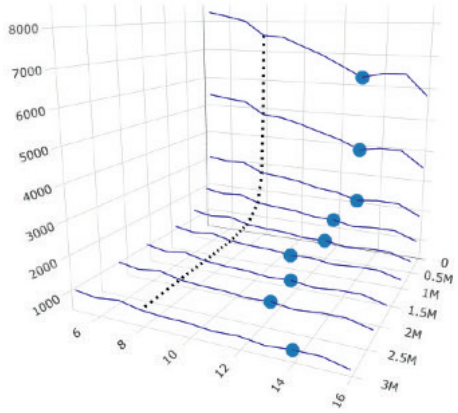


Fig. 2: Average strawberries bioimpedance (y-axis) per frequency j (z-axis) (blue lines) observed over time $t = 5, \ldots, 16$ (x-axis) and detected change points (blue dots). The black dotted line separates the observations used as instrumental variables from those used in the threshold estimation.

5 Conclusion

In this paper we presented an application of the three-way dynamic panel threshold model proposed by [2] in the special case where the dependent variable is self-excited and there are no other regressors. We showed the ability of the model to detect changes in fruit ripeness that is useful to fruit industry for quality control.

Acknowledgements The authors acknowledge the support of the Free University of Bozen-Bolzano via the interdisciplinary research project "New directions in statistical methods for bio-impedance analysis of fruit ripeness (BIOFRUIT)".

References

1. Aminikhanghahi, S. and Cook, D.J. (2017) A survey of methods for time series change point detection, *Knowledge and Information Systems*, 51(2), 339-367.
2. Di Lascio, F.M.L. and Perazzini, S. (2022) Three-way panel threshold model for change point detection. To be submitted.
3. Hansen, L. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, 50(3), 1029-1054.
4. Ibba, P. (2021), Fruit quality evaluation using electrical impedance spectroscopy. PhD thesis, 196, Supervisor: P. Lugli, Free University of Bozen-Bolzano, https://hdl.handle.net/10863/18446.
5. Seo, M. and Shin, Y. (2016) Dynamic panels with threshold effect and endogeneity. *Journal of Econometrics*, 195(2), 169-186.

Bayesian modelling and inference 3

A dynamic power prior approach to non-inferiority trials for normal means with unknown variance

Power prior dinamiche in prove cliniche di non-inferiorità per medie normali con varianza incognita

Francesco Mariani, Fulvio De Santis, Stefania Gubbiotti

Abstract Non-inferiority (NI) trials compare new experimental therapies to standard ones (active control). Since historical information on the control treatment is often available, a Bayesian approach to NI trials allows to exploit results from past studies and, eventually, to improve accuracy of inference. Here, we propose the use of a dynamic power prior: the active control treatment's endpoint is modelled by a power prior distribution, whose informativeness is tuned by a measure of similarity between past and current information. The methodology is evaluated and compared to the frequentist method by simulation; an application to real drug data is available as well.

Abstract *I test clinici di non-inferiorità (NI) confrontano una terapia sperimentale e una standard (controllo attivo). Un approccio di tipo bayesiano ai test NI permette di sfruttare negli studi correnti i risultati degli studi passati e, eventualmente, di migliorare l'accuratezza dell'inferenza. Qui viene proposta una metodologia bayesiana in cui il parametro relativo all'effetto del controllo attivo è modellato con una power priori informativa, la cui informatività è modulata secondo una misura di similarità tra informazione storica e corrente. Mediante studi di simulazione, la metodologia è valutata con criteri frequentisti; viene infine presentata un'applicazione a dati reali.*

Key words: Bayesian clinical trials, Hellinger distance, historical control.

Francesco Mariani
Sapienza University of Rome, Department of Statistics, Piazzale Aldo Moro n.5, 00185 Rome, Italy, e-mail: f.mariani@uniroma1.it

Fulvio De Santis
Sapienza University of Rome, Department of Statistics, Piazzale Aldo Moro n.5, 00185 Rome, Italy e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti
Sapienza University of Rome, Department of Statistics, Piazzale Aldo Moro n.5, 00185 Rome, Italy e-mail: stefania.gubbiotti@uniroma1.it

1 Introduction

Non-inferiority (NI) clinical trials aim to establish if a new experimental therapy is not worse than a standard treatment (active control), which is known to be effective. Since much controversy exists regarding the ethic of placebo-controlled trials [11, 13], NI tests are quite popular because they do not require a placebo arm. As past information on the control arm is often available, the Bayesian paradigm offers an ideal framework for the inclusion of historical information in a current experiment [3, 4, 6, 10, 12]. On the one hand, by borrowing historical data more trial resources can be devoted to the novel treatment while retaining accurate estimates of the current control arm parameters [12]; on the other hand, the potential lack of full homogeneity between current and previous data may recommend their down-weighting. At the same time, regulatory agencies [1, 2] recommend that novel statistical methodologies for analysing trial results parallel to the frequentist strategy [6, 7].

Formally, let us consider a NI trial in which an experimental therapy E is compared to an active control C. The latter is known to be effective since in past studies the control (C_0) has been tested with a placebo P_0. We are interested in studying drug average effects by assuming that the treatment responses in historical and current trials are distributed as mutually independent normal random variables, i.e. $X_i \sim N(\mu_i, \sigma_i^2)$ where μ_i is the unknown parameter of interest, σ_i^2 is the variance, for $i \in \{C_0, P_0, C, E\}$. NI of the experimental therapy with respect to the active control is declared by rejecting the null hypothesis of the test $H_0 : \mu_E - \mu_C \leq -\delta$ vs $H_1 : \mu_E - \mu_C > -\delta$ where $\delta > 0$ is the so-called NI margin. Its choice is not trivial, since it may depend on past performances of the active control and it may be modified according to medical judgment. The test is implemented as follows. First, we determine the p^* credible intervals for $\mu_E - \mu_C$ and for $\mu_{C_0} - \mu_{P_0}$, whose lower bounds are denoted by L_{E-C} and $L_{C_0-P_0}$, respectively. Then, we set $\delta = (1-\lambda) \times L_{C_0-P_0}$, where $\lambda \in [0,1]$ allows to involve potential clinical judgement in the NI margin determination. If $L_{E-C} > -\delta$, we reject H_0 declaring NI of the experimental treatment. We assume that no prior information on $\mu_{P_0}, \mu_{C_0}, \mu_E$ is available. As a result, the initial distributions for μ_{P_0}, μ_{C_0} and μ_E are non-informative. From standard conjugate results, the posterior distributions are also normal conditional on σ_i^2's.

Past data regarding the active control C_0 can be exploited as prior knowledge over μ_C. As a consequence, the prior distribution of μ_C is informative and, following [9], it is modelled by the *power prior*

$$\pi^{PP}(\mu_C|C_0) \propto L(\mu_C; C_0)^{a_0} \times \pi_0(\mu_C), \tag{1}$$

where $\pi_0(\mu_C)$ is a non-informative starting prior, $L(\mu_C; C_0)$ is the likelihood function of μ_C with the observed historical data C_0, $a_0 \in [0,1]$ tunes the amount of borrowed historical information. For $\pi_0(\mu_C) \propto 1$, the power posterior distribution of μ_C is based on Equation (1). Following the dynamic power prior approach of [4], we define

$$a_0(C,C_0) = \kappa(1 - d_H[\pi_c(\mu_C|C), \pi_0(\mu_C|C_0)]),$$

where d_H denotes the Hellinger distance between the posterior distributions of μ_C obtained by updating $\pi(\mu_C)$ with C and C_0, $\kappa \in [0,1]$ is a static coefficient which provides an upper limit to the quantity of information that we borrow. Since d_H is a relative measure, a_0 ranges in $[0,1]$: $a_0 = 0$ corresponds to ignoring historical information, $a_0 = 1$ implies the full incorporation of the past information in the current experiment, whereas intermediate values of a_0 represent partial borrowing. In this paper we generalize this approach by assuming unknown variances: we assume independent Jeffrey's non-informative priors for $\sigma^2_{P_0}$, $\sigma^2_{C_0}$, σ^2_C and σ^2_E; then posterior distributions are independent inverse gamma. The topic of this paper is related to the literature on the following fields: *(i)* Bayesian approaches to NI clinical trials for normal means [6]; *(ii)* power priors and their use in Bayesian design [3, 8, 9]; *(iii)* use of the Hellinger distance in Bayesian analysis of clinical trials [4, 10]. The main contribution of the present article is to apply the dynamic power prior approach to the context previously introduced by [6].

2 Simulation studies

As requested by regulatory agencies [1, 2], frequentist properties of the proposed methodology are analysed via simulation. Moreover, results are compared to the ones obtained through the frequentist fixed-margin approach [6]. Indeed, in the NI margin, $L_{C_0-P_0}$ is computed following the frequentist method even under the Bayesian paradigm; we let λ vary, thus taking into account potential clinical judgements in the determination of δ. Simulations are conducted under the unknown variance case according to the following scheme:

1. **Specify** $n_{P_0}, n_{C_0}, n_C, n_E, \mu_{P_0}, \mu_{C_0}, \mu_C, \sigma^2_{P_0}, \sigma^2_{C_0}, \sigma^2_C, \sigma^2_E$;
2. **Generate historical data** from $X_{P_0} \sim N(\mu_{P_0}, \sigma^2_{P_0})$ and $X_{C_0} \sim N(\mu_{C_0}, \sigma^2_{C_0})$.
3. **NI margin computation**: compute $\delta = (1-\lambda) \times Ł_{C_0-P_0,1-\frac{\alpha}{2}}$ following the frequentist fixed-margin methodology [6];
4. N-steps iteration:
   - **Simulate current data** from $X_C \sim N(\mu_C, \sigma^2_C)$ and $X_E \sim N(\mu_E = \mu_C - \delta + \xi, \sigma^2_E)$;
   - Compute the **Hellinger distance** using current data and historical data;
   - **Gibbs Sampling**: generate M posterior values for μ_E, σ^2_E, μ_C, σ^2_C using posterior distributions. Initial values are set equal to the true parameters values;
   - Once M empirical values for μ_E and μ_C are obtained, compute L^{sim}_{E-C} as the $\frac{\alpha}{2}$ quantile of the empirical distribution of $(\mu_E - \mu_C)$;
5. Once N values for L^{sim}_{E-C} are obtained, compute the fraction of $L^{sim}_{E-C} > -\delta$ and obtain the empirical **Type-I error** (if $\xi = 0$) or the **empirical power** (if $\xi > 0$).

The following two scenarios are considered.

- **Scenario 1**: Let $\mu_{C_0} = 1$, $\mu_{P_0} = 0$, $\mu_C = 1$ and $\mu_E = \mu_C - \delta + \xi$, $\sigma_{C_0} = \sigma_{P_0} = \sigma_C = \sigma_E = 1$, $n_{C_0} = n_{P_0} = 600$, $n_C = n_E = 30$, and $p^* = 95\%$. We simulate

7000 MCMC samples for the parameters of interest and we remove a burn-in of 2000 elements ($M = 5000$). The procedure is repeated for $N = 10000$ times so that we obtain N decisions. The estimated type-I error is the total number of rejections over N. Type-I error is computed for different values of the NI margin δ by changing $\lambda = 0, 0.3, 0.6, 0.9$; moreover, we involve different specifications of the power prior parameter a_0 (Table 1). Figure 1 (left panel) shows the power functions for $\lambda = 0.3$.

- **Scenario 2**: using the same parameters values, sample sizes and credibility/confidence level as in Scenario 1, we now add to the active control historical data a certain percentage of noise ν simulated from a uniform distribution $Unif(-5,5)$. The goal is not only to analyse the procedure under new heterogeneous scenarios, but also to assess the impact of the discount parameter a_0. For $\lambda = 0.3$, type-I error is computed for several noise levels in the historical data ($\nu = 15\%, 12.5\%, 10\%, 7.5\%, 5\%$) and for different values of a_0 (Table 2). Moreover, Figure 1 (right panel) shows the power function for $\nu = 5\%$.

Table 1 Type-I error for Bayesian and frequentist approaches under Scenario 1.

λ	$a_0 = 1$ (full borrowing)	$\kappa = 1$	$\kappa = 0.8$	$a_0 = 0$ (no-borrowing)	frequentist case
0.0	0.0310	0.0261	0.0250	0.0237	0.0253
0.30	0.0271	0.0234	0.0226	0.0219	0.0283
0.60	0.0281	0.0241	0.0241	0.0226	0.0278
0.90	0.0308	0.0260	0.026	0.0255	0.0259

Table 2 Type-I error under Scenario 2.

ν	$a_0 = 1$ (full borrowing)	$\kappa = 0.8$	$\kappa = 1$	$1 - d_H$
15%	0.0979	0.0620	0.0591	0.339
12.5%	0.1084	0.0678	0.0543	0.343
10%	0.0622	0.0454	0.0437	0.346
7.5%	0.0536	0.0399	0.0397	0.349
5%	0.0676	0.0493	0.0476	0.342

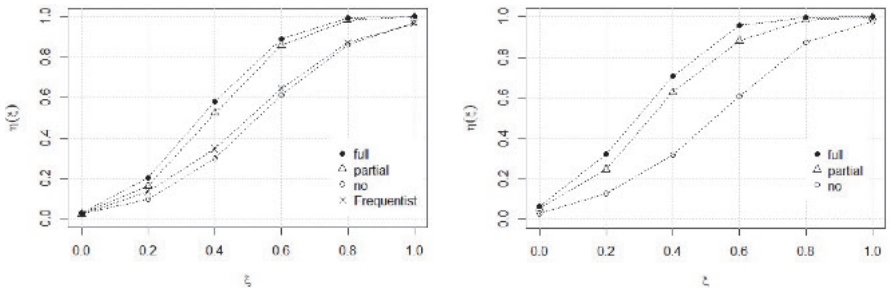


Fig. 1 Power function $\eta(\xi)$ for full, partial, no-borrowing and frequentist cases under Scenario 1 (left panel) and under Scenario 2 when $\nu = 5\%$ (right panel).

3 Application to drug data

In this section, we consider an example used in [6]: the proposed approach is applied to a dataset for drug A which is indicated for the treatment of iron deficiency anemia in adult patients with chronic kidney disease. The parameter of interest in this study is the mean change in hemoglobin (Hgb) from baseline to day 35. Drug A is dosed at 510mg per injection for a total of two daily injections. Results from previous placebo-controlled trials are available. The aim of our study is to show that the Bayesian approach with partial borrowing gives similar results to those of the frequentist approach in a NI test for a new dosage (255 mg four times daily). Results are provided in Table 3.

Table 3 Bayesian and frequentist rejection decisions with drug A (1: reject H_0; 0 otherwise).

	Probability		Decisions		
λ	$\kappa = 0.8$ (partial borrow)	$a_0 = 1$ (full borrow)	Bayesian $\kappa = 0.8$	Bayesian $a_0 = 1$	frequentist
0.4	0.982	0.990	**1**	**1**	**1**
0.5	0.969	0.983	**1**	**1**	0
0.6	0.948	0.971	0	**1**	0
0.7	0.917	0.952	0	**1**	0
0.8	0.873	0.924	0	0	0

4 Discussion

We presented a Bayesian methodology which mirrors the frequentist fixed-margin approach in construction but uses an informative power prior [9] for the active control mean effect. This informative prior depends on a measure of divergence between the historical and the current information, which is quantified by a function of the Hellinger distance: this approach was first introduced by [10] and then developed by [4] for testing NI trials with binary endpoints. The present paper re-adapts these notions to pinpoint a dynamic Bayesian methodology for testing NI trials with normal responses in the case of unknown variances.
Regulatory agencies require new statistical methodologies for the analysis of clinical trials to be evaluated in terms of their frequentist properties, such as the type-I error and the power: in this regard, we simulated data to determine if the proposed method retains the type-I error (size of the test) and if it is as powerful as the frequentist approach. In order to have a clearer idea about the new methodology performances and about the Hellinger distance behaviour, a second numerical simulation with noised data is introduced. On the one hand, under Scenario 1, all the analysed methodologies maintain their size by protecting the type-I error; however, cases in which the historical information is fully or partially borrowed are consistently more powerful than the others. On the other hand, under Scenario 2, for a larger amount of noise in the historical data (15%, 12.5%), none of the methods retains the test size;

nevertheless, for a smaller amount of noise ($10\%, 7.5\%, 5\%$), only the Bayesian approach with partial borrowing maintains its size. Finally, an application to a real case study shows how the proposed and the frequentist methods bring to the same decisions almost all the times, in contrast with the full borrowing methodology.
In conclusion, the proposed methodology controls type-I error and parallels the frequentist strategy; it is simple in construction and its greatest advantage lies on its power to declare NI when the experimental treatment is indeed non-inferior; finally, it allows to wisely borrow past information in the current study by accounting for the discrepancy between the current and the historical data. Results are encouraging and some futher developments can be foreseen: (i) in simulation studies, one might take into account potential randomness of C_0 and study its impact on the performances in terms of type-I error and power; (ii) in noised-data simulations, it would be interesting to modify the historical data with a noise which favours the null hypothesis while the current data favours the non-inferiority hypothesis; (iii) the methodology can be improved by modelling an informative dynamic power prior distribution for the active control variance.

References

1. CDER/CBER/FDA Guidance for Industry. *Non-Inferiority Clinical Trials*. (2016) Available at: https://www.fda.gov/media/78504/download.
2. CDHR/FDA Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Guidance for industry and FDA staff (2010)
3. De Santis F. Power Priors and Their Use in Clinical Trials. *Am Stat*. 60(2), 122-129 (2006)
4. De Santis F., Gubbiotti S. A method for incorporating historical information in non-inferiority clinical trials. *Proceedings of the 50th scientific meeting of the Italian Stat. Society* (2021)
5. EMA Guideline on the choice of non-inferiority margin. EMEA/CPMP/EWP/2158/99 (2005)
6. Gamalo M. A., Wu R., Tiwari R. C. Bayesian approach to non-inferiority trials for normal means. *Stat Methods Med Res*. 25(1):221-40 (2016)
7. Gamalo-Siebers M., Gao A., Lakshminarayanan M., Liu G., Natanegara F., Railkar R., Schmidli H & Song G. Bayesian methods for the design and analysis of noninferiority trials. *Journal of Biopharmaceutical Statistics*. 26:5, 823-841 (2016)
8. Hobbs B.P., Carlin B.P., Mandrekar S.J., Sergent D.J. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 67: 1047-1056 (2011)
9. Ibrahim J.G., Chen M.H., Gwon Y., Chen F. The power prior: theory and applications. *Stat Med*. 34(28):3724-49 (2015)
10. Ollier A., Morita S., Ursino M., Zohar S. An adaptive power prior for sequential clinical trials- Application to bridging studies. *Stat Methods Med Res*, 29: 2282-2294 (2020)
11. Simon R. Are placebo-controlled clinical trials ethical? *Editorial Ann Intern Med*. 133:474–5 (2000)
12. Viele K., Berry S., Neuenschwander B., et al. Use of historical control data for assessing treatment effects in clinical trials.*Pharm Stat*. 13:41-54 (2014)
13. World Medical Association. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*. 2013;310(20):2191–2194 (2013)

Bayesian Change-Point Detection for a Brownian Motion with a Total Miss Criterion

Rilevazione Bayesiana del Punto di Cambiamento per un Moto Browniano con un Criterio di Errore Complessivo

Bruno Buonaguidi

Abstract We study the following problem of sequential analysis: we observe a Brownian motion which has a zero drift initially; at a an unknown and random time θ, known as change-point, the Brownian motion takes a non-zero drift. Since the Brownian motion is observed in real time, we want to estimate θ optimally by means of a stopping time which minimizes a total miss criterion, namely the linear combination between the expected advance in detecting θ wrongly and expected delay of a late detection. This problem is solved in the Bayesian formulation, where θ is assumed to follow an exponential prior distribution.

Abstract *Studiamo il seguente problema di analisi sequenziale: osserviamo un moto Browniano che ha inizialmente un drift nullo; ad un ignoto e casuale istante θ, chiamato punto di cambiamento, il moto Browniano assume un drift non nullo. Poiché il moto Browniano é osservato in tempo reale, vogliamo stimare θ in modo ottimale mediante un tempo di arresto che minimizzi un criterio di errore complessivo, dato dalla combinazione lineare tra l'anticipo e il ritardo attesi nel rilevare θ. Il prblema é risolto nella formulazione Bayesiana, dove θ segue una distribuzione a priori di tipo esponenziale.*

Key words: Brownian motion, change-point/disorder problem, optimal stopping, sequential analysis, total miss criterion.

1 Introduction

In a change-point detection problem for a sequentially observed stochastic process the goal is to estimate the moment θ, known as *change-point* or *disorder time*, at which the process suffers an abrupt modification of its statistical properties. This

Bruno Buonaguidi
Università Cattolica del Sacro Cuore di Milano e-mail: bruno.buonaguidi@unicatt.it

problem arises in several fields: in quality control, where goods produced by a machine are sequentially analyzed in order to identify when the latter moves from an "in-control" to an "out-of-control" state; in internet surveillance problems, where data are analyzed in real time to identify hacker attacks that turn legitimate traffic into malicious traffic; in credit card fraud detection, where the expenditures pattern of a credit card holder is monitored to identify rapidly the moment at which it is hit by fraud (see, e.g., [2]); in finance, where the asset prices are analyzed to detect the appearance of an arbitrage in the market due to insider trading; in epidemiology, where one wants to identify promptly the onset of the spread of a certain disease; in biomedicine, where signals recorded by means of electroencephalograms and electrocardiograms could reveal the appearance of altered health conditions. We refer to [8] for a broader overview.

In this short paper we consider the following problem. Let X be a Brownian motion having a zero drift initially; at the unknown change-point θ, X takes a non-zero and known drift μ. The change-point detection problem for X aims at estimating θ efficiently, by relying on the observation of X only. By efficient estimation we mean that the occurrence of θ is declared so that a risk function, given by the linear combination of the expected costs due to an early and late detection, is minimized. The problem is studied in the Bayesian formulation, where θ is assumed to be independent of X and to follow an exponential distribution with an atom at 0. The risk function we consider is given by

$$R(\pi,\tau) := \mathbb{E}_\pi[(\theta-\tau)^+] + c\mathbb{E}_\pi[(\tau-\theta)^+], \tag{1}$$

where $\pi \in [0,1]$ is the prior probability that θ has already occurred at time 0 and τ is a stopping time of X, which we use to estimate the occurrence of θ and at which we interrupt the observation of X. In the expression above, $\mathbb{E}_\pi[(\theta-\tau)^+]$ is the expected advance in detecting θ wrongly, while $\mathbb{E}_\pi[(\tau-\theta)^+]$ is the expected detection delay since the occurrence of θ; the constant $c>0$ is given and weights the importance attributed to the two sources of cost. We refer to (1) as the expected *total miss* criterion.

When the first addend in (1) is replaced by the probability of a false alarm, namely $\mathbb{P}_\pi(\tau<\theta)$, we obtain the classic Shiryaev change-point disruption problem for a Wiener process (see, e.g., [7, Ch. 4.4]). Let us also observe that the risk function (1) was analyzed in [5] by setting $c=1$. Here, unlike the latter article, we reduce the optimal stopping problem associated to (1) to a new and equivalent optimal stopping problem by abandoning the original measure $\mathbb{P}_\pi$ and we provide the corresponding solution for any $c>0$.

2 The Model

A standard Brownian motion $W := (W_t)_{t\geq 0}$ and an independent and non-negative random variable θ are defined on the probability space $(\Omega,\mathscr{F},\mathbb{P}_\pi)$. The probability

measure $\mathbb{P}_\pi$ is given by

$$\mathbb{P}_\pi := \pi\mathbb{P}^0 + (1-\pi)\int_0^\infty \mathbb{P}^s \lambda e^{-\lambda s} ds, \tag{2}$$

where $\mathbb{P}^s$ is the probability measure under which $\mathbb{P}^s(\theta = s) = 1$, $s \geq 0$, and $\pi \in [0,1]$ and $\lambda > 0$ are two given and fixed values. Then, it readily follows that $\mathbb{P}_\pi(\theta = 0) = \pi$ and $\mathbb{P}_\pi(\theta > t | \theta > 0) = e^{-\lambda t}$, for $t > 0$. The researcher is allowed to observe the process $X := (X_t)_{t\geq 0}$, defined by

$$X_t = \mu(t-\theta)^+ + W_t, \quad \mu \neq 0 \text{ and known.} \tag{3}$$

This means that X is a standard Brownian motion taking a non-zero drift at θ. Then, solving the Bayesian change-point detection problem for X with a total miss criterion means computing the value function

$$V(\pi) := \inf_\tau R(\pi, \tau), \quad \pi \in [0,1], \tag{4}$$

and determining the optimal estimate $\tau_\star$ of θ, at which the infimum in (4) (searched among all the stopping times τ of X) is reached.

To approach the problem (4) we need the following quantities:

$$L_t := \frac{d\mathbb{P}^0}{d\mathbb{P}^\infty}\bigg|_{\mathscr{F}_t^X}, \quad \Pi_t := \mathbb{P}_\pi(\theta \leq t | \mathscr{F}_t^X), \quad \Phi_t := \frac{\Pi_t}{1-\Pi_t}, \tag{5}$$

where $\mathscr{F}_t^X$ is the sigma-algebra generated by X up to $t \geq 0$ and $\mathbb{P}^\infty$ is the probability measure under which the drift of X never changes (i.e., $\mathbb{P}^\infty(\theta = \infty) = 1$). The processes $L := (L_t)_{t\geq 0}$, $\Pi := (\Pi_t)_{t\geq 0}$ and $\Phi := (\Phi_t)_{t\geq 0}$ are known as likelihood ratio process, posterior probability process and posterior probability odds process, respectively, for obvious reasons. Girsanov's theorem and well known arguments (see, e.g., [4]) show that the explicit expressions of L_t and Φ_t for $t \geq 0$ are

$$L_t = \exp\left(\mu X_t - \frac{\mu^2}{2}t\right), \quad \Phi_t = e^{\lambda t} L_t \left(\frac{\pi}{1-\pi} + \lambda \int_0^t \frac{e^{-\lambda s}}{L_s} ds\right). \tag{6}$$

3 Change of Measure and optimal estimation of θ

Under the probability measure $\mathbb{P}^\infty$, the process X from (3) boils down to the standard Brownian motion W and this leads to a substantial simplification of the expressions in (6), which become independent of θ. Moreover, under the measure $\mathbb{P}^\infty$, the problem (4) can be equivalently expressed as

$$V(\pi) = (1-\pi)\left(\frac{1}{\lambda} + c\hat{V}\left(\frac{\pi}{1-\pi}\right)\right), \quad \pi \in [0,1), \tag{7}$$

being

$$\hat{V}(\varphi) := \inf_{\tau} \mathbb{E}_{\varphi}^{\infty} \left[\int_0^{\tau} e^{-\lambda t} \left(\Phi_t - \frac{1}{c} \right) dt \right], \tag{8}$$

where the infimum is taken over all the stopping times τ of X and the probability measure $\mathbb{P}_{\varphi}^{\infty}$ is such that $\mathbb{P}_{\varphi}^{\infty}(\Phi_0 = \varphi) = 1$, for $\varphi \in [0,\infty)$ (see [1] for a formal proof).

From the structure of the integrand in (8), we can conjecture that there exists an optimal point $b \geq 1/c$ such that, as soon as Φ_t crosses b from below, we can interrupt the observation to declare the occurrence of θ. Standard arguments (see, e.g., [6, Ch. 3-4] or [7, Ch. 3]) show that the function $\hat{V}$ and the optimal boundary b solve the free-boundary problem

$$(\mathbb{L}\hat{V} - \lambda \hat{V})(\varphi) = -\left(\varphi - \frac{1}{c} \right), \quad \varphi \in [0,b), \tag{9}$$

$$\hat{V}(\varphi) = 0, \quad \varphi \in [b,\infty) \quad \textit{(instantaneous stopping)}, \tag{10}$$

$$\hat{V}'(b) = 0 \quad \textit{(smooth fit)}, \tag{11}$$

where $\mathbb{L}$ is the infinitesimal generator of Φ and is given by

$$\mathbb{L}_{\Phi} = \lambda(1+\varphi)\frac{\partial}{\partial \varphi} + \frac{1}{2}\mu^2 \varphi^2 \frac{\partial^2}{\partial \varphi^2}. \tag{12}$$

Solution methods for ordinary differential equations related to quickest detection problems (see, e.g., [3, Sec. 4-5]) imply that the solution to (9)+(12) and (10) is given by

$$\hat{V}(\varphi) = A(1+\varphi) - \frac{1}{c\lambda} - \frac{2}{\mu^2}(1+\varphi) \int_{\frac{b}{1+b}}^{\frac{\varphi}{1+\varphi}} \left(\frac{1-r}{r} \right)^{\nu} e^{\nu/r} \int_0^r \frac{y^{\nu-1}}{(1-y)^{\nu+2}} e^{-\nu/y} dy dr, \tag{13}$$

for $\varphi \in [0,b]$ and $\hat{V}(\varphi) = 0$ for $\varphi \in [b,\infty)$. In (13) above, A is a constant to be determined and $\nu := 2\lambda/\mu^2$. The instantaneous and smooth fit conditions (10)-(11) at b yield $A = 1/(c\lambda(1+b))$ with b which solves

$$\frac{e^{\nu(1+b)/b}}{b^{\nu}} \int_0^{\frac{b}{1+b}} \frac{y^{\nu-1}}{(1-y)^{\nu+2}} e^{-\nu/y} dy = \frac{1}{c\nu}. \tag{14}$$

Figure 1 below shows the function $\varphi \mapsto \hat{V}(\varphi)$ on $[0,\infty)$ and the optimal stopping point b. Applying finally the Itó-Tanaka formula to $\hat{V}$ obtained above, which boils down to Itô's formula due to the smooth fit (11), and making use of the optional sampling theorem, it is readily verified that the solution to the free-boundary problem (9)-(11) is unique and coincides with that of the optimal stopping problem (8) and that the optimal estimate $\tau_{\star}$ of θ is given by

$$\tau_{\star} = \inf\{t \geq 0 : \Phi_t \geq b\}. \tag{15}$$

From (6) we see that Φ depends explicitly on the observed path of X and is therefore observable. Then, $\tau_\star$ can straightforwardly be computed.

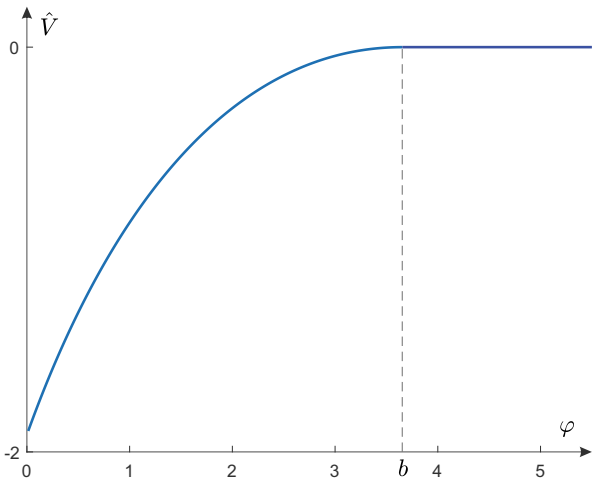


Fig. 1: A computer drawing of the function $\varphi \mapsto \hat{V}(\varphi)$ on $[0,\infty)$ from (8) and (13), and of the optimal stopping boundary point $b = 3.654$ solving (14). We fixed $\mu = 0.5$, $\lambda = 1$ and $c = 0.3$.

References

[1] Buonaguidi, B.: The disorder problem for diffusion processes with the ε-linear and expected total miss criteria. In progress (2022)

[2] Buonaguidi, B., Mira, A., Bucheli, H., and Vitanis, V.: Bayesian quickest detection of credit card fraud. Bayesian Analysis **17**, 261–290 (2022)

[3] Ernst, P. A. and Peskir, G.: Quickest real-time detection of a Brownian coordinate drift. arXiv:2007.14786v1 (2020)

[4] Gapeev, P. V. and Shiryaev, A. N.: Bayesian quickest detection problems for some diffusion processes. Advances in Applied Probability **45**, 164–185 (2013)

[5] Karatzas, I.: A note on Bayesian detection of change-points with an expected miss criterion. Statistics & Decisions **21**, 3–13 (2003)

[6] Peskir, G. and Shiryaev, A. N.: *Optimal Stopping and Free-Boundary Problems*, Lectures in Mathematics ETH Zürich. Basel, Birkhäuser Verlag (2006)

[7] Shiryaev, A. N.: *Optimal Stopping Rules*. New York, Springer-Verlag (1978)

[8] Tartakovsky, A., Nikiforov, I., and Basseville, M.: *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press (2015)

On the comparison of alternative Bayesian measures of posterior discrepancy

Fulvio De Santis and Stefania Gubbiotti

Abstract In this paper we consider the problem of quantifying the degree of conflict between posterior distributions of a parameter of interest based on the same data but induced by two alternative prior distributions. This problem has recently received attention in the literature on clinical trials. Specifically, we contrast two interval-based measures of discrepancy using as a reference the Hellinger distance between the posteriors. We perform the comparison in the context of clinical experiments involving binary outcomes and using the normal approximation for the log odds ratio. We provide a real data example.

Abstract *In questo articolo si considera il problema di quantificare il grado di conflitto tra distribuzioni a posteriori per un parametro di interesse basate sugli stessi dati ma indotte da due diverse distribuzioni a priori. Questo problema è stato recentemente oggetto di attenzione nella letteratura nell'ambito delle prove cliniche. In particolare si confrontano due misure di discrepanza basate su intervalli di credibilità prendendo come riferimento la distanza di Hellinger tra le due distribuzioni a posteriori. L'esame comparativo viene effettuato nel contesto di esperimenti clinici con dati binari in cui si utilizza l'approssimazione normale per il log odds ratio. Le caratteristiche delle misure oggetto del confronto sono illustrate mediante un esempio basato su dati reali.*

Key words: Consensus, Credible intervals, Hellinger distance, Normal approximation, Posterior distribution.

Fulvio De Santis
Sapienza University of Rome, Department of Statistics, Piazzale Aldo Moro n.5, 00185 Rome, Italy
e-mail: fulvio.desantis@uniroma1.it

Stefania Gubbiotti
Sapienza University of Rome, Department of Statistics, Piazzale Aldo Moro n.5, 00185 Rome, Italy
e-mail: stefania.gubbiotti@uniroma1.it

1 Introduction

In Bayesian inference extra-experimental information and uncertainty on model parameters are expressed by prior distributions. Sometimes alternative sources of knowledge exist. For example, in a clinical trial setup where the parameter of interest represents a treatment effect or effects difference, two experts may have conflicting opinions (such as skepticism or optimism). Similarly, when two historical data sets are available, the priors based on these sources may produce discordant evidence. Both situations potentially yield contrasting posterior conclusions. In order to quantify the degree of discordance between posterior distributions induced by different priors, several measures of discrepancy have been recently introduced. In this paper we focus on three approaches. The first two methods, respectively proposed by [5] and [3], are based on credible intervals, whereas the third one is a formal distance between posterior distributions (see [4]). All three methods have been originally introduced to define predictive sample size determination criteria. Conversely, the goal of this paper is to compare the posterior behaviour of the interval-based measures with the formal distance between posteriors, regardless of their use in experimental design. Specifically, the comparison is performed in the context of clinical trials involving binary outcomes and using the normal approximation for the log odds ratio (log OR).

The paper is organized as follows. In Section 2 we introduce the three measures of discrepancy between posterior distributions. In Section 3 we specialize the methods for normal models. Section 4 provides an application to a comparative study of two treatments for diabetic patients with coronary artery disease is considered. Finally, Section 5 reports a discussion.

2 Methodology

Let θ be a real one-dimensional unknown parameter and let $\mathbf{X}_n$ be the data based on n observations with density or probability mass function $f_n(\cdot|\theta)$. According to the Bayesian approach, let us assume that θ is a continuous random variable, $\pi(\theta)$ its prior density, $\mathbf{x}_n$ the observed data and $\pi(\theta|\mathbf{x}_n)$ the corresponding posterior density. We also suppose that $\pi(\theta|\mathbf{x}_n)$ is unimodal so that both highest posterior density (HPD) and equal tails (ET) credible sets are intervals. Assume now that two alternative sources of prior information are available and formalized by two prior densities $\pi_a(\theta)$ and $\pi_b(\theta)$ that induce the two posteriors $\pi_a(\theta|\mathbf{x}_n)$ and $\pi_b(\theta|\mathbf{x}_n)$. The experiment shows consensus if a measure of discrepancy between $\pi_a(\theta|\mathbf{x}_n)$ and $\pi_b(\theta|\mathbf{x}_n)$ is small to a prespecified degree. Let us consider the $(1-\gamma)$-level posterior credible interval $C_j = [L_j, U_j]$, i.e. a subset of the parameter space Θ such that $\int_{C_j} \pi_j(\theta|\mathbf{x}_n)d\theta = 1-\gamma$, for $j = a, b$.

In this article we focus on the following three measures of discrepancy:

$$D_1 = \max\{|L_a - L_b|, |U_a - U_b|\}, \tag{1}$$

$$D_2 = \max\{1 - \mathbb{P}_a(C_b), 1 - \mathbb{P}_b(C_a)\}, \tag{2}$$

$$D_3 = \left(1 - \int_\Theta \sqrt{\pi_a(\theta|\mathbf{x}_n) \cdot \pi_b(\theta|\mathbf{x}_n)}d\theta\right)^{\frac{1}{2}}. \tag{3}$$

D_1 has been proposed in [5], D_2 in [3] and D_3 is the Hellinger distance between the posterior densities (see [6] for its first application in the clinical trial context). Therefore, we have *posterior consensus* at degree $\epsilon_i \in [0, 1]$, for $i = 1, 2, 3$, if $D_i < \epsilon_i$. Notice that D_1 is not a relative measure and depends on the measurement unit of θ (see [2]). On the contrary both D_2 and D_3 range in $[0, 1]$, a fact that makes the choice and the interpretation of ϵ_i easier. Furthermore, while D_1 and D_2 are based on specific posterior summaries, D_3 is an overall distance between distributions.

3 Results for normal model

Assume now that data $X_1, \ldots, X_n$ are summarized by a statistic Y_n that, conditional on θ, is approximately normal with parameters $(\theta, \sigma^2/n)$, σ^2 known. Suppose that $\pi_j(\theta)$ is a normal density with mean μ_j and variance σ^2/n_j, where n_j is the prior sample size, $j = a, b$. From standard conjugate analysis the posterior density of θ is normal with parameters $\bar{\mu}_j = (ny_n + n_j\mu_j)/(n + n_j)$, $\bar{\sigma}_j^2 = \sigma^2/(n + n_j)$, $j = a, b$. Under these assumptions, $C_j = [L_j, U_j] = \bar{\mu}_j \pm z\bar{\sigma}_j$, where z is the $1 - \gamma/2$-level quantile of the standard normal distribution. It is easy to check that

$$D_1 = \max\{|(\bar{\mu}_a - \bar{\mu}_b) - z(\bar{\sigma}_a - \bar{\sigma}_b)|, |(\bar{\mu}_a - \bar{\mu}_b) + z(\bar{\sigma}_a - \bar{\sigma}_b)|\}, \tag{4}$$

$$D_2 = \max\left\{1 - \left[\Phi\left(\frac{\bar{\mu}_b + z\bar{\sigma}_b - \bar{\mu}_a}{\bar{\sigma}_a}\right) - \Phi\left(\frac{\bar{\mu}_b - z\bar{\sigma}_b - \bar{\mu}_a}{\bar{\sigma}_a}\right)\right], 1 - \left[\Phi\left(\frac{\bar{\mu}_a + z\bar{\sigma}_a - \bar{\mu}_b}{\bar{\sigma}_b}\right) - \Phi\left(\frac{\bar{\mu}_a - z\bar{\sigma}_a - \bar{\mu}_b}{\bar{\sigma}_b}\right)\right]\right\}, \tag{5}$$

$$D_3 = \left(1 - \sqrt{\frac{2\bar{\sigma}_a\bar{\sigma}_b}{\bar{\sigma}_a^2 + \bar{\sigma}_b^2}} \cdot \exp\left\{-\frac{1}{4}\frac{(\bar{\mu}_a - \bar{\mu}_b)^2}{\bar{\sigma}_a^2 + \bar{\sigma}_b^2}\right\}\right)^{\frac{1}{2}}. \tag{6}$$

D_3 has been proposed and discussed as a measure of posterior discrepancy in [4].

4 Example

The model described in the previous section can be used for instance in phase III two-arms clinical trials with binary data when θ is the log OR and Y_n is the sample

log OR. Specifically, following [7], Y_n is asymptotically $N(\theta, \sigma^2/n)$ with $\sigma = 2$, so that n can be interpreted as the effective number of observations, i.e. the total number of events occuring in the two arms. In this setup we consider an example based on the real trial data discussed in [1] where coronary artery bypass graft (CABG) is compared with percutaneous coronary intervention (PCI) in diabetic patients with multivessel coronary artery disease in terms of survival (with log OR > 0 favouring CABG with respect to PCI). In [1] the prior distribution for the log OR is based on a metanalysis of 8 historical trials and is combined with the likelihood of the data of the FREEDOM trial. The posterior distribution of the log OR shows evidence in favour of CABG (e.g. posterior mean 0.545 with 95% credible interval [0.342, 0.734] corresponding to a reduction in mortality risk between 29%-52%).

(i)

n	L_a	U_a	L_b	U_b	D_1	D_2	D_3
10	0.349	0.748	0.014	0.406	0.342	0.919	0.755
50	0.372	0.752	0.067	0.441	0.311	0.895	0.723
100	0.396	0.755	0.121	0.475	0.280	0.862	0.686
500	0.495	0.760	0.343	0.606	0.154	0.627	0.477
1000	0.545	0.756	0.447	0.657	0.099	0.450	0.342

(ii)

n	L_a	U_a	L_b	U_b	D_1	D_2	D_3
10	0.007	1.144	-0.262	0.858	0.285	0.166	0.109
50	0.209	1.047	0.059	0.890	0.157	0.114	0.063
100	0.316	0.985	0.219	0.884	0.100	0.091	0.041
500	0.511	0.849	0.486	0.823	0.026	0.060	0.011
1000	0.563	0.806	0.550	0.793	0.013	0.055	0.006

(iii)

n	L_a	U_a	L_b	U_b	D_1	D_2	D_3
10	0.349	0.748	-0.550	1.930	1.181	0.758	0.447
50	0.372	0.752	0.136	1.244	0.492	0.544	0.252
100	0.396	0.755	0.298	1.082	0.327	0.443	0.187
500	0.495	0.760	0.515	0.865	0.105	0.231	0.092
1000	0.545	0.756	0.566	0.814	0.058	0.159	0.061

(iv)

n	L_a	U_a	L_b	U_b	D_1	D_2	D_3
10	0.007	1.144	-0.550	1.930	0.786	0.377	0.136
50	0.209	1.047	0.136	1.244	0.198	0.148	0.027
100	0.316	0.985	0.298	1.082	0.097	0.101	0.012
500	0.511	0.849	0.515	0.865	0.016	0.060	0.002
1000	0.563	0.806	0.566	0.814	0.008	0.055	0.001

Table 1 Posterior credible intervals ($1 - \gamma = 0.95$) and discrepancy measures D_1, D_2, D_3 for increasing values of n (fixed $y_n = 0.69$ and known variance $\sigma^2 = 4$), under the following prior assumptions: (i) π_a historical ($\mu_a = 0.545$, $n_a = 376$) and π_b skeptical ($\mu_b = 0.198$, $n_b = 390$); (ii) π_a discounted historical ($\mu_a = 0.545$, $n_a = 37.6$) and π_b discounted skeptical ($\mu_b = 0.198$, $n_b = 39$); (iii) π_a historical ($\mu_a = 0.545$, $n_a = 376$) and π_b non-informative ($n_b = 0$); (iv) π_a discounted historical ($\mu_a = 0.545$, $n_a = 37.6$) and π_b non-informative ($n_b = 0$).

Let us now discuss the choice of the priors π_a and π_b. As in [4], we first consider the posterior distribution as π_a, e.g. a normal density of parameters $\mu_a = 0.545$, $n_a = 376$. For π_b we follow [1] who also consider a skeptical prior yielding a second posterior distribution for the log OR (e.g. posterior mean 0.198 with 95% credible interval $[0, 0.4]$, corresponding to a reduction in mortality risk between 0%-23%). This results in a normal density of parameters $\mu_b = 0.198, n_b = 390$. To have a better insight on the effect of n_a and n_b on the three discrepancy measures, we also consider priors with the same means μ_a and μ_b but with prior sample sizes discounted, for instance, by a factor 0.1 (thus obtaining $n_a = 37.6$ and $n_b = 39$ respectively). For comparison, we also consider a non-informative prior as π_b ($n_b = 0$) in contrast with the same π_a and its discounted version. In summary, by combining the above mentioned choices of π_a and π_b, we have the following four cases:

(i) π_a historical ($\mu_a = 0.545$, $n_a = 376$) and π_b skeptical ($\mu_b = 0.198$, $n_b = 390$);
(ii) π_a discounted historical ($\mu_a = 0.545$, $n_a = 37.6$) and π_b discounted skeptical ($\mu_b = 0.198$, $n_b = 39$);
(iii) π_a historical ($\mu_a = 0.545$, $n_a = 376$) and π_b non-informative ($n_b = 0$);
(iv) π_a discounted historical ($\mu_a = 0.545$, $n_a = 37.6$) and π_b non-informative ($n_b = 0$).

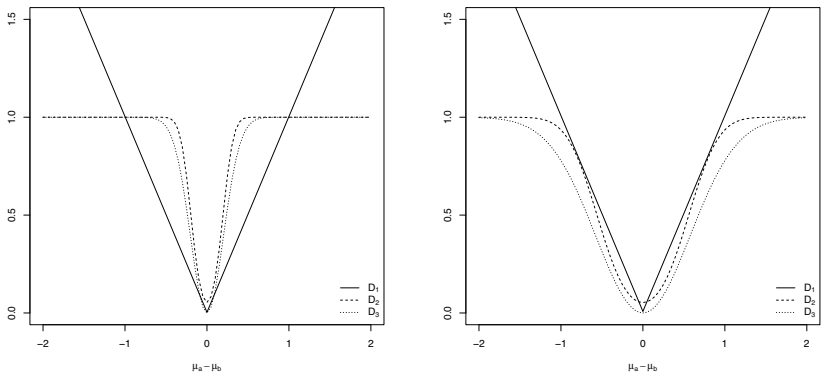


Fig. 1 Plots of D_1, D_2 and D_3 as functions of $\mu_a - \mu_b$ for $n = 10$, $y_n = 0.69$, with $n_a = 376$ $n_b = 390$ (left panel) and $n_a = 37.6$ $n_b = 39.0$ (right panel).

Table 4 reports bounds of 95% posterior credible intervals and values of D_1, D_2 and D_3 for several sample sizes in cases (i)-(iv). Figure 1 shows the plots of D_1, D_2 and D_3 as functions of $\mu_a - \mu_b$ for different choices of the prior sample sizes.
Some comments are in order.

1. As expected, D_i decreases as $n \to \infty$, $i = 1, 2, 3$.
2. As noted in Section 2, $D_i \in [0, 1]$ for $i = 2, 3$, whereas D_1 is not a relative measure (see for instance case (iii) for $n = 10$ where $D_1 > 1$; see also Figure 1).

3. In all the cases we have considered $D_2 > D_3$. As regards D_1, instead, depending on the choice of the prior parameters for each sample size, its values are smaller (see for instance case (i)) or larger (see case (ii) for $n \leq 100$) than those of D_2 and D_3.
4. The more informative the priors (i.e. the larger the values of n_a and n_b), the larger the values of D_i, $i = 1, 2, 3$ (compare (i) vs (ii) and (iii) vs (iv)).
5. In our example the skeptical prior is more in contrast with π_a than the non-informative prior (compare (iii) vs (i) and (iv) vs (ii)). This is shown by the smaller values of D_3, which is the only formal distance between posterior distributions. The values of D_2 follow the trend of D_3, whereas D_1 is not consistent (for instance in case (iii) D_1 is larger than in case (i) for $n \leq 100$).

5 Conclusions

The topic of this article is the comparison of two interval-based measures of conflict between the posterior distributions induced by contrasting priors D_1 and D_2 with the behaviour of the Hellinger distance between posterior distributions D_3. Under the normal model assumptions we have performed the comparison in the context of real clinical trials using several setups of conflicting priors. Results suggest that D_2 has a behavior consistent with D_3, whereas D_1 is more difficult to interpret not only because by definition it is not a relative measure but also because its values do not follow the same trend of D_3 as shown in comment 5 of the previous section. We plan to elaborate on the analytical aspects of the problem in future research.

References

1. Bittl, J.A., He, Y.: Bayesian analysis. A practical approach to interpret clinical trials and create clinical practice guidelines. Circ Cardiovasc Qual Outcomes. **10**(8) (2017)
2. De Santis, F., Gubbiotti, S.: Joint control of consensus and evidence in Bayesian design of clinical trials. *Biometrical Journal* (online) 1-15, doi: 10.1002/bimj.202100035 (2021)
3. De Santis, F., Gubbiotti, S.: A method for incorporating historical information in non-inferiority trials. 50th Meeting of the Italian Statistical Society (Pisa) Book of short papers SIS 2021 (9788891927361)
4. De Santis, F., Gubbiotti, S.: Optimal sample size for evidence and consensus in phase III clinical trials. Optimization in Artificial Intelligence and Data Sciences ODS, First Hybrid Conference, Rome, Italy, September 14-17, 2021 Editors: L. Amorosi, P. Dell'Olmo, I. Lari. Springer 2022 (in press)
5. Joseph, L., Belisle, P.: Bayesian consensus-based sample size criteria for binomial proportions. Statist. Med. **38**(23), 4566 –4573 (2019)
6. Ollier, A., Morita, S., Ursino, M., Zohar, S.: An adaptive power prior for sequential clinical trials - Application to bridging studies. Statistical Methods in Medical Research. **29**(8), 2282–2294 (2020)
7. Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian approaches to clinical trials and health-care evaluation. Wiley, New York (2004)

A Bayesian Test for the comparison of two independent populations

Un test bayesiano per il confronto di due popolazioni indipendenti

Mara Manca, Silvia Columbu and Monica Musio

Abstract In this paper, we propose a testing procedure that allows to compare parameter functions from two independent populations. We address this issue through a test based on the Bayesian Discrepancy Measure, a measure of evidence recently introduced in the literature. This approach is flexible, as it can be adapted to take into account different distributions and different parameter transformations. In addition, this methodology enables us to tackle problems that are not yet covered in the literature.

Abstract *In questo articolo, proponiamo una procedura di test che permette di confrontare funzioni dei parametri di due popolazioni indipendenti. Affrontiamo questo problema attraverso un test basato sulla Misura di Discrepanza Bayesiana, una misura di evidenza recentemente introdotta in letteratura. Tale approccio è flessibile poiché può essere riadattato in modo da prendere in considerazione diverse distribuzioni e diverse funzioni dei parametri. Inoltre, questa metodologia ci permette di affrontare problemi che non sono ancora stati trattati in letteratura.*

Key words: Bayesian test, Bayesian Discrepancy Measure, sharp hypothesis, variance, coefficient of variation, skewness

1 Introduction

In this article we propose a general procedure for testing hypotheses on the comparison of two parameters, or their transformations, from two independent populations. We consider a valutative approach which involves only the specification of one hypothesis without considering the alternative (in the same spirit of Fisher's pure significance test, see [3]). This method is based on a measure of evidence, the Bayesian Discrepancy Measure (BDM), which has been recently introduced in the literature

Dept. of Mathematics and Computer Sciences, University of Cagliari, Via Ospedale, 72, Cagliari

(see [2]). The testing procedure we are going to deal with has a general validity as it allows to perform a variety of tests in a simple way. This problem has been only partially addressed so far, in fact, tests found in the literature usually are defined for specific parameter functions. For example, in the frequentist field, we need "ad hoc" techniques to solve similar problems, which are usually applicable only to specific cases.

In what follows, we propose to use the BDM for comparing variances, coefficients of variation and skewness of two particular independent populations.

The problem of comparing variances has been widely discussed in both the frequentist and the Bayesian fields. Whereas the problem of comparing coefficients of variation, although widely addressed in the frequentist paradigm, has had only few contributions in the Bayesian one (see [1] for all the references). Additionally, to the best of our knowledge, the case of the comparison of two skewness indexes treated here is completely original and has not been addressed before.

2 The Bayesian Discrepancy Measure

Let $X \sim f(x|\theta)$ be a parametric model indexed by a scalar parameter $\theta \in \Theta = \mathbb{R}$, $g(\theta)$ a prior density, $\mathbf{x}$ a sample of *iid* observations and $g(\theta|\mathbf{x}) \propto g(\theta)\, L(\theta|\mathbf{x})$ a posterior density, where $L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ is the likelihood function. Let also $G(\theta|\mathbf{x})$ be the correspondent distribution function and $m_1 = G^{-1}\left(\frac{1}{2}|\mathbf{x}\right)$ the posterior median.
We are interested in testing sharp hypothesis $H : \theta = \theta_H$. The BDM is defined as

$$\delta_H = 1 - 2 \cdot \int_{I_E} dG(\theta|\mathbf{x}) = 1 - 2 \cdot \min\left\{\mathbb{P}(\theta > \theta_H|\mathbf{x}), \mathbb{P}(\theta < \theta_H|\mathbf{x})\right\}, \qquad (1)$$

where the external interval is expressed as

$$I_E = \begin{cases} (\theta_H, +\infty) \text{ if } & \theta_H > m_1 \\ (-\infty, \theta_H) \text{ if } & \theta_H < m_1. \end{cases}$$

If $\theta_H = m_1$ this interval can be expressed equivalently as $(-\infty, m_1)$ or $(m_1, +\infty)$, while, if $\Theta \subseteq \mathbb{R}$ then the interval I_E will be redefined consequently.
The more θ_H is far from the median of the posterior distribution $G(\theta|\mathbf{x})$, the more δ_H is large. In this case, H is not conform to G. On the contrary, the smaller δ_H the stronger is the evidence in favor of H.

The approach described previously can be extended to the multiparameter case. Let $X \sim f(x|\theta)$ be a parametric model indexed by a parameter $\theta = (\theta^1, \cdots, \theta^k) \in \Theta \subseteq \mathbb{R}^k$ and $\varphi = \gamma(\theta)$ be the parameter of interest, on which we would like to perform a test, where $\gamma : \Theta \to \Phi \subseteq \mathbb{R}$ is a known continuous function. Suppose, in addition, that there exists a one-to-one transformation between θ and (φ, ζ), with ζ a noise parameter.
For $k \geq 2$, hypothesis of the type $H : \varphi = \varphi_H$ induce a partition $\{\Theta_a, \Theta_H, \Theta_b\}$ of Θ,

through the function γ, where the subsets are defined as

$$\Theta_a = \{\theta \in \Theta \mid \gamma(\theta) < \varphi_H\}, \quad \Theta_b = \{\theta \in \Theta \mid \gamma(\theta) > \varphi_H\},$$
$$\Theta_H = \{\theta \in \Theta \mid \gamma(\theta) = \varphi_H\},$$

through which the BDM is expressed as

$$\delta_H = 1 - 2 \cdot \int_{I_E} \mathrm{d}G(\theta|\mathbf{x}) = 1 - 2 \cdot \min_{a,b} \left\{ \mathbb{P}(\theta \in \Theta_a \mid \mathbf{x}), \mathbb{P}(\theta \in \Theta_b \mid \mathbf{x}) \right\}, \qquad (2)$$

where now the external set is $I_E = \arg\min_{a,b} \left\{ \mathbb{P}(\theta \in \Theta_a \mid \mathbf{x}), \mathbb{P}(\theta \in \Theta_b \mid \mathbf{x}) \right\}$.
Formula (2) can be naturally extended to cases in which we can find a partition of three elements $\{\Theta_a, \Theta_H, \Theta_b\}$ of the parameter space associated to the hypothesis of interest.
The acceptance or rejection of H, that characterise this test, depends on the value of δ_H and its comparison with a threshold $\omega \in \{0.95, 0.99, 0.995, 0.999, \dots\}$, whose choice depends on the researcher.

3 Comparing parameters of two independent populations

The BDM is suitable for comparing two independent populations.
Let us consider two independent random variables $X_\ell \sim f(x_\ell|\theta_\ell)$, $\ell = 1,2$, a prior density $g_\ell(\theta_\ell)$ and a sample $\mathbf{x}_\ell = \{x_{\ell_1}, \dots, x_{\ell_{n_l}}\}$ of n_ℓ *iid* observations. The posterior density is then $g_\ell(\theta_\ell|\mathbf{x}_\ell) \propto g_\ell(\theta_\ell)\, L_\ell(\theta_\ell|\mathbf{x}_\ell)$. We indicate with θ the joint parameter vector $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 = \Theta \subseteq \mathbb{R}^k \times \mathbb{R}^k$ of the two populations parameters. We are interested in testing the hypothesis

$$H : \varphi_1 - \varphi_2 = 0, \qquad (3)$$

with $\varphi_\ell = \gamma(\theta_\ell)$ the parameters of interest ($\ell = 1,2$), which identify the partition $\{\Theta_a, \Theta_H, \Theta_b\}$ of Θ where

$$\Theta_a = \left\{ \theta \in \Theta \;\middle|\; \varphi_1 < \varphi_2 \right\}, \quad \Theta_b = \left\{ \theta \in \Theta \;\middle|\; \varphi_1 > \varphi_2 \right\},$$
$$\Theta_H = \left\{ \theta \in \Theta \;\middle|\; \varphi_1 = \varphi_2 \right\}.$$

To perform the test we can then use formula (2), where

$$\mathbb{P}(\theta \in \Theta_j \mid \mathbf{x}_1, \mathbf{x}_2) = \int_{\Theta_j} g_1(\theta_1|\mathbf{x}_1)\, g_2(\theta_2|\mathbf{x}_2)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2, \quad j = a, b. \qquad (4)$$

The evaluation of these probabilities demands the computation of multidimensional integrals that can not be solved in a closed form, hence we suggest to approximate them using the Monte Carlo Integration method.

We now consider a variety of distributions and parameters transformations and briefly outline the case-by-case tools needed to calculate the BDM. In particular, we specify the posterior distributions and the subsets of the parameter space that are needed to compute the integrals (4). In all the cases discussed a Jeffreys' prior has been adopted.

Skewness of two Inverse Gaussian populations Let us consider two indipendent Inverse Gaussian random variables $X_\ell \sim IG(x_\ell \mid \mu_\ell, \lambda_\ell)$, $\ell = 1,2$, i.e.

$$f(x_\ell \mid \mu_\ell, \lambda_\ell) = \sqrt{\frac{\lambda_\ell}{2\pi x_\ell^3}} \exp\left\{ -\frac{1}{2}\lambda_\ell \left(\frac{x_\ell - \mu_\ell}{\mu_\ell \sqrt{x_\ell}} \right)^2 \right\}, \quad X_\ell \in \mathbb{R}^+, \ (\mu_\ell, \lambda_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+.$$

Given n_ℓ observations, the posterior distribution of the parameter vector with non-informative prior $g(\mu_\ell, \lambda_\ell) \propto \frac{1}{\sqrt{\mu_\ell^3 \lambda_\ell}}$ is

$$g_\ell(\mu_\ell, \lambda_\ell \mid \mathbf{x}_\ell) \propto \sqrt{\frac{\lambda_\ell^{n_\ell - 1}}{\mu_\ell^3}} \exp\left\{ -\frac{n_\ell \lambda_\ell}{2} \left(\frac{\bar{x}_\ell}{\mu_\ell^2} - \frac{2}{\mu_\ell} + \frac{1}{a_\ell} \right) \right\}, \quad \ell = 1,2,$$

where $\bar{x}_\ell$ and a_ℓ are the arithmetic and harmonic means respectively.
We are interested in comparing the skewness of the two populations. The skewness of a real-valued random variable is defined as the third standardized moment. In this case it assumes the expression

$$\varphi_\ell = 3\sqrt{\frac{\mu_\ell}{\lambda_\ell}}.$$

The hypothesis $H : \varphi_1 - \varphi_2 = 0$ identifies in the parameter space Θ the subsets

$$\Theta_j = \left\{ (\mu_1, \lambda_1, \mu_2, \lambda_2) \in \mathbb{R}^4_+ \;\middle|\; \mu_1 \lambda_2 \lesseqgtr \mu_2 \lambda_1 \right\}, \quad j = a, b.$$

Coefficients of variation of two Negative Binomial populations Let us consider two discrete Negative Binomial populations. It is known that, given $X_\ell | \lambda_\ell \sim$ Poisson$(x_\ell \mid \lambda_\ell)$ with $\lambda_\ell \sim$ Gamma$(\lambda_\ell \mid \alpha_\ell, \beta_\ell)$, then the unconditional random variable X_ℓ follows a Negative Binomial distribution

$$X_\ell \sim NB\left(x_\ell \;\middle|\; \frac{\beta_\ell}{\beta_\ell + 1}, \alpha_\ell \right), \qquad (\alpha_\ell, \beta_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+, \ \ell = 1,2.$$

Assuming the non-informative prior $g_\ell(\alpha_\ell, \beta_\ell) \propto \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1}$, where $\psi^{(1)}(\alpha_\ell) = \sum_{j=0}^{\infty} (\alpha_\ell + j)^{-2}$ is the PolyGamma function (see [5]), the posterior distribution of the parameter vector takes the expression

$$g_\ell(\alpha_\ell, \beta_\ell \mid \mathbf{x}_\ell) \propto \frac{\prod_i (x_{i\ell} + \alpha_\ell - 1)!}{[(\alpha_\ell - 1)!]^{n_\ell}} \left[\frac{\beta_\ell^{\alpha_\ell}}{(\beta_\ell + 1)^{\alpha_\ell + \bar{x}_\ell}} \right]^n \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1}.$$

We want to compare the two populations through their coefficients of variation. For the negative binomial distribution the coefficient of variation is

$$\varphi_\ell = \frac{\sqrt{Var(X_\ell)}}{|E(X_\ell)|} = \sqrt{\frac{\beta_\ell + 1}{\alpha_\ell}}, \quad \ell = 1,2.$$

The hypothesis $H : \varphi_1 - \varphi_2 = 0$ identifies on the parameter space Θ the subsets

$$\Theta_j = \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}^2_+ \times \mathbb{R}^2_+ \;\middle|\; \alpha_2(\beta_1 + 1) \lesseqgtr \alpha_1(\beta_2 + 1) \right\}, \quad j = a, b$$

Variances of two Gamma populations Consider two independent random variables X_ℓ, $\ell = 1,2$, with distributions

$$X_\ell \sim Gamma\big(x_\ell | \alpha_\ell, \beta_\ell\big) = \frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} x^{\alpha_\ell - 1} e^{-\beta_\ell x_\ell}, \quad X_\ell \in \mathbb{R}^+, \; (\alpha_\ell, \beta_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+.$$

We are interested in performing a test on the variances, i.e. $\mathbb{V}ar(X_1) = \mathbb{V}ar(X_2)$ where $\mathbb{V}ar(X_\ell) = \frac{\alpha_\ell}{\beta_\ell^2} = \varphi_\ell$. The posterior distribution of the parameter vector, after assuming the Jeffreys' prior (see [5]), takes the expression

$$g_\ell(\alpha_\ell, \beta_\ell | \mathbf{x}_\ell) = k_\ell \cdot g(\alpha_\ell, \beta_\ell) \cdot L(\alpha_\ell, \beta_\ell \,|\, \mathbf{x}_\ell) \propto \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1} \left(\frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} g^{\alpha_\ell} e^{-\bar{x}_\ell \beta_\ell} \right)^n$$

with k_ℓ the normalising constant. The hypothesis $H : \varphi_1 - \varphi_2 = 0$ identifies on the parameter space Θ the subsets

$$\Theta_j = \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}^2_+ \times \mathbb{R}^2_+ \;\middle|\; \alpha_1 \beta_2^2 \lesseqgtr \alpha_2 \beta_1^2 \right\}, \quad j = a, b.$$

3.1 Example

We present an application to real data for the comparison of skewness indexes of two Inverse Gaussian populations. In the engineering context, river flooding rates have significant economic, social, political implications. The modelling and analysis of such data is an important application of extreme value theory. We consider a study on the Floyd Ever Flood rate data for the years 1935–1954. A more extended version of this dataset was analysed in [4] where it is shown that the exponentiated Weibull model is appropriate for such data, more in general for extreme value analysis, and that the Inverse Gaussian distribution belongs to this family.

In the years 1935–1944 we have that $n_1 = 10$ and $\hat{\varphi}_1 = 2.19$ while in the second group, for the years 1945–1954, $n_2 = 10$ and $\hat{\varphi}_2 = 5.20$. The computation of δ_H, as explained in the prevoius section, depends on the evaluation of two four-dimensional integrals (see (4)), that can be approximated through the Monte Carlo

integration method. Furthermore, since the posterior distribution is known up to a normalizing constant, it is not possible to sample directly from it. We, therefore, applied the random walk Metropolis-Hasting algorithm using a bivariate Normal proposal distribution. In order to obtain a chain that converges to the target distribution, $9 \cdot 10^5$ samples were generated and reduced to $1.2 \cdot 10^5$ after considering a burn-in period and a chain thinning. We, finally, found that the discrepancy measure is $\delta_H \approx 1$, then we can reject the null hypothesis of the skewness indexes equality.

4 Conclusions

In this paper we have proposed a testing procedure to compare parameter functions of two independent populations. Until now, only sharp hypotheses have been considered, but the extension to more complex hypotheses is under study. Theoretical and computational developments in more general contexts, such as linear or generalised models, ANOVA, model selection, are also at an advanced stage of investigation.

References

[1] Bertolino, F., Columbu, S., Manca, M., Musio, M.: Testing the equality of two coefficients of variation: a new bayesian approach (2022). URL https://arxiv.org/abs/2204.10147

[2] Bertolino, F., Manca, M., Musio, M., Racugno, W., Ventura, L.: A new bayesian discrepancy measure (2021). URL https://arxiv.org/abs/2105.13716

[3] Cox, D.R., Hinkley, D.V.: Theoretical statistics. CRC Press (1979)

[4] Mudholkar, G.S., Hutson, A.D.: The exponentiated weibull family: some properties and a flood data application. Communications in Statistics–Theory and Methods **25**(12), 3059–3083 (1996)

[5] Yang, R., Berger, J.O.: A catalog of noninformative priors. Institute of Statistics and Decision Sciences, Duke University (1996)

A contribution to the L. J. Savage problem

Sul problema di L. J. Savage

Francesco Bertolino, Silvia Columbu and Mara Manca

Abstract The aim of this contribution is to provide an answer, from a subjective Bayesian perspective, to the conceptual problem proposed in 1959 by Leonard Jimmie Savage. For this purpose we propose to apply the Bayesian Discrepancy Measure, which is a new evidence measure recently introduced in the literature.
Abstract *Lo scopo di questo lavoro è quello di fornire una risposta, in ottica bayesiana soggettiva, al problema concettuale proposto nel 1959 da Leonard Jimmie Savage. In questa direzione proponiamo di utilizzare la Misura di Discrepanza Bayesiana, una nuova misura di evidenza che è stata recentemente introdotta in letteratura.*

Key words: Bayesian test, elicitation, Bayesian Discrepancy Measure.

1 Problem description and objective solution

During the Summer Course in Varenna (1959) Savage [4] illustrated three isomorphic conceptual experiments. His explicit aim was to argue against the idea that statistical analysis, in order to be objective, cannot take into account the personal convictions of the statistician.

The first experiment involves a German professor (*GP*), expert in Mozart and Haydn music, who claims to be able to distinguish the two composers' sheets by simply listening to a few notes. The second experiment concerns Mrs Muriel Bristol (*MB*), the "Fisher's lady", who states she is able to distinguish the flavour of a cup of tea in which the tea has been poured first and then the milk later, from a cup in which the opposite happened. The protagonist of the third experiment is a drunk soothsayer (*DS*) who claims to be able to guess the results of flipping a coin by placing it on

Dept. of Mathematics and Computer Sciences, University of Cagliari, Via Ospedale, 72, Cagliari

his arm. The three individuals tested on 10 trials always give the correct answer, resulting in 10 successes.

If, in a frequentist setting, we consider the hypotheses

$$\left\{ H_0 : \theta = \frac{1}{2} \; vs \; H_1 : \theta \neq \frac{1}{2} \right\}, \tag{1}$$

with θ probability of success, we obtain a *p-value* $= \frac{1}{2^9} \cong 0.002$ and H_0 is rejected in the same way for all the three experiments.

Although, the experiment results observed suggest identical conclusions on H_0, it is reasonable to not be confident on the abilities claimed by *DS* in the third case. As a consequence, we may think to include our pre-experimental opinions in the formalisation of the problem.

For the moment, we perform the test under a standard objective Bayesian framework using the Bayes Factor. Since Savage does not specify the stopping rule adopted, two different situations can be considered.

- In the first situation, we fix the number of the trials and the observed successes, i.e. $n = s = 10$. Since $S|\theta \sim Bin(s \mid \theta, n)$, we assume the Jeffreys non-informative (proper) prior
$$g_0(\theta) = \frac{1}{\pi} \cdot \frac{1}{\sqrt{\theta\,(1-\theta)}}.$$
Given the likelihood $L(\theta) = \theta^s$, the Bayes factor is
$$BF_{0,1}(n,s) = \frac{\left(\frac{1}{2}\right)^s}{\pi^{-1} \displaystyle\int_0^1 \frac{\theta^s}{\sqrt{\theta(1-\theta)}}\, \mathrm{d}\theta} = 0.0098.$$
This result forces a strong rejection of H_0 and, again, all three experiments come to the same conclusion.
- In the second situation, we set the number $\nu = 10$ of successes and observe $y = 0$ failures. Given that $Y|\theta \sim NB(y \mid \theta, \nu)$ the Jeffreys non-informative prior is assumed
$$h_0(\theta) \propto \frac{1}{\theta\sqrt{1-\theta}}.$$
Since $h_0(\theta)$ is improper, we consider the proper distribution
$$h_b(\theta) = \frac{c_b}{\theta\sqrt{1-\theta}} \cdot \mathbf{1}_{(b,1)}(\theta), \quad b > 0,$$
where c_b, the normalising constant of the distribution, depends on b and is such that $\lim_{b\to 0} c_b = 0$. Provided the likelihood, $L(\theta) = \theta^\nu$, the Bayes factor is

$$BF^b_{0,1}(\nu, y) \;=\; \frac{\left(\frac{1}{2}\right)^{\nu}}{c_b \cdot \displaystyle\int_b^1 \frac{\theta^{\nu}}{\theta\sqrt{1-\theta}}\,\mathrm{d}\theta} \;=\; \frac{\left(\frac{1}{2}\right)^{\nu}}{c_b \cdot B_b\left(\nu, \frac{1}{2}\right)},$$

where $B_b(\cdot,\cdot)$ is the incomplete beta function. Notice that $\lim_{b\to 0} B_b\left(\nu, \frac{1}{2}\right) = B\left(\nu, \frac{1}{2}\right)$ and $\lim_{b\to 0} c_b = 0$ implies $\lim_{b\to 0} BF^b_{0,1}(\nu, y) \to \infty$. Hence, the hypothesis H_0 cannot be rejected. In addition, in this situation we run into the well known Jeffreys-Lindley paradox.

In conclusion, it can be argued that by using a proper prior the hypotheses are rejected in the same way in all the experiments, whereas the use of an improper prior leads to the Jeffreys-Lindley paradox. Furthermore, we obtain opposite results even though the likelihoods are the same.

It should be pointed out that even when dealing with intrinsic and fractional Bayes factors, which are also part of objective Bayesian methods (see [1] and [3]), Savage's three isomorphic experiments still receive equal results.

2 A subjective solution proposal through the BDT

Following Savage's guidelines, we now attempt a subjective Bayesian analysis in which the three isomorphic experiments are considered separately. To this aim, we propose to use a new Bayesian evidence measure recently introduced in the literature (see [2]).

The Bayesian Discrepancy Measure Let $X \sim f(x|\theta)$ be a parametric model indexed by a scalar parameter $\theta \in \Theta = \mathbb{R}$, we indicate with $G(\theta|\mathbf{x})$ the posterior distribution function and with $m_1 = G^{-1}\left(\frac{1}{2}|\mathbf{x}\right)$ the posterior median.
We are interested in testing the sharp hypothesis $H_0 : \theta = \theta_{H_0}$ and we propose to do it by considering the Bayesian Discrepancy Measure (BDM), defined through the distance between the posterior median and the hypotheses as

$$\delta_{H_0} = 2 \cdot \mathbb{P}(\theta \in I_{H_0}|x) = 2 \cdot \int_{I_{H_0}} dG(\theta|x), \tag{2}$$

where I_{H_0} is the discrepancy interval, expressed as

$$I_{H_0} = \begin{cases} (m_1, \theta_{H_0}) \text{ if } & m_1 < \theta_{H_0} \\ \{m_1\} \quad \text{if } & m_1 = \theta_{H_0} \\ (\theta_{H_0}, m_1) \text{ if } & m_1 > \theta_{H_0} \end{cases}. \tag{3}$$

We have values of δ_{H_0} close to 1 when θ_{H_0} is far from the median of the posterior distribution $G(\theta|\mathbf{x})$. Following this idea the Bayesian Discrepancy Test (BDT), to which this measure is associated, after fixing a threshold $\omega \in \{0.95, 0.99, 0.995, \dots\}$ allows to reject H_0 when we have values above it.

The subjective problem Although Savage's three experiments are isomorphic, they should be treated separately. Considering our beliefs on them we specify the following *Beta* priors.

1. Being confident of *GP*'s capabilities we set the mode in $\theta = 1$ and $\mathbb{P}_{GP}(\theta \geq 0.90) = \frac{1}{2}$.
2. Having less confidence in *MB*'s capabilities we set the prior mean $\frac{\alpha_0}{\alpha_0+\beta_0} = \frac{3}{4}$ and $\mathbb{P}_{MB}(\theta \leq \frac{1}{2}) = 0.15$.
3. Being completely dubious about *DS*'s divinatory abilities we set $\mathbb{E}(\theta) = \frac{1}{2}$, which implies that $\alpha_0 = \beta_0$, and $\mathbb{P}_{SD}\left(\frac{1}{2} - 0.05 < \theta \leq \frac{1}{2} + 0.05\right) = 0.90$.

It should be noticed that it is pointless to evaluate the hypothesis $H_0 : \theta = \frac{1}{2}$ in the first two cases, whereas it is appropriate to evaluate it in the third case. Therefore,

1. the prior probability is $\mathbb{P}_{GP}(\theta \geq 0.90) = \frac{1}{2}$, while the posterior probability is $\mathbb{P}_{GP}(\theta \geq 0.90 \mid n,s) = 0.833$. This outcome indicates a consolidation of our initial opinions, but it is still insufficient to determine the protagonist's flawlessness.
2. In this case we move from a prior probability $\mathbb{P}_{MB}\left(\theta \geq \frac{3}{4}\right) = 0.321$ to a posterior probability $\mathbb{P}_{MB}\left(\theta \geq \frac{3}{4} \mid n,s\right) = 0.865$. This result leads us to conclude that *MB* can boast a tea tasting sensibility that not many people possess.
3. In order to verify the hypothesis $H_0 : \theta = \frac{1}{2}$ we evaluate the BDM that, starting from the posterior median $m_1 = 0.5193$, is

$$\delta_{H_0} \;=\; 2 \cdot \int_{1/2}^{m_1} g_1^{DS}(\theta \mid n,s)\,\mathrm{d}\theta \;=\; 0.466\,,$$

   which indicates that there is not enough evidence to reject H_0.

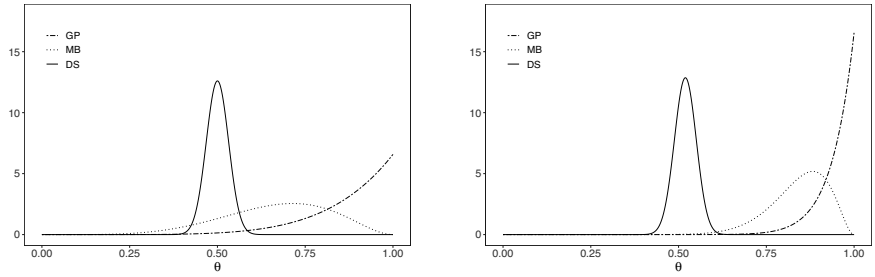
Table 1 shows the prior hyperparameters of the Beta distributions proposed and their posterior values of the conjugate model. The associated distribution plots are shown in Figure 1.

prior			
hyperparameters	*GP*	*MB*	*DS*
α_0	7	6	125
β_0	1	3	125

posterior			
hyperparameters	*GP*	*MB*	*DS*
α_1	17	16	135
β_1	1	3	125

Table 1: Hyperparameters of the prior (left) and posterior (right) distributions.

From a subjective perspective, conclusions about the supposed sensitive faculties of *DS* arise as a combination of an objective fact, $n = s = 10$, and the researcher belief/scepticims towards *DS*, expressed through the choice of the prior hyperparameter α_0.



(a) Prior distributions (b) Posterior distributions

Fig. 1: Probability density functions for *GP*, *MB*, *DS*.

In eliciting the prior we have imposed that $\mathbb{E}(\theta) = \frac{1}{2}$, whence $\alpha_0 = \beta_0$, and we have been conservative in choosing $\alpha_0 = 125$. It should also be taken into account that the higher the disbelief in *DS*, the more concentrated the prior distribution is in $\theta = \frac{1}{2}$.

What happens when the opinions of more than one individual have to be taken into account, i.e. what happens when considering more than one value of α_0? In Figure 2 we have summarised the behaviour of the BDM by varying the hyperparameter values and observed that when α_0 increases, the BDM tends to decrease. In particular, when $\alpha_0 \leq 8.5$ the hypothesis H_0 is rejected since $\delta_{H_0} > 0.95$.

Another aspect that is worth discussing is the dependence of rejection conclusions on the observed sample size. Regarding *DS*'s case, suppose for example to have $n = s = 16$ and consider again the hypothesis test on $H_0 : \theta = \frac{1}{2}$. In this scenario a frequentist analysis, with a p-value of 3×10^{-5}, will again suggest the rejection of the hypothesis. On the other hand, the use of a prior with $\alpha_0 = \beta_0 = 125$, proposed in the former discussion, would not allow such a conclusion as we obtain $\delta_{H_0} = 0.6743$. In this second experimental setting, letting vary α_0 we observe that the threshold of rejection has increased moving from $\alpha_0 = 8.5$ to $\alpha_0 = 26$. In the figure it is also reported the behaviour of the discrepancy measure for a third potential setting where $n = s = 8$.

3 Conclusions

Savage's problem describes a situation in which the pre-experimental information outweighs the one of the experiment. Arguing that under these circumstances any objective procedure (whether frequentist or Bayesian) provides no real help to statistical analysis, Savage suggests that the effort of elicitation is unavoidable.

Moving from Savage's criticism, we proposed to address the problem from a subjective perspective by introducing priors which reflect differing opinions on the three experiments. The BDT has then been applied to test the abilities of *DS* in the

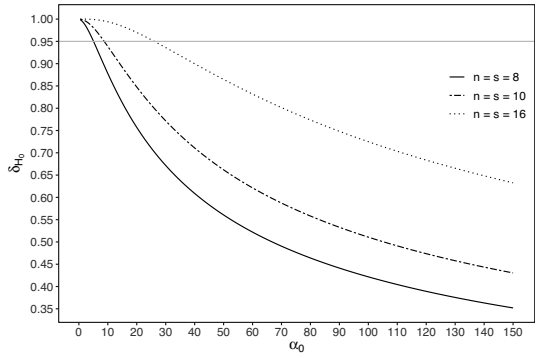


Fig. 2: Sensitivity analysis showing δ_{H_0} values in dependence of α_0.

third experiment. We remark that we have intentionally avoided performing a similar test in the remaining two situations as testing a similar hypothesis would be in contrast with the elicitation. In fact, there is a strict connection between the selection of the hypothesis to test and the opinions expressed through the prior elicitation.

Finally, for Savages' third experiment, we reported a preliminary sensitivity analysis on the hyperparameters' choice. We have also investigated the behaviour of the BDM for different samples sizes with fixed proportion of successes.

Although the subjective analysis was here approached using the BDM, the Bayes factor could also have been applied. We intend to do it in a future work in which the relationship between the BDM and the Bayes factor will be also investigated.

References

[1] Berger, J.O., Pericchi, L.R.: The intrinsic bayes factor for model selection and prediction. Journal of the American Statistical Association **91**(433), 109–122 (1996)

[2] Bertolino, F., Manca, M., Musio, M., Racugno, W., Ventura, L.: A new bayesian discrepancy measure (2021). URL https://arxiv.org/abs/2105.13716

[3] O'Hagan, A., Forster, J.J.: Kendall's advanced theory of statistics, volume 2B: Bayesian inference, vol. 2. Arnold (2004)

[4] Savage, L.J.: La probabilità soggettiva nei problemi pratici della statistica. In: Induzione e statistica, pp. 123–200. Springer (2011)

Methods for Complex Data

Optimization of Delayed Rejection Adaptive Metropolis

Ottimizzazione del Metropolis Adattivo con Rifiuto Ritardato

Daniele Raffo and Antonietta Mira

Abstract We present an improved Delayed Rejection Adaptive Metropolis algorithm that significantly speeds up sampling. The computational optimization consists in a remodulation of the covariance updating formula, using Cholesky factors and a closed-form expression for the log-proposal ratio used in the second stage of delayed rejection, for both Gaussian and Student's t proposal distributions. The original and improved samplers are run on a test distribution with a fixed runtime and their results are compared; empirically, the optimized algorithm has an efficiency higher by an order of magnitude with respect to the original one.

Abstract *Presentiamo un algoritmo Metropolis Adattivo con Rifiuto Ritardato migliorato che accelera significativamente il campionamento. L'ottimizzazione computazionale consiste in una rimodulazione della formula di aggiornamento della covarianza, utilizzando la fattorizzazione di Cholesky e un'espressione chiusa per il rapporto della log-proposta utilizzato nel secondo stadio del rifiuto ritardato, per entrambe le distribuzioni di proposta Gaussiane e Student's T. Il campionatore originale e quello migliorato sono impiegati su una distribuzione di test con un runtime fisso e i loro risultati vengono comparati; empiricamente, l'algoritmo ottimizato ha un'efficienza maggiore di un ordine di grandezza rispetto a quello originale.*

Key words: Bayesian samplers, Delayed Rejection, numerical methods

---

Daneiele Raffo
Sapienza University of Rome - Faculty of Information Engineering, Informatics, and Statistics, Rome, 00185, Italy
and Università della Svizzera Italiana - Faculty of Economics, Lugano, 6900, Switzerland e-mail: daniele.raffo@unine.ch

Antonietta Mira
Università della Svizzera Italiana - Faculty of Economics, Lugano, 6900, Switzerland
and University of Insubria, Varese, 21100, Italy e-mail: antonietta.mira@usi.ch

1 Introduction

Adaptive Metropolis (AM, [1]) is a variation of the standard Random Walk Metropolis (RWM) algorithm that adapts the covariance matrix of the proposal distribution at each simulation step, using the path of the chain up to that step. In RWM every time we reject the proposal and retain the previous position of the Markov chain autocorrelation increases and, as a result, the asymptotic variance of the resulting MCMC estimator also increases. The mechanism of Delayed Rejection (DR, [2]) produces chains that are guaranteed to be less autocorrelated than those obtained by RWM: if at a certain iteration of the algorithm the proposal is rejected, instead of immediately moving to the next iteration, a second stage move is proposed. The second stage proposal distribution can be updated and may depend on the rejected value as well as on the current position of the Markov chain. Delayed Rejection Adaptive Metropolis (DRAM, [3]) blends these two concepts: the covariance of the proposal at the first stage is adapted as in AM, while that of the second stage is computed as a scaled version of the one used in the first stage as in DR.

The Robust Adaptive Metropolis sampler (RAM, [4]) gave us the inspiration for the optimization of DRAM: instead of directly updating the covariance matrix of the proposal distribution, RAM updates the corresponding Cholesky factor and this translates in a significant speed-up in both the update operation itself and the generation of proposal moves.

Furthermore the proposal ratio which is needed at the second stage of the delayed rejection step can be swiftly computed in closed form for Gaussian and Student's t proposal distributions, avoiding the need to compute the density twice. Since computing the density of these proposal distributions requires the inversion of their covariance matrices, this represents a bottleneck for the efficiency of the algorithm, especially in high dimensions.

2 Methods

Instead of the recursive formula reported in [1], for the update of the covariance matrix in the **Adaptive Metropolis** (**AM**) algorithm we refer to the equivalent formula reported on [5], since the latter is easier to compare to the updated routine used in the optimized version of the algorithm. The covariance matrix $C_n \in \mathbb{R}^{d\times d}$, where d is the number of parameters (or, equivalently, the rank of C_n), used in iteration $n = 1,\ldots,N$ is defined as:

$$C_n = \begin{cases} C_0, & n = 1,2 \\ \frac{n-2}{n-1}C_{n-1} + \frac{s_d}{n}(X_n - \bar{X}_{n-1})(X_n - \bar{X}_{n-1})^T, & n > 2. \end{cases} \tag{1}$$

where $C_0 \in \mathbb{R}^{d\times d}$ is a given initial covariance matrix which may reflect prior beliefs (in all our applications, this is set equal to the identity matrix $I_d \in \mathbb{R}^{d\times d}$), $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, the draws $X_n \in \mathbb{R}^d$ being column vectors, and

s_d is a scaling parameter that we take to be $s_d = 2.4^2/d$ following [6]. The sample mean also satisfies a recursive formula:

$$\bar{X}_n = \begin{cases} X_1 & n = 1, \\ \bar{X}_{n-1} + \frac{1}{n}(X_n - \bar{X}_{n-1}) & n > 1. \end{cases} \tag{2}$$

We can reformulate (1) in terms of the Cholesky factor L_n of C_n; we have, for $2 < n \leq n_T$:

$$L_n L_n^T = \frac{n-2}{n-1} L_{n-1} L_{n-1}^T + \frac{s_d}{n}(X_n - \bar{X}_{n-1})(X_n - \bar{X}_{n-1})^T. \tag{3}$$

Solving (3) for L_n is equivalent to perform a rank-one Cholesky update of $\sqrt{\frac{n-2}{n-1}} L_{n-1}$ given a random vector $\sqrt{\frac{s_d}{n}}(X_n - \bar{X}_{n-1})$:

$$L_n = \text{CholeskyUpdate}\left(\sqrt{\frac{n-2}{n-1}} L_{n-1}, \sqrt{\frac{s_d}{n}}(X_n - \bar{X}_{n-1}) \right). \tag{4}$$

The rank-one Cholesky update is a $O(d^2)$ operation (see [7]); hence, it is of the same asymptotic computational complexity as calculating (1), since that involves the sum of two matrices.

At iteration n, a candidate move, Y_1, is generated from a symmetric proposal distribution $q(X_n, C_n)$ and accepted with the usual Metropolis probability

$$\alpha(X_n, Y_1) = \min\left\{ 1, \frac{\pi(Y_1) q_1(Y_1, X_n)}{\pi(X_n) q_1(X_n, Y_1)} \right\} = \min\left\{ 1, \frac{\pi(Y_1)}{\pi(X_n)} \right\}. \tag{5}$$

where $\pi(\cdot)$ is the target distribution. Note that the proposal distribution cancels in the above acceptance ratio because of its symmetry: $q_1(Y_1, X_n) = q_1(X_n, Y_1)$. If we accept we move to Y_1, i.e. $X_{n+1} = Y_1$; otherwise we stay at the current position, i.e. $X_{n+1} = X_n$.

With **Delayed Rejection** (**DR**), if the first proposal move $Y_1 \sim q(X_n, C)$ is rejected, instead of immediately moving to the next iteration and retaining the current position X_n, a second stage move $Y_2 \sim q(X_n, \gamma C)$ is proposed. γ was set to 0.5 in all our applications (i.e., we halve the covariance matrix). The acceptance probability at the second stage depends not only on the current position of the chain, but also on what we have just proposed and rejected:

$$\alpha_2(X_n, Y_1, Y_2) = \min\left\{ 1, \frac{\pi(Y_2) q_2(Y_2, Y_1)[1 - \alpha_1(Y_2, Y_1)]}{\pi(X_n) q_2(X_n, Y_1)[1 - \alpha_1(X_n, Y_1)]} \right\}. \tag{6}$$

Here, the factor $Q = \frac{q_2(Y_2, Y_1)}{q_2(X_n, Y_1)}$ is called proposal ratio and cannot be simplified further. However, it can be computed in closed form for log-Gaussian and log-Student's

t proposals, respectively:

$$Q_{\mathscr{MVN}_{\log}} = -\frac{1}{2}(Y_2 - Y_1)^T \Omega (Y_2 - Y_1) + \frac{1}{2}(X_n - Y_1)^T \Omega (X_n - Y_1), \tag{7}$$

$$\begin{aligned} Q_{\mathscr{MVT}_{\log}} = &-d \log\left(1 + \frac{1}{d}(Y_2 - Y_1)^T \Omega (Y_2 - Y_1)\right) \\ &+ d \log\left(1 + \frac{1}{d}(X_n - Y_1)^T \Omega (X_n - Y_1)\right). \end{aligned} \tag{8}$$

In an iteration of the vanilla implementation of the **Delayed Rejection Adaptive Metropolis** (**DRAM**), the worst-case scenario in terms of computational complexity happens when the first proposal is rejected and the second proposal must be sampled and the acceptance probability must be computed. Table 1 summarizes the operations that are carried out, their computational complexity and how many times each of them is executed; it is shown how the total complexity of the vanilla algorithm is $5O(d^3) + O(d^2) + 2K$[1] (where K varies with the specific target distribution).

Operation	Complexity	Times
Computing target $\pi(X)$	K	xx
Sampling $Y \sim q(X,C)$	$O(d^3)$	xx
Computing the density at a point $q_j(X,Y)$	$O(d^3)$	xxx
Update covariance matrix C_n	$O(d^2)$	x

Table 1 Summary of operations carried out by the vanilla algorithm.

However, if the Cholesky decomposition L_n of the covariance matrix C_n is available beforehand, as in our optimization, we can sample the first proposal Y_1 having distribution $q(X_n, C_n)$ in the following way:

$$Y_1 = X_n + L_n U_1, \tag{9}$$

and the second proposal (if required) as:

$$Y_2 = X_n + \sqrt{\gamma} L_n U_2, \tag{10}$$

where $U_1, U_2 \sim q(\mathbf{0}, I_d)$ are random vectors sampled from the standardized version of the current proposal distribution. Obviously, sampling from a standardized distribution does not require finding the inverse of the covariance matrix, hence it is an $O(1)$ operation, and sampling with in the method in Equations (9) and (10) is therefore an $O(d^2)$ operation, as it involves a matrix-vector multiplication.

Computing the log-proposal ratio eliminates the need of computing the densities $q(X,Y)$ and, as it only involves matrix-vector multiplications (see Equations (7) and (8)), it is $O(d^2)$.

[1] Note that, at each iteration with $n \geq 2$, we have computed $\pi(X_n)$ at the precedent iteration.

We can then summarize the operations carried out in the worst-case scenario by the optimized algorithm in Table 2, and see how the computational cost is brought down to $4O(d^2)+2K$.

Operation	Complexity	Times
Computing target $\pi(X)$	K	xx
Sampling as $Y = X + LU$	$O(d^2)$	xx
Computing log proposal ratio $Q_{\log}$	$O(d^2)$	x
Update Cholesky factor L_n	$O(d^2)$	x

Table 2 Summary of operations carried out by the optimized algorithm.

3 Results

Optimized DRAM was tested against its vanilla implementation, both with Gaussian and Student's-t proposal distribution, on the Rosenbrock probability distribution, also called Banana distribution for its particular shape. The distribution is defined in d dimensions and has the following kernel [8]:

$$f(x) \propto \exp\left(\sum_{i=1}^{d-1}\Big[-100(x_{i+1}-x_i^2)^2-(x_i-1)^2\Big]\right), \tag{11}$$

with $x=(x_1,...,x_d)\in\mathbb{R}^d$ being the random vector. It is a unimodal distribution whose most of the density lies in a narrow, parabolic region of the parameter space. The global maximum is $f(x^*)=1$ at $x^*=(1,\ldots,1)$. In particular, the samplers were used to approximate the distributions for $d=2,3$ on the hypercube $x_i\in[-5,10]$ for all $i=1,\ldots,d$.

We measured the number of iterations performed in 2 seconds of runtime by the different versions of the algorithm. The starting point for each simulation was $x_0=0\in\mathbb{R}^d$. We did not use burn-in as our main focus was to evaluate the efficiency of the "adaptive" iterations. AM and DR were used at each step. Each simulation run was repeated $k=1000$ Monte Carlo times to analyze the distribution of the results and assess their variability. The results are reported in Table 3.

In the 2D case, the mean number of iterations is increased by a factor of about 15-fold for both Gaussian and Student's-t proposal, from the vanilla version to the optimized version; in the 3D case, this gain is already about 75-fold, with the number of iterations performed by the optimized version almost unchanged from 2D to 3D. This highlights how the computational advantages are greater in higher dimensions. Furthermore, the variability of the results is much larger when using the classical implementation.

	2-dimensional				3-dimensional			
	Normal		Student's-t		Normal		Student's-t	
	Van	Opt	Van	Opt	Van	Opt	Van	Opt
Min.	52	10025	55	7495	44	8520	41	9214
1st Qu.	656	12620	563	11821	105	12027	90	11417
Median	855	12739	726	11938	158	12146	144	11519
Mean	884	13367	754	12499	169	12714	153	12087
3rd Qu.	1070	12874	926	12055	216	12282	197	11621
Max.	2957	24585	2420	21990	964	22108	1092	21382

Table 3 Summary statistics of the number of iterations performed by the compared algorithms in 2 seconds. Van = vanilla version; Opt = optimized version of the algorithm.

4 Discussion

We showed how the optimization of the DRAM algorithm improves its speed significantly; this makes it a valid choice in applied settings, especially when the target distribution has a complex shape and/or the acceptance rate of alternative sampling algorithms is low. Optimized DRAM has been used successfully in the estimation of a Markov-Switching GARCH model for the computation of the conditional volatility of financial assets, and proved to be more efficient than other samplers, also producing chains with a higher degree of stationarity.

References

1. Haario, H., Saksman, E., Tamminen, J: An Adaptive Metropolis Algorithm. Bernoulli, 7(2), 223–242 (2008)
2. Tierney L, Mira A: Some adaptive monte carlo methods for Bayesian inference. Stat Med. 15-30;18(17-18):2507-15. (1999)
3. Haario, H., Laine, M., Mira, A. et al. DRAM: Efficient adaptive MCMC. Stat Comput 16, 339–354 (2006)
4. Matti, V.: Robust adaptive Metropolis algorithm with coerced acceptance rate. Statistics and Computing. 22 (2010)
5. Laine, Marko: Adaptive MCMC Methods With Applications in Environmental and Geophysical Model. Meteorol. Inst. Contrib.. 69. (2008)
6. Gelman, A., Roberts, G. O., Gilks, W. R. In Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M.: Efficient Metropolis jumping rules. Bayesian Statistics, pages 599-608. Oxford University Press, Oxford (1996)
7. Krause, O., Igel, C.: A More Efficient Rank-one Covariance Matrix Update for Evolution Strategies. In Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII (FOGA '15). Association for Computing Machinery, New York, NY, USA, 129–136. (2015)
8. Pagani, F, Wiegand, M, Nadarajah, S: An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. Scand J Statist.; 1– 24 (2021)

Dealing with multicollinearity and outliers in multinomial logit model: a simulation study

Una soluzione alla multicollinearità e ai valori anomali nel modello logit multinomiale: uno studio di simulazione

Ida Camminatiello and Antonio Lucadamo

Abstract Multinomial logit model is a powerful tool for modeling the dependence relationship between a set of quantitative regressors and a categorical response variable with more than two categories. However, it necessitates careful consideration of multicollinearity among the regressors and examination for outliers. For dealing with these problems in the multinomial logit model, an approach based on robust principal components has been developed. The robust approach has achieved satisfactory results on real data. This paper aims to validate such an approach by a simulation study.

Abstract *Il modello logit multinomiale é un potente strumento per modellare la relazione di dipendenza tra un insieme di regressori quantitativi e una variabile dipendente nominale con più di due categorie. Tuttavia, richiede un'attenta considerazione della multicollinearitá tra i regressori e l'osservazione dei valori anomali. Per affrontare questi problemi nel modello logit multinomiale, é stato sviluppato un approccio basato sulle componenti principali robuste. L'approccio robusto ha ottenuto risultati soddisfacenti su dati reali. Questo paper mira a convalidare tale approccio mediante uno studio di simulazione.*

Key words: principal component analysis, multinomial logit model, outliers, simulation study.

Ida Camminatiello
University of Campania, Capua (CE), Italy, e-mail: ida.camminatiello@unicampania.it

Antonio Lucadamo
University of Sannio, Benevento, Italy, e-mail: antonio.lucadamo@unisannio.it

1 Introduction

Multinomial logit (MNL) model allows explaining a response variable on nominal scale in terms of the predictors (McFadden, 1974). The model is widely used in many fields, however, it can be affected by multicollinearity and outlier problems.
In presence of multicollinearity, the stepwise selection of the predictors can be performed. Principal component multinomial regression (Aguilera, Escabias, 2008; Camminatiello, Lucadamo, 2010; Lucadamo, Leone, 2015) has been proposed as an alternative way of solving consequences of multicollinearity.
This estimation method solves the multicollinearity problem; however, the results can be affected by extreme values in the continuous predictors, thus there is a need for some robust estimation methods. Camminatiello and Lucadamo (2018) developed an approach based on robust principal components (PCs), called robust principal component multinomial regression (RobPCMR), so that the effect of both multicollinearity and outliers is minimized in the estimation of the MNL model parameters (second section). The robust approach applied for assessing judges' performances (Camminatiello, Lucadamo 2018) showed good results. However, an extensive simulation study is needed to verify that RobPCMR is resistant to many types of contamination, to compare the results with other robust approaches for PCs proposed in the literature, and to select the optimal dimension of the model. For this purpose, the third section presents a first simulation study. The last section concludes with some remarks and perspectives.

2 Robust principal component multinomial regression

When multicollinearity among the regressors exists, the estimation of MNL model becomes inaccurate because of the need to invert nearsingular and ill-conditioned information matrices (Ryan, 1997). Therefore, the covariates of the MNL model could be substituted by a reduced number of PCs of the regressors. Because the PCs are based on the eigenvectors of the empirical covariance matrix, they are very sensitive to anomalous observations. Several methods for robustifying principal component analysis (PCA) have been proposed in literature. RobPCMR considers a recently developed method, called ROBPCA (Hubert, Rousseeuw and Vanden Branden, 2012), for its interesting properties. In a nutshell RobPCMR can be synthesized as follow.
Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_p]$ is the $n \times p$ matrix of the p predictor variables and $\mathbf{y}$ the $n \times 1$ vector of categorical response variable with more than two categories observed on n statistical units.
At first step, RobPCA creates the robust PCs of the regressors $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_p]$ which are linear combinations of the original variables, $\mathbf{Z} = \mathbf{XV}$. At second step the MNL model is carried out on the set of p robust PCs. At third step, the number of robust PCs, $a \leq p$, to be retained in the model, is selected according to different tools (Camminatiello, Lucadamo, 2010). Finally the MNL model is carried out on the chosen subset of robust PCs. The MNL model can be expressed in terms of a robust

PCs as:

$$\pi_i^{(a)}(c) = \frac{exp(\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jc}^{(a)})}{\sum_{b=1}^{s} exp(\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jb}^{(a)})} = \frac{exp(\sum_{k=1}^{a} z_{ik} \gamma_{kc}^{(a)})}{\sum_{b=1}^{s} exp(\sum_{k=1}^{a} z_{ik} \gamma_{kb}^{(a)})} \quad (1)$$

where $\gamma_{kb}^{(a)} = \sum_{j=1}^{p} v_{kj} \beta_{jb}^{(a)}$ are the robust coefficients to be estimated on the subset of a robust PCs and $\beta_{jb}^{(a)}$ are the robust parameters obtained after the extraction of the a components. Finally, the robust MNL parameters can be expressed in function of original variables $(\mathbf{x_1}, ..., \mathbf{x_p})$

$$\mathbf{Z^{(a)}} \gamma^{(\mathbf{a})} = \mathbf{X V^{(a)}} \gamma^{(\mathbf{a})} = \mathbf{X} \beta^{(\mathbf{a})} \quad (2)$$

where $\beta^{(a)} = V^{(a)} \gamma^{(a)}$ is the matrix of robust parameters expressed in function of original variables; $\mathbf{Z}^{(a)}$ is the matrix of selected robust PCs; $\gamma^{(a)}$ is the matrix of robust parameters on a robust PCs for the s alternatives; $\mathbf{V}^{(a)}$ is the matrix of robust eigenvectors for the selected components.
We observe that the approach can be applied to low - $n > p$ - as well as high dimensions - $n \leq p$.
To measure the performance of a model, several criteria can be applied (Camminatiello, Lombardo, Durand, 2017; Camminatiello, Lucadamo, 2010). Here, we focus on rate of well classified which we expect higher compared to Principal component multinomial regression (PCMR) and MNL.

3 A simulation study.

The goodness of the proposed methodology has been investigated through some simulation studies. Analyses have been performed for data with and without outliers and 100 simulations for each scenario have been considered.

Genuine data have been generated according to the following scheme:

1. Building of the matrix $(\mathbf{X})$ of explicative variables: each variable has been generated from a normal distribution with mean 0 and standard deviation 1. Different correlation structures among variables have been considered.
2. Generation of the dependent variable using the multinomial distribution to define the membership to each group.

Contamination has been introduced by replacing a proportion $\varepsilon = 0.1, 0.2$ and 0.4 of genuine observations, by outliers generated from a normal distribution with the same variance than before, but mean equal to 20. The scenario we introduce in this paper is the following:

- number of observations: $n = 1000$
- number of predictors: $p = 20$

- number of modalities of response variable: 3
- correlation among predictors: 0.9

To evaluate the goodness of RobPCMR we can consider table 1, in which the correct classification rate for classical multinomial logit model, principal component multinomial Regression and RobPCMR with different level of resistence (α) to outliers are compared. In this analysis we decided to retain in the analysis, for all the methods, only the significant variables or components, according to a stepwise selection based on AIC.

Table 1 Correct classification rate (%) for different contamination levels

	Methods				
	Multinom.	PCMR	RobPCMR($\alpha = 0.90$)	RobPCMR($\alpha = 0.80$)	RobPCMR($\alpha = 0.60$)
No contamination					
Estimation set	75.66	75.64	74.60	74.53	74.64
Validation set	72.06	72.78	71.84	72.50	72.20
$\varepsilon = 0.1$					
Estimation set	73.68	72.31	71.86	71.50	71.59
Validation set	69.73	68.81	70.01	70.93	70.15
$\varepsilon = 0.2$					
Estimation set	71.74	71.32	69.96	69.96	70.11
Validation set	67.56	67.83	68.03	68.36	67.86
$\varepsilon = 0.4$					
Estimation set	67.79	67.59	66.06	66.39	66.07
Validation set	62.54	61.29	63.77	62.98	63.63

Looking at the values in table, we can see that the rate of well classified for the estimation sample is always higher for the classical multinomial logit model, but we are interested almost to what happens for the validation set. When there is not contamination, it is easy to notice that the PCMR works better than other methods, but the RobPCMR, for all levels α, lead to good results. When we consider data with outliers, the RobPCMR shows, for the validation set, higher performance than other methods, independent of the resistence level. Obviously deeper simulation studies are needed to better understand the validity of the proposed method.

4 Remarks and perspectives

In this paper, we presented a first simulation study to verify that RobPCMR is a valid method for dealing with multicollinearity and outlier problems in MNL. However, further simulation studies are needed to show that RobPCMR is resistant to many types of contamination, to compare the results with other robust approaches for PCs proposed in the literature, to consider different correlation structures among predictors and variations in the numbers of observations, explicative variables, and modalities of the dependent variable and, to select the optimal dimension of the

model.
Moreover, the rate of well classified is a performance measure widely applied in literature, but other criteria - such as the estimator variance or the predicted residual error sum of squares - could be discussed. Finally, a robust method for ordinal logistic regression could be proposed.

References

1. Aguilera, A., Escabias M.: Solving Multicollinearity in Functional Multinomial Logit Models for Nominal and Ordinal Responses. In: Dabo-Niang, S., Ferraty, F. (eds.) Functional and Operatorial Statistics. Contributions to Statistics, pp. 7-13 Physica-Verlag HD (2008)
2. Bastien, P., Esposito Vinzi, V. Tenenhaus, M.: PLS Generalised Linear Regression. Computational Statistics Data Analysis **48**, 17–46 (2005)
3. Camminatiello, I., Lombardo, R., Durand, J.F.: Robust partial least squares regression for the evaluation of justice court delay. Qual Quant (2017) doi: https://doi.org/10.1007/s11135-016-0441-z
4. Camminatiello, I., Lucadamo, A.: Estimating multinomial logit model with multicollinear data. Asian Journal of Mathematics and Statistics **3** (2), 93–101 (2010)
5. Camminatiello I., Lucadamo A.: A robust multinomial logit model for evaluating judges' performances. In: Abbruzzo A., Brentari, E., Chiodi, M., Piacentino, D. (eds) Book of short papers SIS (2018)
6. Hosmer, D. W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons, Inc. New York (2000)
7. Hubert, M., Rousseeuw, P.J., Vanden Branden, K.: ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics (2012) doi: 10.1198/004017004000000563
8. Lucadamo, A., Leone, A.: Principal component multinomial regression and spectrometry to predict soil texture. J Chemometr. **29** (9), 514-520 (2015).
9. McFadden, D.: Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P. (ed), Frontiers in Econometrics, pp. 105-142 Academic Press, New York (1974).
10. Nur Aidi, M., Purwaningsih, T.: Modeling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island. International Journal of Statistics and Applications **3**(1), 1-8 (2013)
11. Ryan T.P.: Modern Regression Methods, Wiley, New York (1997).

A tool to validate the assumptions on ratios of nearest neighbors' distances: the Consecutive Ratio Paths

Uno strumento per validare le assunzioni sui rapporti di distanze fra vicini più vicini: le traiettorie di rapporti consecutivi

Francesco Denti and Antonietta Mira

Abstract The estimation of the intrinsic dimension is an essential step in many data analyses involving, for example, dimensionality reduction. Likelihood-based estimators, which rely on the distributions of the ratios of distances between nearest neighbors, have been recently proposed. However, these distributional results depend on several assumptions. One of the most important is the local homogeneity of the point process characterizing the data-generating mechanism. By exploiting a recent theoretical result, we develop the Consecutive Ratio Paths, a graphical tool to assess the validity of the local-homogeneity assumption in a dataset. This tool is also helpful to uncover the presence of multiple latent manifolds, a potential indicator of the existence of heterogeneous intrinsic dimensions.

Abstract *La stima della dimensione intrinseca è un passaggio essenziale in molte analisi dei dati che comportano, per esempio, la riduzione della loro dimensionalità. Recentemente sono stati proposti stimatori basati sulla verosimiglianza, costruiti partendo dalle proprietà delle distribuzioni dei rapporti delle distanze tra i vicini più vicini. Questi risultati distributivi dipendono da diverse ipotesi. Una delle più importanti è l'omogeneità del processo di punto che caratterizza il meccanismo di generazione dei dati. Sfruttando un recente risultato teorico, proponiamo di monitorare le traiettorie di rapporti consecutivi, fornendo uno strumento grafico per valutare la validità dell'ipotesi di omogeneità locale in un set di dati. Questo strumento è utile anche per scoprire la presenza di molteplici manifold latenti, evidenziando l'eventuale esistenza di dimensioni intrinseche eterogenee.*

Key words: intrinsic dimension, Pareto distribution, nearest neighbors, model-based estimation, graphic tool

Francesco Denti
Università Cattolica del Sacro Cuore, Milan; e-mail: francesco.denti@unicatt.it

Antonietta Mira
Università della Svizzera italiana, Lugano and University of insubria, Como

1 Introduction

Dimensionality reduction techniques are valuable tools to ensure the feasibility of statistical analyses involving large datasets. These methods are used to obtain parsimonious but meaningful representations of the data over a latent (potentially non-linear) manifold with dimensionality lower than the original one, preserving as much information as possible. We call the dimensionality of the latent manifold the *intrinsic dimension* (`id` from now on). It is essential to rely on an accurate `id` estimate to perform effective dimensionality reduction. Recently, [3] introduced the `TWO-NN` model, a likelihood-based `id` estimator that builds on the properties of the ratios of distances between data points and their first two nearest neighbors (NNs). The `TWO-NN` model has been recently extended in [2], where the authors propose an estimator that relies on the ratios of NN of generic order, called `Gride`. This modeling framework produces reliable estimates for the `id` given that all the underlying modeling assumptions, on which we elaborate more in the next section, are satisfied. However, it is not always trivial to assess the correctness of these assumptions to perform model validation. In this paper, we propose the Consecutive Ratio Paths (`CoRaP`), a graphical tool to assess the compliance of the data to the assumptions mentioned above. The paper proceeds as follows. First, Section 2 briefly presents the modeling framework and the distributional results, which are the foundations of the `CoRaP`, and delineates our contribution. Then, Section 3 discusses the results on simulated data, explaining how to interpret the plots that we propose to assess the assumptions. Finally, Section 4 concludes the article, discussing potential future directions.

2 The modeling background

Consider a dataset $\boldsymbol{X} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^D,\ \forall i$. The dataset $\boldsymbol{X}$ is viewed as a realization of a Poisson point process with constant density ρ defined over $\mathbb{R}^d$. We also postulate that the data points take place on a manifold of dimension $d < D$. Then, given a distance function δ taking values in $\mathbb{R}^+$, let $r_{i,l} = \delta(x_i, x_{i,(l)})$ be the value of this distance between observation i and its l-th NN $x_{i,(l)}$. [3] proved that the ratios between the second and first order NNs' distances is distributed as

$$\mu_{i,2} = \frac{r_{i,2}}{r_{i,1}} \sim Pareto(1,d), \qquad \mu_{i,2} \in (1,+\infty), \qquad i = 1,\dots,n. \tag{1}$$

Equation (1) states that one can recover the `id` of a dataset, d, by estimating the scale parameter of a Pareto distribution using the sample of ratios $\boldsymbol{\mu}_{i,2} = \{\mu_{i,2}\}_{i=1}^n$. One can only prove the result in (1) if the density ρ is constant (*global homogeneity*). However, the derived `id` estimator is empirically valid as long as the density is approximately constant on the scale defined by the distance from the furthest of the NNs, $r_{i,2}$. We refer to this weakened assumption as *second order local homogeneity*.

Thus, we can summarize the principal modeling assumptions of the `TWO-NN` in the following three points: (a) the computed ratios of NN distances are independent, (b) the underlying latent manifold on which the data take place is unique, and (c) the data generating mechanism is characterized by a density function that is *second order locally homogeneous*. [2] extended the result in Equation 1, showing that, under the same theoretical framework, one can obtain that, for the ratios of the l-th to the $(l-1)$-th order NN distances, the following is true:

$$\mu_{i,l} = \frac{r_{i,l}}{r_{i,l-1}} \sim Pareto(1,(l-1)\cdot d), \quad \mu_{i,l} \in (1,+\infty) \quad i = 1,\dots,n, \quad l = 1,\dots,L, \tag{2}$$

where $L < n$. Moreover, it can be proved that the entries of the vector $(\mu_{i,2},\dots,\mu_{i,L})$ are all independent.
Equation (2) is important because it provides the link between the `TWO-NN` modeling framework and the likelihood estimators proposed by [4] and corrected in [5]. The estimator derived from (2) would require even more stringent hypotheses since the intensity function has to be approximately constant on the scale defined by the distance from the furthest of the NNs for each point i, $r_{i,L}$, with $L > 2$ (called *L-th order local homogeneity*). These strong hypotheses are unlikely to be completely satisfied in many real scenarios. Nonetheless, in the next section, we show how the distributional result stated in Equation (2) can be exploited to assess the validity of the underlying modeling assumptions.

2.1 The Consecutive Ratios Paths

The previous results states that, under ideal conditions, the consecutive ratios of distances are independent and Pareto distributed. Given well-known properties of the Gamma and Pareto distributions, we consider the transformation $\eta_{i,l} = (l-1)\cdot \log(\mu_{i,l})$, which ensures that $\eta_{i,l} \sim Exp(d)$. Then, for every observation $i = 1,\dots,n$, we compute the vector of running means $\bar{\boldsymbol{\eta}}_i$ over the increasing orders of NNs $l = 2,\dots,L$ defined as

$$\bar{\boldsymbol{\eta}}_i = (\bar{\eta}_i(2),\dots,\bar{\eta}_i(L)), \quad \text{where} \quad \bar{\eta}_i(l) = \sum_{k=2}^{l} \eta_{i,k}/(l-1). \tag{3}$$

We call the vector $\bar{\boldsymbol{\eta}}_i$ the *Consecutive Ratio Path* for the i-th observation (`CoRaP`$_i$), since it summarizes the evolution of the behavior of the distances between x_i and its NNs. If all the required assumptions are met, for each combination (i,l) we have $(l-1)\cdot\bar{\eta}_i(l) \sim Gamma(l-1,d)$, implying that $\mathbb{E}[\bar{\eta}_i(l)] = 1/d$ and $\mathbb{V}[\bar{\eta}_i(l)] = 1/((l-1)d^2)$. In other words, the ideal `CoRaP`$_i$ is a collection of random variables characterized by a constant expected value across i and l, identically equal to the reciprocal of the `id`, and by decreasing variance. Moreover, for l^* large enough, we can exploit the Central Limit Theorem, stating that $\bar{\eta}_i(l^*) \approx \mathcal{N}\left(1/d, 1/((l^*-1)d^2)\right)$,

where with $\mathcal{N}(\mu, \sigma^2)$ indicates a Normal distribution with mean μ and variance σ^2.
Our goal is to compare the behavior of the observed $\texttt{CoRaP}_i$ with the expected theoretical counterpart to detect violation of hypotheses. The ideal case (**Case A**) is displayed in the top left panel of Figure 1, and it is obtained simulating $Exp(d = 5)$ random numbers populating a matrix with $n = 1{,}000$ observations and considering $L = 750$ NN order. We can appreciate how the `CoRaP` concentrate around the true value $d = 1/5$. Also, the left panel of Figure 2 showcases the distribution of $\bar{\eta}_i(1{,}000)$, for which the Gaussian asymptotic approximation holds very well.

3 Detecting violation of assumptions via CoRaP

We discuss the interpretation of the `CoRaP` with the help of three additional synthetic datasets, generated as follows:

- **Case B**: 1,000 data points sampled from a five-dimensional Normal distribution (one manifold, unique `id`);
- **Case C**: 1,000 data points sampled from two bi-dimensional, non-overlapping Normal distributions (two manifolds with identical `id`);
- **Case D**: 1,000 data points sampled from two different Normal distributions with `id` of 2 and 7, respectively (two manifolds with different `ids`).

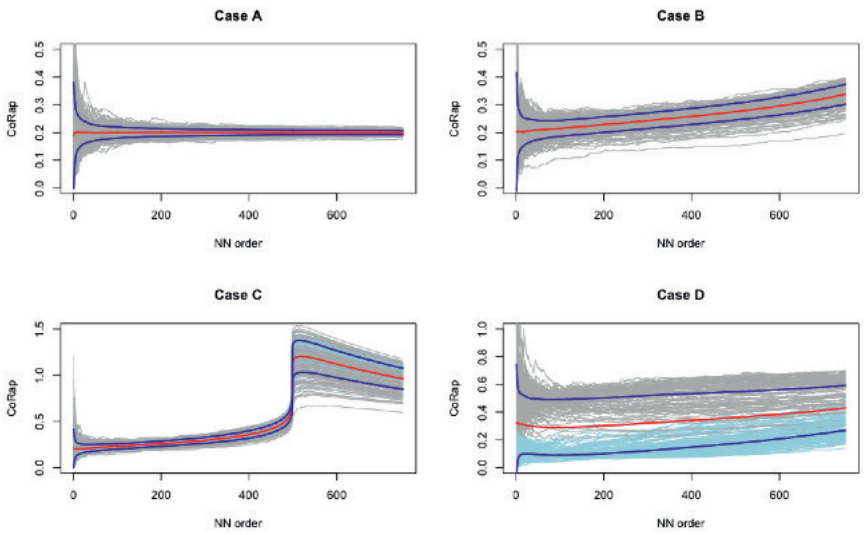


Fig. 1 The `CoRaP` obtained from datasets generated under the four different scenarios (Cases A, B, C, and D) are depicted in gray and/or light blue. The red lines represent the overall ergodic mean (computed point-wise), and the blue lines help quantify the uncertainty ($\pm$ std.dev.).

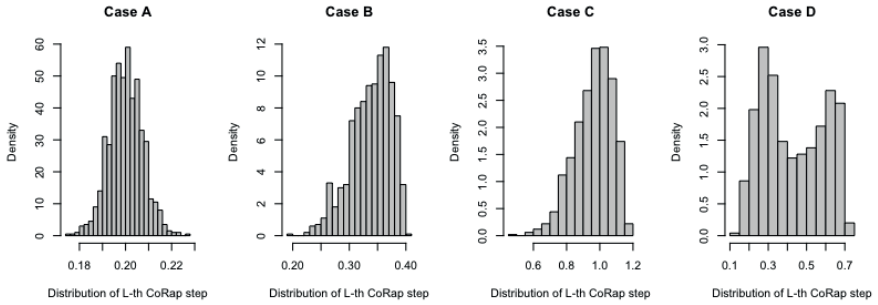


Fig. 2 Histograms of the distributions $\bar{\eta}_i(750)$, under ideal theoretical conditions (Case A) and the other three scenarios considered in the simulation study (B, C, and D).

Figure 1 displays the `CoRaP` for all the observations in the four datasets up to the NN order of $L = 750$. Similarly, Figure 2 reports the histograms of the distributions for $\bar{\eta}_i(L),\ \forall i$. In both figures, the remaining three panels illustrate characteristically different behaviors that the `CoRaP` might manifest in applications. From the top-right panel (Case B) in Figure 1, we notice how the inhomogeneity induced by the tails of the Normal distribution is reflected by a slight departure from the ground truth ($d = 1/5$) as L grows. The shift in the distribution is evident from the corresponding histogram (second panel) in Figure 2.
The bottom two panels (Cases C and D) in Figure 1 unveil more interesting patterns. The former describes data characterized by the same `id` but grouped in different clusters. The corresponding `CoRaP` capture this behavior showing a discontinuity point at $L = 500$. Similarly, the latter panel presents the case where the data possess more than one latent manifold: the resulting `CoRaP` are separated into two clusters, which is evident from the different starting points for low values of l. Each group presents well-behaved `CoRaP`, but pooling the data together produces misleading results (e.g., the overall ergodic mean is not representative of the data). The presence of multiple manifolds is also captured in the right panel of Figure 2: the resulting distribution presents multimodality.

4 Conclusion

In this paper, we introduced the `CoRaP`, a model-based, intuitive graphical summary of a dataset. Monitoring the evolution of the `CoRaP` allows to detect departures from the theoretical assumptions required by the `TWO-NN` and `Gride` models. In particular, we were able to detect the presence of different manifolds with heterogeneous `id`, which represents an important violation of the assumption. This result is a precious indicator that models tailored to identify, for a given dataset, the presence of multiple latent manifolds – each one with its own `id` – should be used. An example

of this type of models is `Hidalgo` [1]. We will extend and improve the preliminary results presented in this contribution. For example, the multimodality encountered in the last histogram of Figure 2 can be exploited to devise a fast algorithm for multiple `id` estimation leveraging on EM for mixtures of Gaussians. Moreover, the model-based nature of our results paves the way for the derivation of statistical tests concerning the topological structure of datasets.

References

[1] Michele Allegra, Elena Facco, Francesco Denti, Alessandro Laio, and Antonietta Mira. Data segmentation based on the local intrinsic dimension. *Scientific Reports*, 10(1):1–27, 2020.

[2] Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. Distributional Results for Model-Based Intrinsic Dimension Estimators. *ArXiv preprint*, 2021.

[3] Elena Facco, Maria D'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):1–8, 2017.

[4] Elizaveta Levina and Peter J Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In L K Saul, Y Weiss, and L Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, 2005.

[5] D.J.C. MacKay and Z. Ghahramani. Comments on 'Maximum Likelihood Estimation of Intrinsic Dimension' by E. Levina and P. Bickel (2004). *Comment on personal webpage*, 2005.

Dimensionality reduction and visualization for interval-valued data via midpoints-ranges principal component analysis

Riduzione della dimensionalità per variabili a intervallo attraverso l'analisi delle componenti principali su centri e raggi

Viviana Schisa, Alfonso Iodice D'Enza, and Francesco Palumbo

Abstract Interval data proved to be a useful coding in big data applications since many individuals can be collapsed into a single unit, so that each cell of the table contains a bounded interval. This contribution presents the Principal Component Analysis (PCA) for interval-valued data based on midpoints and radii approach, where the symbolic standardization is used, along with an ad-hoc Procrustean rotation.

Abstract *I dati ad intervallo rappresentano una codifica efficiente di grandi quantità di dati: numerose osservazioni possono essere sintetizzate in una, che contiene l'intervallo dei valori osservati sulle singole osservazioni. Il presente contributo descrive l'analisi in componenti principali per dati a intervallo, basata sull'approccio centri-raggi e la standardizzazione simbolica.*

Key words: Interval-Valued Data; Principal Component Analysis; Symbolic Standardization; Procrustean Rotation.

1 Introduction

In several statistical analysis problems, uncertainty and imprecision in the data suggest considering interval-valued variables instead of the more classical single-valued variables. There are also situations where data are natively recorded as intervals: e.g., describing a species of plants or animals, and more in general when the data refers to groups of units and not to single units [1]. Therefore, each single interval *datum* is represented by $[x] \equiv [\underline{x}, \overline{x}]$, $\overline{x} \geq \underline{x}$, and $\underline{x}$ and $\overline{x}$ indicate the min and max values of the interval. Then *interval algebra* approach postulates that the knowledge about the interval is limited to its extremes values, and from a geometric point of view, $[x] \subset \mathbb{R}$ (is a closed bounded interval in $\mathbb{R}$). The generic interval value $[x]_i$ is

Department of Political Sciences, Università degli Studi di Napoli Federico II, Naples, Italy
e-mail: schisaviviana@gmail.com, iodicede@unina.it, fpalumbo@unina.it

defined as

$$[x]_i = [\underline{x}_i, \overline{x}_i] \qquad\qquad i = 1, \ldots, n,$$

and can equivalently be represented in the center $x_i^c = \frac{1}{2}(\overline{x}_i + \underline{x}_i)$ and the range $x_i^r = \frac{1}{2}(\overline{x}_i - \underline{x}_i)$ notation. So the interval $[x]_i \equiv \{x_i^c, x_i^r\}$. The initial formalization of the PCA for interval data is by [8]: since then, several contributions into the interval-valued data analysis were published, and some of these are referenced in the following suggested reconsidering in a new light the MR-PCA. Recently, from the same authors, an article proposed a new MR-PCA algorithm that is a significant improvement of the method [4]. Moreover, it is worth mentioning that the routines of the new MR-PCA are now available written in the CRAN-R meta-language: no MR-PCA code was officially or unofficially delivered up to now.

The present contribution shortly summarizes the changes in the new MR-PCA algorithm (see for details [4]), and it reports a case study based on the well-known `abalone` dataset (`http://www.uci.edu/`).

2 Interval-valued variables pre-treatment

Data pre-treatment represents a crucial aspect in any factorial analysis. In classical PCA, the most classical pre-treatment consists in centering by the arithmetic mean $\bar{x}$ and scaling the variables by the standard deviation s. Therefore, all involved variables have the mean equal to 0 and variance and standard deviation equal 1. Moreover, scaling produces dimensionless variables and each principal component results as a weighted sum of the scaled variables. The classical pre-treatment cannot be adopted when dealing with interval-valued variables, namely centers and ranges. First, the definition of the classical standard deviation no longer applies to interval data; second, centers and ranges must be treated independently but consistently to reconstruct the best-approximated interval data on the factorial axes. In their revised MR-PCA, to pre-process interval-valued data consistently with their intrinsic nature, [4] used the definition of the *symbolic mean* and the *symbolic covariance matrix* Σ_{symb} for interval-valued variables [7], which guarantees advantages with respect to the original proposal.
Let $[\underline{x}_{ij}, \overline{x}_{ij}]$ be the interval-valued realization of the j-th variable for the i-th observation $(j = 1, ..., p;\ i = 1, ..., n)$, where $\underline{x}_{ij} \leq \overline{x}_{ij}$. Then the *symbolic* mean $\bar{w}_j$ and the *symbolic* variance $\sigma^2_{j_{symb}}$ of $[x]_j$ are defined according to the following formulae:

$$\overline{w}_j = \frac{1}{2n}\sum_{i=1}^{n}(\underline{x}_{ij} + \overline{x}_{ij}) \ \text{ and } \ \sigma^2_{j_{symb}} = \frac{1}{3n}\sum_{i=1}^{n}(\underline{x}^2_{ij} + \underline{x}_{ij}\overline{x}_{ij} + \overline{x}^2_{ij}) - \overline{w}^2_j. \qquad (1)$$

It is worth noting that $\bar{w}_j$ corresponds to the mean of the midpoints for the j-th variable. The variance (1) can be extended to the bivariate case for the covariance for j and j':

$$\sigma_{jj'_{symb}} = \frac{1}{6n}\sum_{i=1}^{n}[2(\underline{x}_{ij}-\bar{w}_j)(\underline{x}_{ij'}-\bar{w}_{j'}) + (\underline{x}_{ij}-\bar{w}_j)(\bar{x}_{ij'}-\bar{w}_{j'}) + (\bar{x}_{ij}-\bar{w}_j)(\underline{x}_{ij'}-\bar{w}_{j'}) + 2(\bar{x}_{ij}-\bar{w}_j)(\bar{x}_{ij'}-\bar{w}_{j'})] \quad (2)$$

Then, the *symbolic covariance matrix* Σ_{symb} has diagonal and extra diagonal elements (1) and (2), respectively.
Let $\tilde{\mathbf{X}}^c$ and $\tilde{\mathbf{X}}^r$ be the midpoint and range matrices $(n \times p)$; in the following $\mathbf{X}^c$ and $\mathbf{X}^r$ denote the centered versions. We propose the *symbolic standardization* of the midpoints and radii as follows:

$$\mathbf{Z}^c = \mathbf{X}^c\mathbf{S}^{-1} = \left(\tilde{\mathbf{X}}^c - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^c\right)\mathbf{S}^{-1};$$
$$\mathbf{Z}^r = \mathbf{X}^r\mathbf{S}^{-1} = \left(\tilde{\mathbf{X}}^r - \frac{1}{n}\mathbf{1}\mathbf{1}'\tilde{\mathbf{X}}^r\right)\mathbf{S}^{-1},$$

where $\mathbf{S}$ is a diagonal matrix and $s_{jj} = \sqrt{\sigma^2_{jj_{symb}}}$, for $j = 1, \ldots, p$.

3 Centers and ranges joint representation on the factors

The other relevant issue in interval-valued variables PCA is the interval data reconstruction into the factor space. To this aim, several approaches have been proposed in the literature [2, 3, 5, 7]. The leading principle in the MR-PCA consists in assuming that the position of each interval data on the factor space depends on the center only. As a consequence, the ranges must be represented as supplementary information in the way that it must be as much as possible consistent with the centers factor subspace. Then the standardized and centered ranges are projected after a suitable rotation that maximize the Tucker congruence index. The revised version proposes a more efficient and more consistent algorithm starting from the one introduced by [8], but maximizing the average congruence [10] with respect to all factors (and not just with the first one, as in the first version).
The rotation matrix $\mathbf{T}$ is such that $\mathbf{Z}^c$ is as close as possible to the rotated version of $\mathbf{Z}^r$. Therefore the definition of $\mathbf{T}$ is obtained solving the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{T}}: \quad & tr\left(\mathbf{Z}^c - \mathbf{Z}^r\mathbf{T}\right)\left(\mathbf{Z}^c - \mathbf{Z}^r\mathbf{T}\right)' \\ & tr\left(\mathbf{Z}^c\mathbf{Z}^{c\prime}\right) + tr\left(\mathbf{Z}^r\mathbf{T}\mathbf{T}'\mathbf{Z}^{r\prime}\right) - 2tr\left(\mathbf{Z}^{c\prime}\mathbf{Z}^r\mathbf{T}\right) \quad s.t. \quad \mathbf{T}\mathbf{T}' = \mathbf{I}. \end{aligned} \quad (3)$$

It is easy to see that the problem in (3) can be re-written as

$$\max_{\mathbf{T}}: tr\left(\mathbf{Z}^{c\prime}\mathbf{Z}^r\mathbf{T}\right) \qquad s.t. \qquad \mathbf{T}\mathbf{T}' = \mathbf{I}, \quad (4)$$

which is equivalent to the maximize the correlation coefficient among the columns of the centers matrix and the corresponding columns of the ranges matrix:

$$\sum_{j=1}^{p} \frac{\mathbf{t}_j' \mathbf{Z}^{r\prime} \mathbf{z}_j^c}{(\mathbf{t}_j' \mathbf{Z}^{r\prime} \mathbf{Z}^r \mathbf{t}_j)^{1/2} (\mathbf{z}_j^{c\prime} \mathbf{z}_j^c)^{1/2}},$$

where $\mathbf{t}_j$ and $\mathbf{z}_j^c$ are the j-th column of the $n \times p$ matrices $\mathbf{T}$ and $\mathbf{Z}^c$, respectively. The solution of the optimization problem in (4) is obtained via the iterative procedure proposed by [6].
The coordinates for centers (midpoints) and ranges (radii) are

$$\boldsymbol{\Psi}^c = \mathbf{Z}^c \mathbf{U}_{(d)}^c \quad \text{and} \quad \boldsymbol{\Psi}^r = \mathbf{Z}^r \mathbf{T} \mathbf{U}_{(d)}^r,$$

where $\mathbf{U}_{(d)}^c$ and $\mathbf{U}_{(d)}^r$ are the first d columns of the eigenvector matrices of $\mathbf{Z}^{r\prime}\mathbf{Z}^r$ and $\mathbf{Z}^{r\prime}\mathbf{Z}^r$, respectively.
The projection of the i^{th} range on the α^{th} component is

$$\psi_{i,\alpha} = [(\psi_{i\alpha}^c - \psi_{i\alpha}^r), (\psi_{i\alpha}^c + \psi_{i\alpha}^r)],$$

where $\psi_{i,\alpha}^c$ and $\psi_{i,\alpha}^r$ are the coordinates of the i^{th} center and range on the α^{th} axis. Finally, for interval-valued data the analogue of the PCA relative contributions is given by

$$ctr\_rel_{i\alpha} = \frac{\sum_\alpha (|\psi_{i,\alpha}^c| + |\psi_{i,\alpha}^r|)^2}{\sum_{j=1}^{p} (|z_{i,j}^c| + |z_{i,j}^r|)^2}.$$

4 Application: Abalone data set

The `abalone` data set consists of 4177 observations described through 9 variables: sex (discarded), length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and number of rings. The observations were aggregated according to the number of rings of each abalone. Note that rings indicate the age of the mollusk and that some groups consist of only one sample implying that units are characterized by *tiny (degenerated) intervals* [8] in the available data set. The example shows that tiny intervals do not impact the analysis feasibility.

For the sake of space, the article reports only the most significant results; in particular, the attention focuses on the first factorial plane plots. To illustrate the MR-PCA results on the Abalone data set, it is worth starting from the maps of the loadings in Figure 1, where arrows represent loadings and the spread of the maps is proportional to the overall variability accountable to centers and ranges, respectively. The more significant proportion of variability comes from the centers. The left-hand side plot in Figure 1 helps to understand the centers' correlation structure and explain the units positioning in Figure 2. In contrast, the right-hand side plot

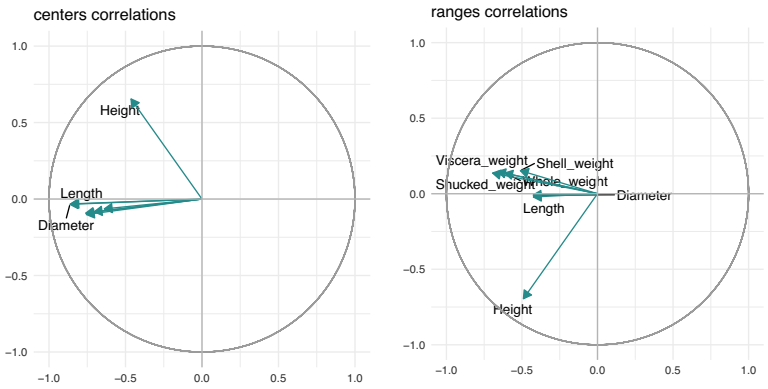


Fig. 1 Symbolic covariance-based standardization: two-dimensional maps of the centers (midpoints) and ranges (radii) loadings.

shows ranges' correlation structure and their orientations and lengths in the plot. In Figure 2, rectangles transparency refers to the quality of the representation, measured by the ratio between the length of the projections of the rotating range and the original length of the standardized range. In this example, the centers-ranges correlation structure allows for high values of quality of the representation on the first two dimensions. Looking at the MR-PCA map, the ranges orientations on the factorial plan allow tracing back the variations to the source variables: e.g., *rings8* and *rings10* ranges have the same direction meaning that they depend on the same variables; *rings3* and *rings8* ranges have opposite orientations. Returning on the un-

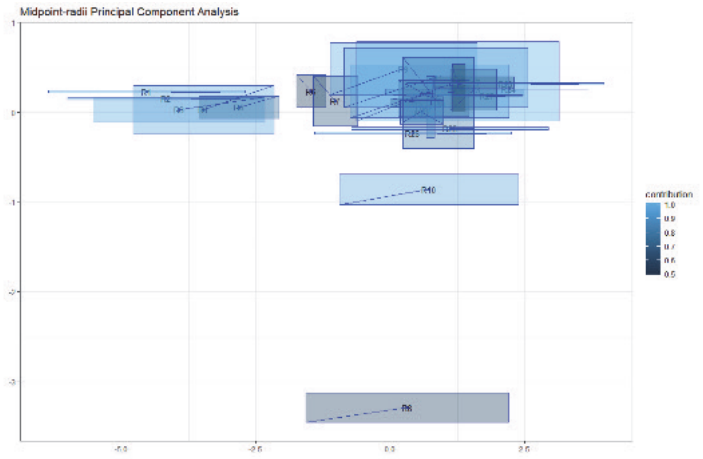


Fig. 2 MR-PCA: observations map (82.90% explained inertia). Procrustean rotation of ranges. Darker rectangles indicate lower relative contributions

rotated ranges representations and having the ranges correlations map, it is possible to identify the variables that have mainly affected the range length and orientations.

5 Concluding remarks and future perspectives

Before the contributions of the Symbolic Data Analysis group [1], at the end of the nineties, the treatment of interval-valued variables was relegated to some applications in the numerical analysis to deal with the round-off error caused from the fixed-point CPUs [9]. Uncertainty and imprecision were and still are, treated under the fuzzy logic paradigm [11]. However, the quick growth of the available data is inducing new interest through interval data analysis. In particular in the processes control. The number of citations to the early papers focused on PCA for interval-valued variables proves it. This new interest also touched our sensibility. First, we realized that software to perform the MR-PCA does not exist; second, we realized that the methodology needed a revision. The article[4] introduces these novelties. However, to validate the efficacy and the consistency, new theoretical results are expected, accompanied by empirical evidence, as the present contribution is.

References

1. Bock, H.H., Diday, E.: Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media (1999)
2. Chouakria, A., Diday, E., Cazes, P.: Vertices principal components analysis with an improved factorial representation. In: Advances in data science and classification, pp. 397–402. Springer (1998)
3. Giordani, P., Kiers, H.A.: A comparison of three methods for principal component analysis of fuzzy interval data. Computational Statistics & Data Analysis **51**(1), 379–397 (2006)
4. Iodice D'Enza, A., Schisa, V., Palumbo, F.: Principal component analysis for interval data: common approaches and variations. Italian Journal of Applied Statistics **33**(3), 249–270 (2022)
5. Irpino, A., Lauro, C., Verde, R.: Visualizing symbolic data by closed shapes. In: Between data science and applied data analysis, pp. 244–251. Springer (2003)
6. Kiers, H.A., Groenen, P.: A monotonically convergent algorithm for orthogonal congruence rotation. Psychometrika **61**(2), 375–389 (1996)
7. Le-Rademacher, J., Billard, L.: Symbolic covariance principal component analysis and visualization for interval-valued data. Journal of Computational and Graphical Statistics **21**(2), 413–432 (2012)
8. Palumbo, F., Lauro, C.N.: A PCA for interval-valued data based on midpoints and radii. In: New developments in Psychometrics, pp. 641–648. Springer (2003)
9. Rohn, J.: Inner solutions of linear interval systems. Lecture Notes in Computer Science (212), 157–158 (1986)
10. Tucker, L.R.: A method for synthesis of factor analysis studies. Tech. rep., Educational Testing Service Princeton Nj (1951)
11. Zadeh, L.A.: Fuzzy logic–a personal perspective. Fuzzy Sets and Systems **281**, 4–20 (2015). Special Issue Celebrating the 50th Anniversary of Fuzzy Sets

Data-driven design-based mapping of forest resources

Mappatura di risorse forestali in approccio da disegno

S. Franceschi, R.M. Di Biase, L. Fattorini, M. Marcheselli and C. Pisani

Abstract The mapping of forest resources in a study region is approached when the region is partitioned into spatial units by means of a completely data driven, design-based sampling strategy. When auxiliary variables are available for all the units, the prediction of the densities of an interest attribute can be performed by using an assisting model. Under these circumstances, the model residuals are interpolated using the inverse distance weighting interpolator with a data-driven smoothing parameter selection, and the density of the attribute for each unit is obtained by summing prediction and interpolated error. Finally, densities are rescaled to match the traditional total estimate with the sum of mapped values. The uncertainty is accounted for by a bootstrap procedure. A simulation study is performed and a case study is presented.
Abstract *La mappatura di risorse forestali per le unità spaziali che compongono una regione di studio viene affrontata con un approccio data-driven basato sul disegno. Se sono disponibili variabili ausiliarie, la previsione del valore dell'attributo di interesse può essere effettuata con strategie assistite da modello. I residui del modello vengono previsti utilizzando l'interpolatore inverse distance weighting dove il parametro di smorzamento viene selezionato con una procedura data-driven e le densità dell'attributo vengono ottenute sommando le previsioni e gli errori interpolati. Inoltre le densità vengono riscalate in modo tale che la stima del totale ottenuta sommando i valori previsti coincida con quella ottenuta con approcci tradizionali. L'incertezza delle stime è valutata attraverso una procedura bootstrap. Uno studio di simulazione viene effettuato e viene presentato un caso di studio.*

Key words: selection rule, smoothing parameter, regression estimator, harmonization, pseudo-population bootstrap

---

Sara Franceschi, Lorenzo Fattorini, Marzia Marcheselli and Caterina Pisani
Dipartimento di Economia Politica e Statistica e-mail: lorenzo.fattorini@unisi.it,e-mail: sara.franceschi@unisi.it,e-mail: marzia.marcheselli@unisi.it,e-mail: caterina.pisani@unisi.it

Rosa Maria Di Biase
Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, e-mail: rosamaria.dibiase@unimib.it

1 Introduction

Wall-to-wall maps of forest attributes are essential information sources for forest management and most of the methodologies applied for mapping rely on model-based inference, owing to the impossibility to estimate values at non-sampled units without any assumption.

We present an alternative approach to forest mapping, in which an assisting model is used, but uncertainty stems from the adopted sampling scheme, in accordance to a model-assisted perspective [1]. However, any interpolator adopted for mapping can achieve statistical soundness only if it is proven to be design-based asymptotically unbiased and consistent (DBAU&C). Conditions ensuring DBAU&C, derived by [3] for finite populations of spatial units when the sole spatial information is available, are extended to allow for the exploitation of auxiliary sources of information, such as those provided by satellites and aircraft-based sensors, available for each spatial unit (e.g., pixels) partitioning the study area.

In particular, we propose a completely data-driven, design-based strategy for mapping forest attributes. The first step is the selection of a model using an Akaike-type criterion for choosing an appropriate set of available auxiliary variables. Using the least-squares method, the predictions of the attribute densities are obtained for each spatial unit. Then, the inverse distance weighting (IDW) interpolator, which relies on Tobler's first law of geography [9], is adopted for interpolating the residuals for non-sampled units on the basis of the residuals that are known for each sampled unit. Following [5], the leave-one-out cross-validation is used to select the smoothing parameter adopted in the IDW interpolator. Subsequently, the interpolated values at non-sampled units are obtained by summing model predictions and interpolated residuals.

Finally, we rescaled the resulting densities in such a way that the total obtained from the interpolated values matches traditional total estimates [7]. Moreover, a bootstrap technique is used to estimate relative root mean squared errors of the predicted values.

2 Data-driven mapping

Consider a study area partitioned into N spatial units and suppose to be interested in the reconstruction of the whole map of the spatial population, that is the estimation of the density of an interest attribute for each spatial unit, by using a suitably selected sample of units. It is worth noting that usually the size of the spatial units are known and therefore the estimation of the density is theoretically equivalent to the estimation of the total amount in each spatial unit. Furthermore, suppose that a large set of covariates is available, as it is common in forest inventories. To select the assisting model for obtaining predictions, we performed a variable selection using an Akaike-type criterion, as suggested by [2]. Indeed, strongly correlated auxiliary

variables can induce instability in the model while they can exhibit a poor prediction capability when weakly correlated to the interest variable.

Let f_j and $\mathbf{x}_j$ respectively be the density of the interest attribute and the vector of selected covariates for spatial unit j, in such a way that densities can be expressed as $f_j = \boldsymbol{\beta}^t \mathbf{x}_j + e_j(\boldsymbol{\beta})$. The choice of $\boldsymbol{\beta}$ can be made using a least-square criterion to minimize the residual sum of squares. Denoting by $\mathbf{b}$ the ordinary least-squares solution for $\boldsymbol{\beta}$, if $\mathbf{b}$ was known, the residuals for each sampled unit would be $e_j(\mathbf{b}) = f_j - \mathbf{b}^t \mathbf{x}_j$, while the residuals for non-sampled units could be obtained using the IDW interpolator. In particular, the j-th interpolated residual would be given by

$$\hat{e}_j(\mathbf{b}) = Z_j e_j(\mathbf{b}) + (1 - Z_j) \sum_{i=1}^{N} w_{i,j}(\alpha) e_j(\mathbf{b})$$

where Z_j is equal to 1 when the j-th unit is sampled and 0 otherwhise and $w_{ij}(\alpha) = Z_i d_{i,j}^{-\alpha} / \sum_{l=1}^{N} Z_l d_{l,j}^{-\alpha}$, with $d_{i,j}$ denoting the distance between the centroid of unit i and unit j, and $\alpha > 2$ is the smoothing parameter. As consequence, the resulting interpolated value for the density for unit j would be $\hat{f}_j(\mathbf{b}) = \mathbf{b}^t \mathbf{x}_j + \hat{e}_j(\mathbf{b})$.

Unfortunately, the least-square solution $\mathbf{b}$ is unknown, involving the knowledge of the density for each unit. Nevetherless, as suggested by [8], an estimate $\hat{\mathbf{b}}$ can be obtained as a function of Horvitz-Thompson estimators. Then, IDW interpolation can be performed with $\mathbf{b}$ replaced by its sampling estimator $\hat{\mathbf{b}}$ on the basis of the observed residuals $e_j(\hat{\mathbf{b}}) = f_j - \hat{\mathbf{b}}^t \mathbf{x}_j$ so that the interpolated densities are given by $\hat{f}_j(\hat{\mathbf{b}}) = \hat{\mathbf{b}}^t \mathbf{x}_j + \hat{e}_j(\hat{\mathbf{b}})$.

As to the asymptotic properties, $\hat{f}_j(\hat{\mathbf{b}})$ is consistent because, as the extent of the spatial units partitioning the study area decreases and their number increases, $\hat{\mathbf{b}}$ converges to $\mathbf{b}$ and, consequently, $\hat{f}_j(\hat{\mathbf{b}})$ converges to $\hat{f}_j(\mathbf{b})$, which in turn is DBAU&C under conditions derived in [3]: i) the existence of a Riemann integrable function giving the density of the interest attribute at any point of the study area; ii) some sort of regularity in the shape of the spatial units; iii) the use of an asymptotically balanced spatial sampling scheme. Commonly adopted sampling schemes ensuring DBAU&C are simple random sampling without replacement, one-per-stratum stratified sampling and systematic sampling.

Moreover, the IDW interpolator depends on the smoothing parameter α determining the roughness of the surface of interpolated residuals. As suggested by [5], the value of α can be selected by means of the leave-one-out cross-validation and the corresponding IDW interpolator is proven to remain DBAU&C.

Finally, it is worth noting that, in a design-based setting, the total of the interest attribute is commonly estimated by traditional estimators, such as the regression estimator, which necessarily give rise to different total estimates with respect to the one achieved by summing the interpolated values for all the spatial units. To obtain non-discrepant results, a harmonization of the estimated map can be achieved by rescaling density estimates in analogy with [7].

As to the evaluation of precision, each step of the proposed mapping strategy (Akaike-type selection of the assisting model, choice of the smoothing parameter and harmonization of maps with the total estimate) should be considered, since all

of them are sample dependent. The use of bootstrap seems to be the sole way for facing the complexity of this data-driven mapping procedure. More precisely, the map of the interpolated values is taken as a pseudo-population from which bootstrap samples are drawn using the same sampling scheme adopted to produce the original sample, as suggested by [5] and [4]. Under the conditions ensuring DBAU&C, the estimated map converges to the true map, so that the bootstrap distribution should converge to the true distribution, also providing consistent estimators of its mean squared errors.

3 Simulation study

The performance of the proposed data-driven mapping strategy was empirically assessed by means of a simulation study performed on a real survey region located in Calabria (Southern Italy). The values of several auxiliary variables and the value of growing stock volume (interest attribute) were available for each pixel partitioning the study region.

In order to check the improvement as the number of pixels partitioning the study area increases and their size become smaller, the study area was partitioned considering three different pixels sizes and the values of interest and auxiliary attributes were aggregated within those pixels.

From each of the three partitions, $10,000$ samples were independently selected by means of one-per-stratum sampling (OPSS) with a constant sampling fraction of 4%. For each sample, selection of auxiliary variables was performed, the growing stock volume estimate for each pixel was derived by using the IDW interpolator, where the value of the smoothing parameter was obtained by means of leave-one-out cross-validation and, finally, harmonization was implemented. Furthermore, for each simulation run, $1,000$ bootstrap samples were selected from the estimated map by the same OPSS scheme adopted for extracting the original sample, then bootstrapped maps were used to achieve the bootstrap root mean squared error estimates. From the resulting Monte Carlo distributions, several perfomance indexes were computed. Results show that usually 4 or 5 covariates are selected out of the 11 originally available, suggesting that the selecting rule is likely to choose parsimonious models. Also for the choice of the smoothing parameter, small values are the most commonly selected, probably owing to the smoothness of the error surfaces to be interpolated (see Fig. 1). Furthermore, relative bias values and relative root mean squared errors quickly decrease in minima, means and maxima as the spatial grain decreases, with a balance between underestimation and overestimation, showing also a relevant spatial autocorrelation of negative and positive values (see Tab. 1).

As to the bootstrap root mean squared error estimator, underestimation seems to be prevalent. Nevertheless, its tendency to be conservative is more apparent as the spatial grain decreases. Indeed, the number of pixels in which the ratio between the expectation of the bootstrap root mean squared error estimator and the true root

Table 1 Minima, means and maxima of the absolute bias (AB), root mean squared errors (RMSE) and its ratio with the expectation of bootstrap root mean squared error (RAT) achieved for the three populations considered in the simulation study.

	AB			RMSE			RAT		
POP	MIN	MEAN	MAX	MIN	MEAN	MAX	MIN	MEAN	MAX
POP1	0.381	27.400	86.010	28.86	58.59	133.80	0.348	0.652	0.919
POP2	0.038	22.770	81.215	12.93	35.99	92.29	0.221	0.592	1.100
POP3	0.003	14.243	65.035	6.00	23.43	73.88	0.263	0.730	1.809

mean squared error is greater than 1 increases as the spatial grain decreases. It is probable that with a thinner partition conservativeness over the whole area can be achieved.

4 Case study

The proposed mapping strategy was applied to provide the map of densities of wood volume in the forest estate of Rincine (Central Italy), an area partitioned into square cells of $23 \times 23m^2$. For each cell, the values of several auxiliary variables were available. For sampling purposes, the area was partitioned into 50 blocks of cells and, following OPSS, one cell was randomly selected within each block, determining a sample of 50 units. The wood volume of each tree in the selected cells was reckoned to achieve the wood volume density by three size classes: small, medium and large trees.

For each class, variable selection was performed using the procedure adopted in the simulation. Mapping was performed for each size class harmonizing the total of the interpolated values with the regression estimates of the total over the whole survey area. Subsequently, the three estimated maps were taken as pseudo-populations

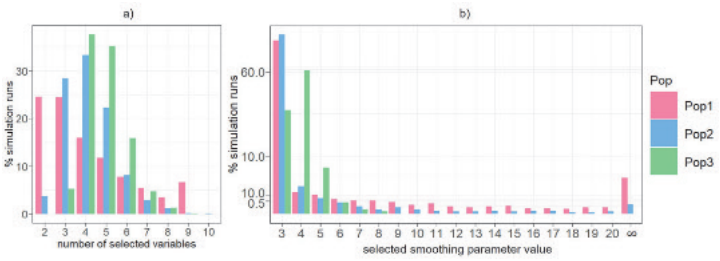


Fig. 1 For each of the three considered populations a) frequencies (expressed as percentage of the number of simulation runs) of the number of variables adopted in the predicting model; b) frequencies (expressed as percentage of the number of simulation runs) of the selected smoothing parameter.

from which 10,000 bootstrap samples were selected by the OPSS scheme adopted to select the original sample. From each bootstrap sample, the three bootstrap maps were achieved for each class repeating the usual steps and the total map was achieved by sum. Then, bootstrap root mean squared errors were computed for each of the four maps. Bootstrap samples were also adopted to achieve the bootstrap root mean squared error of the regression estimator, incorporating uncertainty determined from the variable selection.

The maps of estimated densities within cells show a relevant level of uncertainty for small trees that decreases for medium trees and becomes satisfactory for large trees and for the total. High uncertainty for small trees is mainly due to the high variability in the number of small trees (many of which from scattered forest natural regeneration) within the selected cells, with many cells with 0 small trees and few with more than 150 trees. Consequently, as expected (see e.g. [6]), the precision is deteriorated by the highly clustered spatial pattern.

In order to highlight the improvement when exploiting auxiliary sources of information, those maps were compared to those obtained using the sole spatial information and selecting the value of the smoothing parameter from data as in [5]. From this comparison it is evident that the proposed mapping strategy reduces the excessive smoothing produced by IDW interpolation based on the sole spatial information.

5 Final remarks

Despite being data-driven, the proposed mapping strategy is based on several subjective choices, such as the choice of the linear model for performing regression and the choice of the criterion for implementing the leave-one-out cross-validation. However, it should be also noticed that these choices only impact on the sample statistics adopted for mapping, while the precision of the map is objectively determined by the sampling scheme used for selecting the units partitioning the study area.

A further issue is that two-phase sampling schemes are usually adopted in forest inventories, whereas one-phase OPSS was here proposed. However, [4] proved the consistency of the IDW interpolator if the two-phase scheme continues to provide spatially balanced samples. As consequence, the procedure can be applied also in two-phase large-scale forest inventories.

The proposed design-based, data-driven strategy for mapping forest resources exploiting auxiliary variables, guides the user from the preliminary choice of the assisting model to the final map and the estimation of its precision, through model selection, exploitation of selected covariates for mapping, choice of the smoothing parameter for IDW interpolation, harmonization with traditional estimates of totals and bootstrap resampling from the estimated map. Moreover, this strategy seems to be suitable also for mapping environmental attributes besides forestry.

References

1. Breidt, F.J., Opsomer, J.D.: Model-assisted survey estimation with modern prediction techniques. Stat. Sci. (2017) doi: 10.1214/16-STS589
2. Burman, P., Nolan, D.: A general Akaike-type criterion for model selection in robust regression. Biometrika (1995) doi: 10.2307/2337352
3. Fattorini, L., Marcheselli, M., Pratelli, L.: Design-based maps for finite populations of spatial units. J. Am. Stat. Assoc. (2018) doi: 10.1080/01621459.2016.1278174
4. Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C., Pratelli, L.: Two-phase sampling strategies for design-based mapping of continuous spatial populations in environmental surveys. Ann. Appl. Stat. (2021) doi: 10.1214/20-AOAS1392
5. Fattorini, L., Franceschi, S., Marcheselli, M., Pisani, C., Pratelli, L.: Design-based spatial interpolation with data driven selection of the smoothing parameter. Submitted (2022)
6. Gregoire, T. G., Valentine, H. T.: Sampling Strategies for Natural Resources and the Environment. Chapman & Hall, New York (2008)
7. Marcelli, A., Fattorini, L., Franceschi, S.: Harmonization of design-based mapping for spatial populations. Stoch. Environ. Res. Risk Assess. (2022) doi: 10.1007/s00477-022-02186-2
8. Särndal, C. E., Swensson, B., and Wretman, J.: Model Assisted Survey Sampling. Springer, Berlin (1992)
9. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. (1970) doi: 10.2307/143141

Environmental data and Climate change

Ensemble model output statistics for temperature forecasts in Veneto

Ensemble model output statistics per la previsione delle temperature in Veneto

Gaetan Carlo, Giummolè Federica, Mameli Valentina, Siad Si Mokrane

Abstract Post-processing methods are nowadays widely used for limiting the impact of errors in ensemble forecast of meteorological variables. Ensemble model output statistics are an easy-to-apply technique for post-processing, based on a linear regression model. In this paper we use an ensemble model output statistic for the forecast of daily maximum temperatures in Veneto. We calculate estimative and calibrated predictive distributions for a time period of three years. We then compare the different predictive distributions by means of the log-score, the continuous ranked probability score and the coverage of the corresponding predictive quantiles. We show that the calibrated approach improves on the estimative ones as regards both mean scores and coverage probabilities.

Abstract *Al giorno d'oggi, metodi di post-processing vengono ampiamente utilizzati per limitare l'impatto di errori nelle previsioni di variabili meteorologiche che utilizzano ensemble. Gli ensemble model output statistics sono una tecnica di post-processing basata sul modello di regressione lineare. Nel nostro lavoro li abbiamo utilizzati per la previsione delle temperature massime giornaliere in Veneto. Abbiamo calcolato diverse distribuzioni predittive estimative e calibrate per un periodo di tre anni e le abbiamo confrontate fra loro utilizzando log-score, continuous ranked probability score e probabilità di copertura dei rispettivi quantili. I risultati mostrano la superiorità delle distribuzioni predittive calibrate rispetto alle corrispondenti distribuzioni estimative, secondo tutti i criteri considerati.*

Key words: Calibration, Coverage probability, Ensemble model output statistics, Post-processing, Predictive distributions, Scoring rules, Weather forecast.

Gaetan Carlo, Giummolè Federica, Siad Si Mokrane
Ca' Foscari University of Venice, Via Torino 155 Mestre - Venezia, e-mail: gaetan@unive.it, giummole@unive.it, simokrane.siad@unive.it

Valentina Mameli
University of Udine, Via Tomadini 30 Udine, e-mail: valentina.mameli@uniud.it

1 Introduction

Weather predictions are a critical component of decision support for a wide variety of applications. As a result, there is a significant demand for accurate weather predictions from a diverse variety of stakeholders, including the general public, the corporate sector, and government agencies that issue weather warnings. Forecasts have gradually improved over the previous decades, partly due to breakthroughs in numerical weather prediction (NWP) ([1]; [8]). The advancement of powerful high-performance computers enables more detailed weather simulations. Despite these advances in NWP, forecasts generated by physics-based models, usually provided as forecast ensembles, exhibit systematic bias and are often underdispersive ([2], [7]). Statistical post-processing methods are nowadays widely used for further refining, improving, and calibrating NWPs. The ensemble model output statistics (EMOS, [5]) is among one of the most popular post-processing techniques used to calibrate the ensemble forecasts. EMOS is based on a simple regression model with parameters depending on the ensemble forecasts. It is able to correct for systematic biases and dispersion errors.

This article proposes an adjustment of EMOS based on a bootstrap calibration procedure proposed by [3] and further extended to the EMOS context in [4]. In order to evaluate the performance of the calibrated EMOS and develop a comparison with the classic EMOS, we consider a real case study dealing with maximum daily temperatures at measurement sites located in the Veneto region, northern Italy. The evaluation and comparison of methods are based on measures of goodness for calibration and sharpness, the most desirable properties that characterise predictive models [6]. In particular, we compare different predictive models by means of the log-score, the continuous ranked probability score (CRPS) and the coverage of the corresponding predictive quantiles. These analyses exhibit the accuracy of the calibrated EMOS, with respect to the classic EMOS, in the presented application, and highlight the great potentiality of this new technique to provide calibrated and sharp predictive models.

The paper is organised as follows. In Section 2 we outline the methodology employed in this research. In Section 3 we introduce the data and we assess the performance of the calibrated EMOS in comparison with the classic EMOS. Finally, in Section 4 we present some concluding remarks.

2 The method

The most popular post-processing techniques used to calibrate the ensemble forecasts are EMOS that allow for probabilistic forecasts of weather variables ([5]) in the form of Gaussian predictive distributions. More formally, it is assumed that a weather variable Z depends on the ensemble forecasts $X_1, \ldots, X_m$ in such a way that its mean is equal to $\beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$ and its variance is equal to $\gamma + \delta S^2$, where $S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2$ denotes the ensemble variance and $\beta_1, \ldots, \beta_m$, γ and

δ are non-negative unknown coefficients. Therefore, under normality assumptions, the distribution of Z is $N(\beta_0+\beta_1X_1+\ldots+\beta_mX_m, \gamma+\delta S^2)$.

For fitting the unknown EMOS coefficients, $\beta_0, \beta_1, \ldots, \beta_m, \gamma, \delta$, [5] proposed to use the method of minimum CRPS estimation, which finds the estimates by minimising the CRPS. The CRPS estimates are then substituted to the unknown parameters in the EMOS normal distribution obtaining what is known as an estimative distribution for the future weather quantity Z. In the introduction we have referred to this procedure as the classic EMOS. Here we also consider for comparison the EMOS estimative distribution obtained with the maximum likelihood estimates (MLE), that minimises the log-score. Unfortunately, estimative distributions may perform poorly, especially when the sample size is small with respect to the number of members of the ensemble. In particular, the important requirement of calibration is hardly attained by estimative distributions. As explained in [6], calibration is a sort of consistency between a predictive distribution and future observations and it can be graphically assessed by means of the histogram of the probability integral transformed (PIT) values. Moreover, a well calibrated predictive distribution provides prediction limits with coverage probability close to the nominal value.

In order to obtain calibrated predictive distributions, we consider the approach presented by [3] and recently adapted to the EMOS context by [4] that propose a simple simulation experiment. Suppose that we want to make prediction on the unobservable variable Z with distribution $G(z;\theta)$ depending on an unknown parameter θ. In our case $Z \sim N(\beta_0+\beta_1X_1+\ldots+\beta_mX_m, \gamma+\delta S^2)$, so that $\theta=(\beta_0,\beta_1,\ldots,\beta_m,\gamma,\delta)$. Let $\hat{\theta}$ be the MLE for θ or an asymptotically equivalent alternative, such as the CRPS estimator, based on an observed sample.

A suitable adjustment of the classic EMOS that yields calibrated predictive distributions can be obtained using the following simple bootstrap procedure:

$$G^{boot}(z;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} G\{z_\alpha(\hat{\theta}^b);\hat{\theta}\}|_{\alpha=G(z;\hat{\theta})}, \tag{1}$$

where $z_\alpha(\theta)$ denotes the α-quantile of Z and $\hat{\theta}^b$, $b=1,\ldots,B$, are the estimates obtained by a parametric bootstrap from the estimative distribution of Z. The goodness of the approximation depends on the efficiency of the bootstrap simulation procedure. We also note that in the case under study it is straightforward to draw random samples and calculate quantiles and probabilities from normal estimative distributions. This makes the bootstrap calibration procedure very easy to apply for improving on EMOS estimative distributions.

3 A case study

For this study, we focus on maximum daily temperature forecasts at surface stations in the Veneto region, northern Italy. Our period of interest spans 3 years from 16 August 2009 to 17 August 2012. Historical maximum daily tempera-

ture forecasts were acquired from the website of the Italian national system for the collection, processing and dissemination of climate data, created by ISPRA (http://www.scia.isprambiente.it/). Ensemble predictions were available from World Climate Research Programme. We use the Coupled Model Intercomparison Project Phase 6 system (CMIP6) that consists of over 100 models of which 29 ensemble members cover the Veneto region. The data in the CMIP6 archive are downscaled to station level using elevation as a base for interpolation. In order to assess and compare the performance of different EMOS predictive distributions, we analyse maximum daily temperatures for the station Cavallino-Treporti (Longitude: 12.48642°, Latitude: 45.45805°), located on the Venetian lagoon. After removing missing observations from the selected station, the sample contains 1091 observations of daily temperatures and 26 ensemble members. According to [5], we use a sliding window of 40 observations as training set, with the remaining 1051 days available as test set. In this temporal window the generating process is supposed to be stationary. First, the EMOS parameters are estimated by optimising both the log-score and the CRPS over the sliding training period. Then, for each of the 1051 days available as test set, the performance of the two estimative distributions obtained by estimating with the log-score and the CRPS, and the corresponding calibrated distributions, are evaluated by means of different measures of goodness for calibration and sharpness. Table 1 summarises the mean and standard deviations of the log-score and the CRPS for the four predictive models. The superior performance of the calibrated models is reflected in the smaller values of the two scores.

Table 1 Log-score and CRPS values of the four predictive distributions. Standard errors in brackets. Est log denotes the estimative EMOS with log-score estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

	Est log	Cal log	Est CRPS	Cal CRPS
Log-score	3.22	2.51	3.65	2.58
	(0.09)	(0.03)	(0.14)	(0.06)
CRPS	1.79	1.64	1.82	1.66
	(0.05)	(0.04)	(0.05)	(0.04)

For properly measuring calibration and sharpness (concentration) of the different predictive models, we have also obtained coverage probabilities and mean lengths of central intervals of level 66.7% (Table 2). Moreover, the coverage probability of upper prediction limits of levels 90%, 95% and 99% for the different predictive models have been obtained, being fundamental quantities for the evaluation of environmental risk (Table 3). The effect of calibration can be clearly seen from coverage probabilities for calibrated predictive models, much closer to target values than those for the estimative ones. Of course, shorter central prediction intervals for the estimative models are due to the loss of corresponding coverage.

Finally, Figure 1 shows the PIT histograms for the four considered predictive models. The effect of calibration can be observed in the flat histogram very close to the uniform one for the calibrated predictive models. Instead the U-shaped

histograms of the estimative models are the consequence of the excessive underdispersion.

Table 2 Coverage probabilities and mean lengths of the central prediction interval of level 66.7% for the estimative predictive models with MLE and CRPS estimates, and corresponding calibrated models. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

	Est log	Cal log	Est CRPS	Cal CRPS
α=0.667	0.402	0.646	0.375	0.648
	(0.015)	(0.015)	(0.015)	(0.015)
Mean length	3.197	5.419	2.905	5.334
	(0.023)	(0.040)	(0.024)	(0.044)

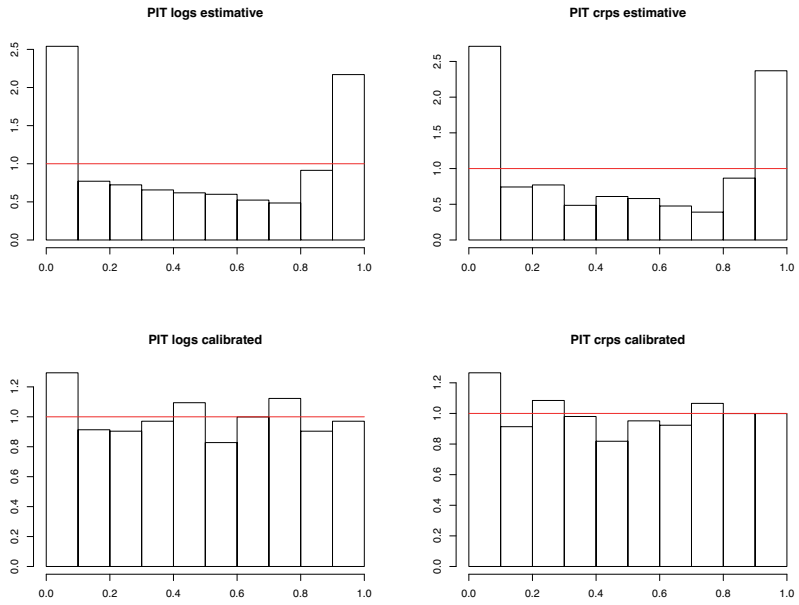


Fig. 1 PIT histograms of the estimative predictive models based on the MLE and the CRPS estimates, and the corresponding calibrated models.

Table 3 Coverage probabilities of upper prediction limits for the estimative predictive models with MLE and CRPS estimates, and the corresponding calibrated models. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

α	Est log	Cal log	Est CRPS	Cal CRPS
0.90	0.783	0.903	0.763	0.900
	(0.013)	(0.009)	(0.013)	(0.009)
0.95	0.835	0.952	0.814	0.952
	(0.011)	(0.007)	(0.012)	(0.007)
0.99	0.912	0.991	0.882	0.990
	(0.009)	(0.003)	(0.010)	(0.003)

4 Conclusions

This work proposes a comparison between the classic EMOS and a calibrated EMOS based on a bootstrap procedure, applied to a real case study of temperatures in the Veneto region, Italy. In particular, we have used measurements from a single station, Cavallino-Treporti, in the Venice lagoon. Future work will consider a spatial analysis of temperature data collected all over the Veneto region, including the coast around the Venice lagoon, the low Venetian plan and the Dolomite mountains. Indeed, the Veneto region's height fluctuates from sea level (and also below sea level) up to about 3,300 m, with a corresponding wide variation in the temperatures.

References

1. Bauer, P., Thorpe, A., Brunet, G.: The quiet revolution of numerical weather prediction. Nature, **525**, 47–55 (2015).
2. Buizza, R.: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Mon. Wea. Rev., **125**, 99–119 (1997).
3. Fonseca, G., Giummolè, F., Vidoni, P.: Calibrating predictive distributions. Journal of Statistical Computation and Simulation, **84**, 373–383 (2014).
4. Giummolè, F., Mameli, V.: Comparing predictive distributions in EMOS. Proceedings of the Italian Statistical Society, 1–6 (2020).
5. Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman , T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CPRS Estimation. Monthly Weather Review, **133(5)**, 1098–1118 (2005).
6. Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 243–268 (2007).
7. Haiden, T., Janousek, M., Vitart, F., Ferranti, L., Prates, F.: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. **588**, 56 pp. (2019).
8. Raftery, A., Gneiting, T., Balabdaoui, F., Polakowski, M.: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. Quart. J. Roy. Meteor. Soc., **145**, 12–24 (2019).

State of the urban Environment in Italy. A comparative analysis of selected composite indicators

Lo Stato dell'ambiente urbano in Italia. Analisi a confronto per la selezione degli indicatori sintetici

Lecardane Giuseppe, Istat, lecardan@istat.it

Abstract The consequences of climate change, depletion of water resources, urban pollution and other environmental crisis involve all the contexts in which human activity is expressed. In this paper, main statistical methods are compared for the synthesis of representative indicators of the set of environmental phenomena. A multidimensional study that offers insights to subjects involved in development policies and environmental sustainability in the urban area.

Abstract *Le conseguenze dei cambiamenti climatici, il depauperamento delle risorse idriche, l'inquinamento urbano e le altre manifestazioni di crisi ambientale coinvolgono in modo trasversale tutti i contesti in cui si esprime l'attività umana. Nel presente lavoro sono messi a confronto alcuni principali metodi statistici per la sintesi di indicatori rappresentativi dell'insieme dei fenomeni ambientali per il monitoraggio delle città italiane. Uno studio multidimensionale che offre spunti di approfondimento a soggetti impegnati su politiche di sviluppo e sostenibilità ambientale nel territorio urbano.*

Key words: environment, synthetic index, urban area.

1. Introduction

Environment quality and social well-being are closely interconnected on a collective and individual level. It is, in fact, a relationship that invests values of primary importance, such as those relating to human health and safety, heritage and resources to be passed on to future generations. Therefore, statistical information, for an in-depth knowledge of environmental issues, it is relevant for everyone. The

[1] Lecardane Giuseppe, Istat; lecardan@istat.it

effectiveness of any action in the environmental field implies an awareness of citizens and the choice of appropriate behaviors. Complex and multidimensional nature of environmental phenomena requires the identification and measurement of indicators for the implementation of more effective and incisive information programs. Due to the number of indicators that represent the different dimensions of the state of the urban environment to be measured, it is also necessary to find a synthesis process to improve comparison and analysis of the observed phenomena. In fact, the synthesis has the advantage of performing simpler and faster analyzes especially in comparative terms and in addition of summarize heterogeneous and multidimensional phenomena.

In this paper, we will proceed to experimental comparison of some main weighting approaches for the composite indicator-construction methods referring to the data on urban environmental quality on issues such as water, air, energy, noise, waste, mobility and urban green for 110 provincial capitals (Istat, 2020).

The aim of this work is not to establish which approach is preferable to another but to analyze the robustness and sensitivity of the results from the different composite methods used. Analysis of the state of the urban environment therefore provides useful measuring tools with the appeal to an awareness of the need for a change of course towards planning a more urban sustainability.

The paper is structured as follows: description and application of the main composite methods used; comparison of the results obtained through cograduation matrices of the rankings, correlation matrices and dispersion matrices of the values obtained with the different methods; summary conclusions.

2. Methodology

The multidimensional nature of the phenomenon and its measurability through a system of elementary indicators allows the targeted construction of a composite technique which is able of acquiring the multiple aspects. Therefore, there is a need to experiment composite methods of elementary data to improve the measurement and communicability of the results.

To identify measurement model the *formative model* was followed, according to which individual indicators are causes of an underlying latent variable, rather than its effects, they are not interchangeable, and their correlations are not explained by the model. Regarding the latest data on the urban environmental (Istat 2020), a set of elementary indicators on environmental themes (water, air, energy, noise, waste, mobility, and urban green) were selected (Tab. 1). These indicators have a high variability and little correlation with each other, characteristics suitable to achieve the aims. It's the basis for the aggregation process through the construction and comparability of some main composite methods. Elementary indicators have been normalized and standardized to obtain data purified from units of measurement and comparison process.

Standardized deviation in the composite index allows the construction of a robust measure and not very sensitive to remove a single elementary index (Mazziotta et al., 2013).

Table 1: Environmental indicators selected, survey Istat "Environmental data of cities". 2020

Environmental Issues	Environmental indicators and polarity (+/-)
Water	a1(-). Household water bill per capita *(liters per day)* a2(-). Household water bill total per capita *(liters per day)* a3(-). Water network losses *(%)*
Air	b4(+). Fixed air quality monitoring stations *(per 100,000 inhabitants)* b5(-). Composite indicator of atmospheric pollution *(average values exceeding threshold limit concentration of PM10, PM2,5, NO2 and O3)*
Energy	c6(+). Extension of the thermal solar panels installed on the municipal buildings *(m^2 per 1,000 inhabitants)* c7(+). Total power of photovoltaic solar panels owned by the municipal administration (kW per 1,000 inhabitants) c8(+). Charging columns for electric cars by type *(per 10 km^2)*
Mobility	d9(-).Motorization rates for cars by municipality *(vehicles in circulation per 1,000 inhabitants)* d10(+).Electric vehicle circulating by municipality *(per 1.000 vehicle circulating)* d11(-).Pollution potential index of vehicle circulating by municipality *(high/medium pollution potential vehicle per 100 medium/low pollution potential vehicle)*
Municipal waste	e12(+).Door-to-door municipal waste collection for households *(%)* e13(-).Municipal road waste collection for households *(%)*
Noise	f14(+).Complaint presented by citizens on noise pollution by municipality *(per 100,000 inhabitants)*
Urban green	g15(+). Tree cadastre by municipality *(Tree per 100 inhabitants)* g16(+).Density of urban green in the municipalities *(% on the municipal area)* g17(+). Urban green in the municipalities *(m^2 per inhabitants)*

Source: Istat

In addition, *polarity* (positive or negative) of the relationship between indicator and phenomenon was specified. Finally, standardized indicators were aggregated. Following, steps to calculate composite index by comparing the following methods.

Given the matrix $X=\{x_{ij}\}$ with n rows (units) and m columns (indicators), composite methods have the following mathematical properties:

Mean Z-scores *(MZ)*

$Mz_i = \frac{\Sigma_{j=1}^{m} z_i j}{m}$ $Z=\{z_{ij}\}$ transformed matrix for unit *i* and indicator *j*

con $z_{ij} = \pm \frac{(x_{ij} - M_{xj})}{s_{x_j}}$ if the indicator *j* has positive or negative polarity.

M_{xj} e S_{xj} arithmetic mean and deviation standard of indicator *j*.

The *MZ* allows transformation of indicators j into standardized deviations and aggregation with the arithmetic mean.

Mean R-Indices *(MR)*

$Mz_i = \frac{\Sigma_{j=1}^{m} r_{ij}}{m}$ $R=\{r_{ij}\}$ transformed matrix for unit *i* and indicator *j*

con $r_{ij} = \begin{cases} \frac{(x_{ij}-Min_{xj})}{(Max_{xj}-Min_{xj})} \\ \frac{(Max_{xj}-x_{ij})}{(Max_{xj}-Min_{xj})} \end{cases}$ Min_{xj} e Max_{xj} indicator *j*

The *MR* allows standardization with min-max method of the indicators j and aggregation with the arithmetic mean.

Adjusted MPI *(AMPI)*

$MPI_{ci}^{\pm} = M_{ri} \pm S_{r_i} cv_i$

con $r_{ij} = \begin{cases} \frac{(x_{ij}-Mi\ _{xj})}{(Max_{xj}-Min_{xj})} 60 + 70 \text{ if the indicator } j \text{ has positive polarity} \\ \frac{(Max_{xj}-x_{ij})}{(Max_{xj}-Min_{xj})} 60 + 70 \text{ if the indicator } j \text{ has negative polarity} \end{cases}$

$M_{r_i} = \frac{\Sigma_{j=1}^{m} r_{ij}}{m}$ $S_{r_i} = \sqrt{\frac{\Sigma_{j=1}^{m} (r_{ij}-M_{ri})^2}{m}}$ $cv_i = \frac{S_{ri}}{M_{r_i}}$

The *AMPI* is a non-compensatory (or partially compensatory) composite index and allows min-max standardization of the indicators j and aggregation with the arithmetic mean penalized by the "horizontal" variability of the indicators themselves. Normalized values are approximately in the range (70; 130), where 100 is the reference value[1].

3. Results

From the exploratory data analysis, indicators show a pronounced variability (many CV values are close to 1) and little correlated with each other (tabb. 2 and 3), characteristics suitable to achieve the aims. To identify a composite index that represents multidimensionality of the urban environment, some main methods were compared using the transformation of the indicators to obtain data purified from units of measurement and their variability. Figure 1 shows cartograms of the four approaches used in 110 provincial capitals. Result of the analysis is almost uniform for all methods, with the subdivision of decreasing territorial trialism Northern, Center and Southern Italy.

Table 2: Average and variability measures of environmental indicators. Provincial capitals. 2020

	a1	a2	a3	b4	b5	c6	c7	c8	d9	d10	d11	e12	e13	f14	g15	g16	g17
Arithmetic mean	206,9	150,8	37,3	2,7	19,4	4,4	259,3	2,4	668,1	1,5	129,0	72,0	47,0	12,2	14,8	14,0	42,7
Standard deviation	37,5	27,8	15,2	2,0	19,8	18,4	261,7	5,7	69,9	0,8	18,8	34,4	40,4	20,5	14,3	14,7	61,0
Coefficient of variation	0,2	0,2	0,4	0,7	1,0	4,2	1,0	2,4	0,1	0,5	0,1	0,5	0,9	1,7	1,0	1,1	1,4

Source:Istat data processed

[1] In the Bienaymé-Cebycev theorem, terms of the distribution within the interval (70; 130) constitute at least 89 percent of the total terms of the distribution.

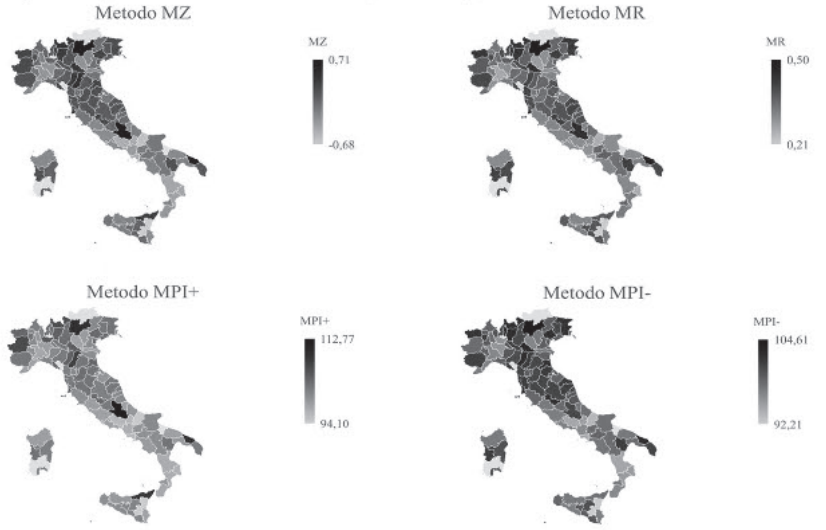
Table 3: Correlation matrix of environmental indicators. Provincial capitals. 2020

	a1	a2	a3	b4	b5	c6	c7	c8	d9	d10	d11	e12	e13	f14	g15	g16	g17
a1	1	0,78	-0,23	-0,05	0,38	-0,01	-0,27	0,34	-0,19	0,33	-0,12	0,21	-0,10	0,05	-0,01	0,09	0,05
	a2	1,00	-0,16	-0,10	0,33	-0,06	-0,32	0,39	-0,16	0,23	-0,02	0,24	-0,18	0,08	-0,05	0,06	0,00
		a3	1,00	0,09	-0,42	0,07	-0,02	-0,20	0,24	-0,30	0,31	-0,20	-0,01	0,07	-0,35	0,02	-0,03
			b4	1,00	-0,09	0,09	0,15	-0,16	0,14	-0,05	0,00	0,26	-0,09	-0,12	-0,05	-0,10	0,15
				b5	1,00	-0,12	-0,14	0,35	-0,33	0,47	-0,51	0,13	0,10	0,04	0,38	0,07	0,01
					c6	1,00	0,05	-0,05	0,20	-0,03	0,03	0,03	-0,02	-0,04	-0,09	0,19	-0,01
						c7	1,00	-0,24	0,24	-0,20	0,05	0,02	-0,06	-0,18	-0,02	-0,17	-0,05
							c8	1,00	-0,33	0,30	-0,15	0,03	-0,12	0,08	0,10	-0,03	-0,07
								d9	1,00	-0,27	0,27	0,20	-0,31	-0,10	-0,10	-0,24	0,05
									d10	1,00	-0,55	0,13	0,02	0,09	0,28	0,08	0,30
										d11	1,00	-0,03	-0,21	-0,07	-0,37	-0,02	-0,12
											e12	1,00	-0,42	-0,10	0,01	-0,05	0,14
												e13	1,00	0,19	0,18	0,08	0,04
													f14	1,00	0,07	0,32	0,02
														g15	1,00	-0,02	0,08
															g16	1,00	-0,11
																g17	1,00

From the ranking of the four composite indicators, it is possible to observe the positioning of the Italian municipalities based on the state of environment which decreases towards the higher ranks. Trento has the best urban environmental performance while Catania is the city with the highest negative impact.

In ranking of the top five cities with low environmental impact, Bolzano, Sondrio, Mantova and Bergamo are also distinguished, cities of small and medium population size in Northern Italy. At the bottom of the ranking with a greater environmental pressure Isernia, Napoli, Frosinone and Campobasso in Southern Italy (Tab. 4). Table 5 shows rank differences compared by means of the absolute difference and Spearman's rank correlation coefficient. Sensitivity analysis shows similar results in the comparison between MR-MPIc$_+$ method and the MR-MPIc$_-$ method with absolute average rank differences 5.19 and 6.36 positions respectively with a strength of the relationship directly proportional and close to 1 (0.98 and 0.96). Linear relationship with R-values is very high too (fig. 2).

Figure 1: Map of composite indicators. Provincial capital. 2020 *(*)*



(*) Chromatic provincial areas refer to their provincial capitals.

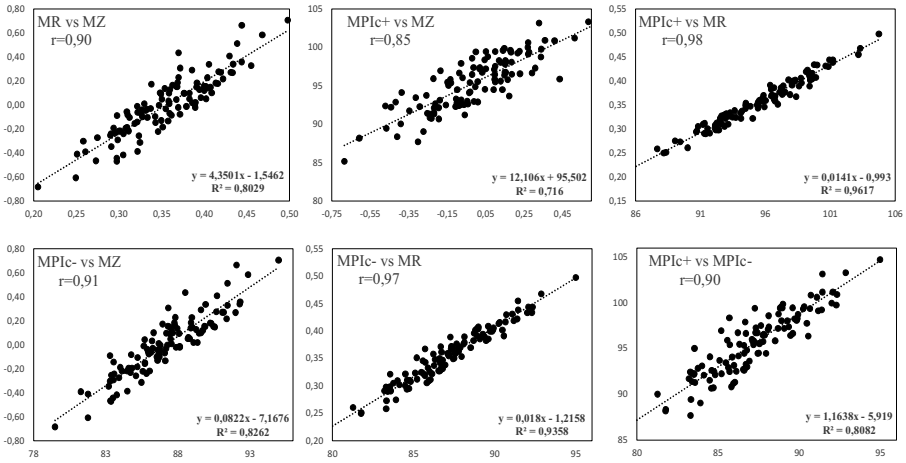
Table 4 - Ranking of the five best and worst environmental performances - Provincial capital. 2020

Provincial capital	MZ		MR		MPIc+		MPIc-	
	N.	Rank	N.	Rank	N.	Rank	N.	Rank
Better environmental performance								
Trento	0,71	1	0,50	1	104,76	1	95,00	1
Bolzano-Bozen	0,51	4	0,44	6	101,20	5	91,45	7
Sondrio	0,41	6	0,43	10	100,85	7	90,72	12
Mantova	0,36	7	0,44	5	100,93	6	92,33	3
Bergamo	0,34	8	0,43	7	99,77	11	92,29	4
Worse environmental performance								
Isernia	0,44	106	0,30	98	92,19	92	83,51	100
Napoli	0,47	107	0,27	105	89,44	105	83,33	103
Frosinone	0,47	108	0,30	97	92,35	89	83,38	102
Campobasso	0,61	109	0,25	109	88,17	108	81,77	108
Catania	0,68	110	0,21	110	85,14	110	79,49	110

Table 5 - Sum of ranking differences between composite methods used.

Measures	Ranking differences					
	MZ-MR	MZ-MPIc+	MZ-MPIc-	MR-MPIc+	MR-MPIc-	MPIc+-MPIc-
Absolute average rank diff.	11,60	14,35	10,44	5,19	6,36	11,29
Cograduation index ρ	0,89	0,83	0,91	0,98	0,96	0,89

Figure 2 – Linear relationship compared between composite methods used



4. Conclusions

Study on the multidimensional aspects of the urban environmental through comparison of some composite methods offer an important contribution to the interpretation of the phenomenon. Geography of the environmental state and urban anthropic pressure highlights an unbalanced and negative configuration for most of

the southern cities. At the other end of distribution, higher environmental performances are recorded, especially in the northern small and medium-sized urban areas where investments in environmental projects are constantly growing.

References

1. Istat: Dati ambientali nelle città. Qualità dell'ambiente urbano. Statistiche Report. I.Stat. Roma (2020).
2. Lecardane G.: Lo Stato dell'Ambiente. Misure statistiche e Indicatori di qualità ambientale. E-book Istat. Roma. (2013).
3. Lecardane G., Arcarese, G.: L'ambiente Urbano. I dati sulla qualità ambientale nelle nostre città. Aracne editrice srl. Roma (2009).
4. Mazziotta, M., Pareto, A. A Non-compensatory Composite Index for Measuring. Well-being over Time. Cogito. Multidisciplinary Research Journal, Vol. V, 4 (2013).
OECD. Handbook on Constructing Composite Indicators. OECD Publications, Paris (2008)

A Functional Data Analysis approach for Climate Model Selection : the case study of Campania Region

Selezione di un modello climatico attraverso l'analisi dei dati funzionali: il caso di studio della Regione Campania

Villani V., Romano E., Mercogliano P.

Abstract In recent years climate projections are available in large number of applicative fields, so the difficulty lies in using them for climate change impact studies. The main aim of this work is to propose a Functional Data Analysis clustering strategy to select a single regional climate model. We will refer to climate projections available within the EURO-CORDEX program and select the one that has the best ability to simulate the observed climate in the studied area.

Abstract Dato il rapido aumento del numero di proiezioni climatiche disponibili negli ultimi anni e la difficoltà di utilizzarle tutte per gli studi sull'impatto del cambiamento climatico, l'obiettivo di questo lavoro è selezionare un unico modello climatico regionale, tra quelli disponibili nell'ambito del programma EURO-CORDEX, che abbia la migliore capacità di simulare il clima osservato nell'area di studio. A tal fine, è stata proposta una strategia di clustering per dati funzionali**.**

Key words: Climate changes, Impact studies, High-resolution climate projections, Functional Data Analysis

[1] Villani V., Regional Models and geo-Hydrological Impacts (REMHI) Division, Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Caserta, 8100, Italia; veronica.villani@cmcc.it:

Romano E., Dipartimento di Matematica e Fisica, Universitá degli Studi della Campania Luigi Vanvitelli, Caserta, Italy; elvira.romano@unicampania.it:

Mercogliano P., Regional Models and geo-Hydrological Impacts (REMHI) Division, Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Caserta, 8100, Italia; paola.mercogliano@cmcc.it:

1 Introduction

The rapid increase in the number of available climate projections in recent years, especially by adopting high resolution climate models, and the difficulty (in terms of computation costs) in using all of them for the local assessment of climate change impacts studies, make crucial to select a subset of climate models representative of the area of interest. The use of the mean value (called the ensemble mean) of this subset for subsequent climate change impact studies is in direct conflict with the aim of evaluating extreme scenarios to be used in the context of such studies. In fact, in the ensemble mean of climate models, information on climate extremes could be lost. The aim of this work is to select a regional climate model, among those available under the EURO-CORDEX program (Hennemuth et al, 2017; Jacob et al, 2014; Jacob et al, 2020), which has the best skill in simulating the observed climate over the study area. To do this, we used Functional Data Analysis (FDA). Particularly, the FDA allowed us to study the commonalities and dissimilarities between the results of the various simulations and the trend of the station data. Such dissimilarities have been evaluated by calculating the functional distance between the coefficients of the first derivatives of the 18 functional simulations and of each of these with the coefficients of the first derivative of the observed functional data. Since, considering the coefficients of the first derivatives of the functional data, the distance is not influenced by the bias of the simulations considered. Hence, in the case where the calculated distance is small, the functional data show a similar form of temporal changes in the series. Subsequently, a cluster analysis was performed on the dissimilarity matrix obtained as mentioned above, and a characterization of the variability within the classes was performed to identify the simulation showing a more similar form of temporal changes in the series with respect to the time series of observations. The annual cumulative precipitation variable was used, both calculated with the daily series of station data available over the Campania region and with the data of 18 EURO-CORDEX models. As works of literature in which the FDA has been applied to climatological data we find: Holtanova et al. (2019) which used FDA to distinguish time series data related to climatology; Burdejová and Härdle (2019) have suggested a dynamic functional factor for the FDA of climatological data, including precipitation and temperature; Ghumman et al. (2020) which aims to answer the research question that what are the commonalities and dissimilarities in the observed data time series and the results of various GCMs. In this paper we want to show how this approach can be used to get insight into climate data modelling. The aim is to improve the process of climate models selection by comparing the mutual dissimilarities between simulations and observed data using a functional data analysis clustering approach.

2 Data description

This work is based on the use of observational datasets and high-resolution climate projections currently available on the Italian territory. Regarding these latter, CORDEX regional climate model (RCM) simulations for the European domain (EURO-CORDEX) at 0.11° resolution (~12 km) forced by different global climate models (GCM) were used. EURO-CORDEX is the European branch of the international CORDEX initiative, which is a program sponsored by the World Climate Research Program (WRCP, http://wcrp-cordex.ipsl.jussieu.fr/) to organize an internationally coordinated framework to produce improved regional climate change projections for all land regions world-wide. More information on the EURO-CORDEX initiative are available at the following link http://www.euro-cordex.net[1].
RCMs are numerical climate models able to perform an explicit simulation of some of the most important physical dynamics of the climate system, for a limited area. They are nested on the GCMs, usually using a one-way nesting procedure: this technique consists of using the outputs from GCM simulations, conveniently interpolated to provide initial and driving lateral boundary conditions for a high-resolution simulation only for a part of the Earth surface. GCM are forced by Representative Concentration Pathway (RCP) scenarios.
The ensemble of climate simulations used in this work consists of eighteen GCM-RCM combinations carried out in the frame of EURO-CORDEX, as shown in Table 1. The Representative Concentration Pathway scenarios used in this activity are RCP4.5 (Intermediate emissions) and RCP8.5 (High emissions), adopted by the IPCC for its fifth Assessment Report (AR5) in 2014. RCPs are time and space dependent trajectories of concentrations of greenhouse gases and pollutants resulting from human activities, including changes in land use. RCPs provide a quantitative description of concentrations of the climate change pollutants in the atmosphere over time, as well as their radiative forcing in 2100 (specifically, RCP4.5 achieves an overall impact of 4.5 watts per square meter by 2100 and RCP8.5 of 8.5 watts per square meter) (IPCC, 2014a; Moss et al, 2010).
Furthermore, one climate variable has been taken into account: precipitation.

Table 1: List of the eighteen EURO-CORDEX simulations used in this work.

Driving GCM	GCM Member[2]	RCM
CNRM-CM5	r1i1p1	ALADIN53
CNRM-CM5	r1i1p1	RCA4
CNRM-CM5	r1i1p1	CCLM4-8-17

[1] Extended guidelines for the use of the EURO CORDEX data are available at the link: http://www.euro-cordex.net/imperia/md/content/csc/cordex/euro-cordex-guidelines-version1.0-2017.08.pdf

[2] r1i1p1, r2i1p1, r3i1p1 and r12i1p1 are ensemble members in the driving global model calculation. R is an abbreviation for realization (starting point of the calculation), i for initialization method and p for physics version (https://www.medcordex.eu/cordex_archive_specifications_2.2_30sept2013.pdf).

EC-EARTH	r12i1p1	CCLM4-8-17
EC-EARTH	r12i1p1	RACMO22E
EC-EARTH	r12i1p1	RCA4
EC-EARTH	r1i1p1	RACMO22E
EC-EARTH	r3i1p1	HIRHAM5
IPSL-CM5A-MR	r1i1p1	WRF331F
IPSL-CM5A-MR	r1i1p1	RCA4
HadGEM2-ES	r1i1p1	CCLM4-8-17
HadGEM2-ES	r1i1p1	RACMO22E
HadGEM2-ES	r1i1p1	RCA4
MPI-ESM-LR	r1i1p1	CCLM4-8-17
MPI-ESM-LR	r1i1p1	REMO2009
MPI-ESM-LR	r1i1p1	RCA4
MPI-ESM-LR	r2i1p1	REMO2009
NorESM1-M	r1i1p1	HIRHAM5

3 Methodology

The main aim of this work is to explore the potential of using functional data analysis (FDA) to facilitate the comparison between simulated climate data and observed climate data. Indeed, what is unique about functional data is the possibility of also using information on the rates of change or derivatives of the curves. Slopes, curvatures, and other characteristics made available can be used because these curves are intrinsically smooth, and we can use this information in many useful ways. In this paper, we use the FDA to analyse the variations and differences over time of the simulated and observed precipitation variable and identify the simulated data that has a trend closer to that of the observed data.

The strategy proposed to identify a regional climate model, among those available under the EURO-CORDEX program, which shows the most similar form of temporal changes in the series compared to the historical series of observations, consists of an analysis with several sequential steps.

The first step is to select the grid points of the models that are representative of the study area. For this purpose, 4 methods have been used, so that a first error was not introduced precisely due to the choice to use only one of these methods. The first method consists in identifying the grid point of the model closest to the station point under examination (closest point method); the second in identifying a 3x3 box centered on the station point under examination (box 3x3 method); the third is the IDW (Watson and Philip, 1982) a deterministic spatial interpolation method which assumes that each point of the model grid has an influence that decreases with increasing distance from the station point (IDW method); the fourth is the Kriging (Bohling, 2005) a geostatistical interpolation method applying weights in measured points, according to a (moderately) data-driven weighting function (Kriging method). Those weights are based not only on distance between points (like in

IDW), but also on the variation between measured points as function of distance. In this work, the ordinary Kriging approach is adopted.
The second step was to convert the observed precipitation data and the simulated precipitation data representative of the study area, obtained with the 4 previously described methodologies, into functional data. Therefore, this step consists in construct the smooth functional curves from their corresponding simulated and observed data obtained discretely over time. In this work, aggregated data on an annual scale were considered for both simulated and observed data. Let there be N curves and let $y_i(t)$ denote the value of the i^{th} curve at time $t \in T$ (for $i = 1, \ldots, N$). The method used to represent smooth functions $y_i(t)$ over time $t \in T$ is through linear combinations of known basis functions as follows:

$$y_i(t) = \sum_{k=1}^{K} c_{ik} \Phi_k(t) = \mathbf{c}_i^{\mathrm{T}} \mathbf{\Phi}(t), \forall\, t \in T, i = 1, \ldots, N$$

where $\Phi_k(t)$ (for $k = 1, \ldots, K$) is the k^{th} known basis functions and c_{ik} is the corresponding coefficient. In the vector-matrix notation, $\mathbf{\Phi}(t)$ is the vector of all K basis functions and $\mathbf{c}$ is a vector that contains all K coefficients. The choice of the basis function is based on the characteristics of the data and the nature of the smooth curve (Ramsay and Silverman, 1997). The most appropriate basis for periodic functions defined on a time interval is the Fourier basis. The degree of smoothness imposed on the curve $y_i(t)$ is controlled by the number K of basis functions. A large K implies more flexibility and smoothness in the estimated curve. Once the basis Fourier object was set up, it was then combined with the discrete observed and simulated annual precipitation data to create the functional precipitation data objects.
The next step is to calculate the first derivative of each functional object created, expressed in the following form:

$$y_i'(t) = \sum_{k=1}^{K} c_{ik}^{'} \Phi_k(t) = \mathbf{c}_i'^{\mathrm{T}} \mathbf{\Phi}(t), \forall\, t \in T, i = 1, \ldots, N$$

The functional distance was then calculated between the coefficients of the first derivatives of the 18 functional simulations and of each of these with the coefficients of the first derivative of the observed functional data, to obtain the functional distance matrix showing the magnitude of dissimilarities. Let the i^{th} and j^{th} estimated curves $y_i(t)$ and $y_j(t)$ be expressed as a linear combination of Fourier basis functions with coefficient vectors $\mathbf{c}_i$ and $\mathbf{c}_j$ respectively. The distance between the curves can then be written as

$$d_{ij}^0 = \left(\mathbf{c}_i - \mathbf{c}_j\right)^{\mathrm{T}} W \left(\mathbf{c}_i - \mathbf{c}_j\right)$$

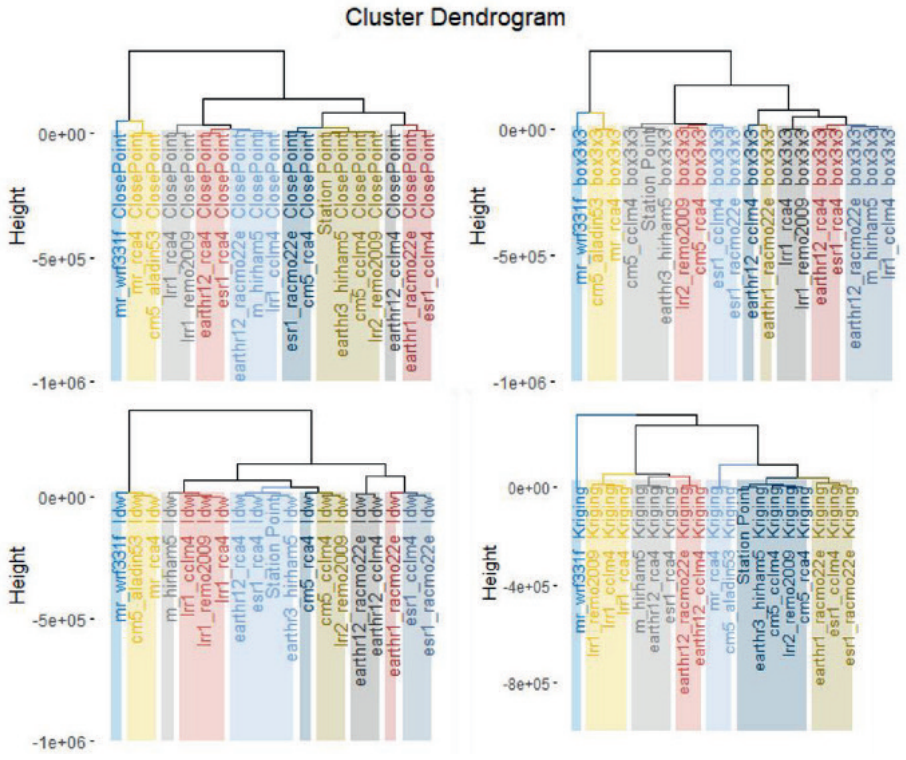
and $W = \int \mathbf{\Phi}(t)\, \mathbf{\Phi}^{\mathrm{T}}(t) dt$, which is a symmetric square matrix of order K, and K is the number of known basis functions. For each set of basis functions, W can be evaluated by using numerical integration, if necessary, and the functional distance matrix D with entries d_{ij} , as defined above, can be computed. Similarly, the distance between the first derivatives of the curves $y_i(t)$ and $y_j(t)$ is given as follows:

$$d_{ij}^{1} = \left(\mathbf{c'}_i - \mathbf{c'}_j\right)^{\mathrm{T}} W \left(\mathbf{c'}_i - \mathbf{c'}_j\right)$$

and $W = \int \mathbf{\Phi}(t)\, \mathbf{\Phi}^{\mathrm{T}}(t) dt$, which is a symmetric square matrix of order K, and K is the number of known basis functions. The distances d_{ij}^{1} are usually not affected by the bias, whereas the common bias of the simulated data is considered to have impacts on the similarity assessed by distances d_{ij}^{0}. Thus, smaller values of that distances show a similar shape of the temporal changes in the data series.

The last step consisted in applying hierarchical clustering in order to search for clusters of models that show a more similar shape in temporal variations compared to the same time series of the observations. By way of example, Figure 1 shows the dendrogram generated by the hierarchical clustering on the functional distance matrix obtained respectively for each method used to extract the grid points of the climate models representative of the study area, according to RCP4.5 scenario.

Figure 1: Dendrogram generated by the hierarchical clustering on functional distance matrix: closest point method (top left), box 3x3 method (top right), IDW method (bottom left), Kriging method (bottom right).



References

1. Fioravanti G., Fraschetti P., Perconti W., Piervitali E., Desiato F.: Controlli di qualità delle serie di temperatura e precipitazione, ISPRA, 2016.
2. Ghumman A.R., Rauf A., Haider H., Shafiquzamman M.: Functional data analysis of models for predicting temperature and precipitation under climate change scenarios. J. Water Clim. Change, v. 11, n.4, p. 1748-1765, 2020.
3. Holtanová E., Mendlik T., Kolácek J., Horová I. & Mikšovský J. (2019) Similarities within a multi-model ensemble: functional data analysis framework article in geoscientific model development. Geoscience Model Development 12, 735–747.
4. Hennemuth TI, Jacob D, Keup-Thiel E, Kotlarski S, Nikulin G, Otto et al J (2017) Guidance for EUROCORDEX climate projections data use. Version1.0 - 2017.08. Link: https ://euro-corde x.net/imperia/md/conte nt/csc/corde x/euro-corde x-guide lines - version1.0-2017.08.pdf
5. IPCC (2014a) Climate change 2014: mitigation of climate change. contribution of working Group III to the fifth assessment report of the intergovernmental panel on climate change. In: Edenhofer O, Pichs-Madruga R, Sokona Y, Farahani E, Kadner S, Seyboth K, Adler A, Baum I, Brunner S, Eickemeier P, Kriemann B, Savolainen J, Schlomer S, von Stechow C, Zwickel T , Minx JC (eds). Cambridge University Press, Cambridge.
6. ISPRA, 2012: Linee guida per l'analisi e l'elaborazione statistica di base delle serie storiche di dati idrologici, Stato dell'Ambiente 32/2012.
7. Jacob D, Petersen J, Eggert B et al. (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research, Reg Environ Chang 14:563–578. DOI:10.1007/s10113-013-0499-2
8. Jacob, D., Teichmann, C., Sobolowski, S. et al. Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. Reg Environ Change 20, 51 (2020). https://doi.org/10.1007/s10113-020-01606-9
9. Moss RH, Edmonds JA, Hibbard KA, Manning MR, Rose SK, van Vuuren DP, Carter TR, Emori S, Kainuma M, Kram T, Meehl GA, Mitchell JFB, Nakicenovic N, Riahi K, Smith SJ, Stouffer RJ, Thomson AM, Weyant JP, Wilbanks TJ (2010) The next generation of scenarios for climate change research and assessment, Nature 463:747-756.DOI:10.1038/nature08823.
10. Ramsay J.O. and Silverman BW. (1997) Functional Data Analysis. Springer: New York.

Evolution of scientific literature on climate change: a bibliometric analysis

Evoluzione della letteratura scientifica sul cambiamento climatico: un'analisi bibliometrica

Gianpaolo Zammarchi[1], Giulia Contu[1], Maurizio Romano[1]

Abstract Climate change represents one of the biggest challenges of our time. We conducted a scientometric analysis of articles on climate change indexed in Web of Science to identify the most productive countries and relevant sources. Through a topic modeling analysis of the abstracts conducted using Latent Dirichlet Allocation (LDA), we identified the main topics and their evolution during time. Through an analysis of 263,197 articles published from 1989 to 2022, we showed an increase of articles related to consequences on living beings and habitats as well as to potential actions against climate change. Our analysis provides a comprehensive picture of studies conducted on this relevant topic in the last thirty years.

Abstract *Il cambiamento climatico rappresenta una delle sfide maggiori del nostro tempo. Abbiamo condotto un'analisi scientometrica degli articoli sul cambiamento climatico indicizzati su Web of Science, per identificare i paesi che hanno prodotto il maggior numero di articoli e le fonti più rilevanti. Tramite un'analisi di topic modeling degli abstract condotta tramite Latent Dirichlet Allocation (LDA), abbiamo identificato i temi principali e la loro evoluzione. L'analisi di 263.197 articoli pubblicati tra il 1989 e il 2022 ha permesso di evidenziare un aumento degli articoli relativi alle conseguenze sugli esseri viventi e gli habitat e alle possibili azioni contro il cambiamento climatico. La nostra analisi offre una panoramica degli studi condotti su questo importante tema negli ultimi 30 anni.*

[1] Gianpaolo Zammarchi, University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, Cagliari; e-mail: gp.zammarchi@unica.it

Giulia Contu, University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari; e-mail: giulia.contu@unica.it

Maurizio Romano, University of Cagliari, Department of Economics and Business Sciences, Viale Sant'Ignazio 17, 09123, Cagliari; e-mail: romano.maurizio@unica.it

Key words: climate change, scientometrics, topic modelling, LDA, general sentiment decomposition

1 Introduction

The expressions “climate change” and “global warming” refer to the rapid and recent global rise in temperatures caused by the increase in CO_2 and methane levels due to human activities (e.g. agriculture, industrial production, overbuilding and deforestation). In 2018, the Intergovernmental Panel on Climate Change (IPCC) reported that human activities have increased global temperature from 0.8 to 1.2°C (since the pre-industrial era) and, if current levels of greenhouse gas emissions will not be reduced, the temperature will increase constantly. For these reasons, under the 2015 Paris Agreement, several nations agreed to keep warming under control, but in order not to cross the threshold of 1.5°C, it would be necessary to halve emissions by 2030 and to achieve net-zero emissions by 2050 [1]. Climate change has also profound repercussions on economic outcomes. In fact, extreme events (i.e. floods, droughts, heat waves) and other effects of the climate change (i.e. rising temperatures and sea-level) can cause wealth destruction or reduction of income [2], especially on minorities and other vulnerable groups. In addition, climate change is a problem that affects all countries. Some of its main and most severe repercussions concern anomalous heat waves and other extreme weather events, increasing number of wildfires, ice melting, deserts expansion and extinction of animal species. Human beings will also be in danger as migration flows, diseases, floods, food and water shortages and economic losses will occur in the next decades [3]. It is therefore understandable why the climate change literature has expanded rapidly in recent years [4].

In this work we conducted a scientometric analysis of scientific studies investigating climate change. Through topic modeling of retrieved abstracts, we discuss the most explored aspects as well as trends in their temporal evolution.

2 Materials and methods

We conducted a bibliometric analysis on the Web of Science Core Collection (WoS) online database, searching for studies mentioning the terms “climate change” OR “global warming” in titles, abstracts, author keywords or keywords plus (“Topic” option in WoS), published up to 29th of January 2022. A document type restriction (only Articles) and a language filter (only English) were applied. From each identified article, the following characteristics were extracted: title, abstract, keywords, authors’ affiliations, year of publication, journal title and the number of citations. We used the Bibliometrix package [5] version 3.1 in R version 4.1.1 to

display the annual scientific production, identify the most relevant sources and the most active countries. Next, the corpus of documents was preprocessed with tm [6] and quanteda [7]. A co-cluster analysis of 1,000 randomly sampled abstracts was conducted with blockcluster [8]. We tested a range from 1 to 5 clusters to find the number that allows to maximize the pseudo-likelihood. Next, we used Latent Dirichlet Allocation (LDA) in topicmodels [9] to perform topic modeling on the whole dataset. Topic names of the clusters were defined using K-means clustering over the word embeddings representation [10] in Python for dimensionality reduction (merging words with different representation but the same meaning) [11].

3 Results

We retrieved 263,197 scientific articles published from 1989 to 2022. The scientific production observed an annual percentage growth rate of 26.18 (Figure 1).

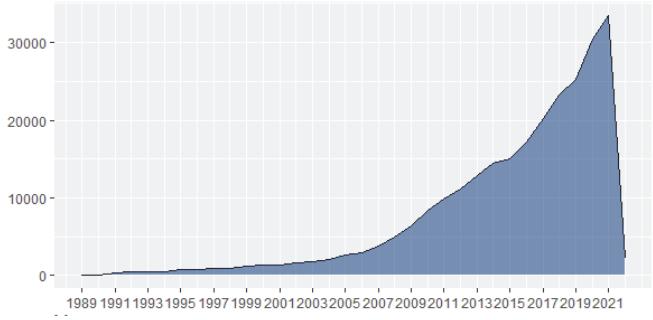


Figure 1. Annual scientific production growth of articles on climate change

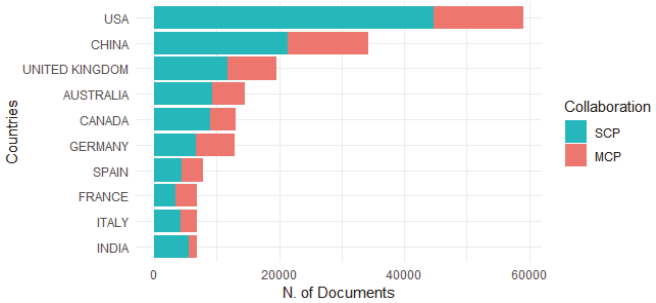


Figure 2. Most productive countries based on the affiliation of the corresponding author

United States were the country in which most articles were published, based on the affiliation of the corresponding authors, followed by China and United Kingdom (Figure 2). Single country publications (SCP) and multiple country publications (MCP) (i.e. articles with international collaborations) are also shown. The five most relevant sources based on the number of published articles are shown in Table 1.

Table 1: Most relevant sources based on the number of published articles

Journal name	Articles
Science of the Total Environment	4,272
Sustainability	3,705
Global Change Biology	3,640
PLOS ONE	3,633
Climatic Change	3,474

Analyses conducted with blockcluster on 1,000 randomly sampled abstracts showed that a five-cluster partition was able to maximize the pseudo-likelihood (Table 2).

Table 2: Pseudo-likelihood for clusters in a range from 1 to 5

Metric	1 cluster	2 clusters	3 clusters	4 clusters	5 clusters
Pseudo-likelihood	-374,555	-373,198	-409,632	-372,792	-340,785

Based on this, we set to five the number of clusters for the topic modeling analysis conducted with LDA on the whole dataset. The first ten words for each topic based on LDA are shown in Table 3. The first 100 words for each topic were used to assign topic names based on the word embeddings representation. The most representative five words based on this representation are also shown in Table 3.

Table 3: Topics and relative words identified by LDA and word embeddings

Topics	LDA	Word embeddings
Topic 1 "Ice"	model, changes, water, temperature, data, precipitation, models, results, study, surface	ice, seasonal, lake, winter, surface
Topic 2 "Community"	species, temperature, may, population, populations, effects, environmental, changes, distribution, habitat	community, populations, habitat, biodiversity, sea
Topic 3 "Fire"	soil, forest, carbon, water, vegetation, growth, increased, temperature, plant, drought	fire, season, response, forest, nitrogen
Topic 4 "Energy"	energy, emissions, carbon, production, environmental, gas, results, study, emission, system	energy, heat, power, fuel, electricity
Topic 5 "Natural resources"	management, research, environmental, adaptation, can, policy, study, risk, development, paper	natural, resources, water, human, environmental

Since LDA is a probabilistic model, any abstract will be assigned to the topics with different probabilities. For each abstract, we selected the topic with the highest probability and then plotted the number of abstracts assigned to each topic (Figure 3). We can observe that Topic 1 ("Ice") is the most represented among our articles, followed by Topic 5 ("Natural Resources") and Topic 2 ("Community").

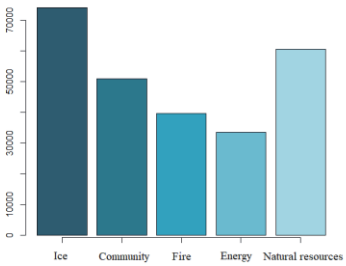


Figure 3: Number of articles for each topic

Next, we plotted the temporal evolution of probability distributions for each topic in order to observe their change over time (Figure 4). It can be observed how the probability of Topic 1 decreased during the years, the probability of Topics 2 and 5 observed an increase while that of Topics 3 and 4 remained somewhat stable, with the latter observing a slight increase from 2020.

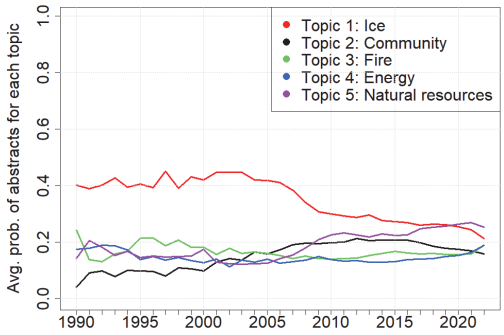


Figure 4: Number of abstracts assigned to each topic

4 Discussion

In this article, we conducted a bibliometric analysis of studies investigating climate change. We reported how the literature has grown, especially in the last two decades. This growth can be explained by the increased relevance that this topic has gained since the Kyoto Protocol signed in 1997. In particular, in recent years, activists such as Greta Thunberg have attracted the attention of newspapers and TV, also shifting the attention of public opinion towards this topic.

We built a corpus using the abstracts of the articles retrieved from the search and used a topic modeling approach to identify the main topics present in the corpus. When we plotted the temporal evolution of the topics' probabilities over the entire time span, we found Topic 2 (including words related to community) and Topic 5

(including words related to natural resources and potential actions) to slightly increase. The latter finding might be explained by the growing international efforts into providing policies and instruments to manage and reduce climate change. On the other hand, while overall Topic 1 (including words related to ice, water and precipitations) was the most represented in our dataset, its probability showed a decrease from 40% to around 20%. Studies assigned to Topics 3 (including words related to fire and vegetation) or Topic 4 (including words related to energy) remained somewhat stable. As future developments, we plan to compare results obtained using different methods for topic modeling, such as seeded LDA.

To conclude, our analysis of the large body of scientific literature on climate change provides an updated and comprehensive picture of studies conducted on this relevant topic in the last thirty years and highlights relevant trends with regards to investigated topics.

References

1. Tollefson, J.: IPCC says limiting global warming to 1.5 [degrees] C will require drastic action. Nature, **562**, 172-174 (2018)
2. Deryugina, T., Hsiang, S.M.: Does the environment still matter? Daily temperature and income in the United States. NBER Working Paper 20750 (2014)
3. TOL, Richard S.J.: The economic effects of climate change. Journal of economic perspectives. **23**, 29—51 (2009)
4. Edenhofer, O (ed.).: Climate change 2014: mitigation of climate change. Cambridge University Press (2015)
5. Aria, M., Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping. Journal of Informetrics. **11**, 959--975 (2017)
6. Feinerer, I., Hornik, K., Meyer, D.: Text Mining Infrastructure in R. Journal of Statistical Software. **25**, 1--54 (2008)
7. Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A.: quanteda: An R package for the quantitative analysis of textual data. J. Open Source Softw. **3**, 774 (2018)
8. Singh Bhatia P., Iovleff S., Govaert G.: blockcluster: An R Package for Model-Based Co-Clustering. J. Stat. Softw., **76** (9), 1--24 (2017).
9. Grün, B., Hornik, K.: topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software. **40**, 1--30 (2011)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in Proceedings of the 26th International Conference on Neural Information Processing Systems–Vol 2, Curran Associates Inc. (USA), 3111–3119 (2013)
11. Romano, M., Mola, F., Conversano, C.: Decomposing tourists' sentiment from raw NL text to assess customer satisfaction. In: ASA 2021 Statistics and information systems for policy evaluation. 147-151, FIRENZE:Firenze University Press, ISBN: 9788855183048, (2021)

Energy and material demand of the Italian Regions

Il fabbisogno di materia e di energia delle regioni italiane

Flora Fullone, Giulia Iorio, Assunta Lisa Carulli

Abstract

Material and energy consumption exerts a general environmental pressure. The economy is in a physical exchange relation with the natural environment via resource and energy flows. There are international statistical standard for measuring material and energy flow and since these concepts are aligned with those of the System of National Accounts, the environmental information can then be integrated with economic information. The work describes the main dynamics of energy and resources consumption, and the regional results obtained. It also highlights that, despite the differences between the methodological references and the sources used for the indicators of materials and energy consumption, there is a coherence at regional level between the two statistical measures.

Abstract

Le pressioni sull'ambiente esercitate dalla crescita economica possono essere descritte attraverso il consumo di materia ed energia. I conti dei flussi di materia, sono finalizzati a dare una misura del consumo di risorse, confrontabile con il PIL per condurre analisi relative alla dimensione ambientale, in parallelo a quella economica. Analogamente la dimensione energetica è descritta attraverso i bilanci energetici realizzati dall'ENEA. Il lavoro descrive le principali dinamiche del consumo di risorse, ed i risultati regionali ottenuti, inoltre evidenzia che, nonostante i riferimenti metodologici e le fonti utilizzate per la costruzione degli indicatori del consumo di materia e di energia siano molto diverse fra loro, si riscontri una coerenza a livello regionale fra le due misure statistiche.

Key words: resources, energy, total energy supply, domestic material consumption

[1] Flora Fullone, ISTAT; email: ffullone@istat.it
Giulia Iorio, ENEA; email: giulia.iorio@enea.it
Assunta Lisa Carulli, ISTAT; email: carulli@istat.it

1 Introduction

The indicators of material and energy consumption are provided in the context of sustainable development goals indicators as they represent some environmental pressures exerted by the socio-economic system. Energy consumption is assessed based on the regional energy balances (BER) provided for by law no. 10 of January 9, 1991. The BERs are consistent with the national energy balance and are developed by applying the Eurostat methodology. Similarly, the regional material consumption is evaluated based on the Eurostat methodology of the Economy-wide material flow accounts (EW-MFA), a statistical accounting framework that records, in thousands of tons per year, the material flows in and out of an economy. The work highlights different regional consumption patterns in relation with economic measure of GVA per capita and GVA composition. It also notes some correlations between the indicators relating to material and energy resource demand on regional scale.

2 Regional energy balance

The Regional Energy Balance (BER) allows to see the total amount of energy extracted from the environment, traded, transformed and used by end-users in a year giving an immediate image of the energy situation. BER

- quantifies energy supply and the energy consumption of the whole economy and of individual sectors;
- allows to observe the evolution over time of energy trends;
- allows to make a comparison with the national energy situation highlighting its diversity and problems;
- allows to evaluate the interrelationships with the socio-economic system through energy indicators such as energy intensity;
- allows to monitor the achievement of regional objectives (burden sharing).

BER is the starting point for regional energy strategies and allows monitoring impacts of energy policies.

ENEA produces Regional Energy Balances: the BERs are consistent with the national energy balance and they are developed by applying the EUROSTAT methodology[1] by adopting EUROSTAT definitions and by following EUROSTAT guidelines.

BER reports all statistically significant energy products and their production, transformation and consumption by different types of economic actors (industry, transport, etc.).

The energy flows are grouped into sections that illustrate how the energy is made available and consumed.

In a first step the energy products are evaluated in physical units and only in a second step the quantities produced and consumed are converted into a common energy unit, by multiplying all the data by the appropriate conversion factor: tons of

oil equivalent. The quantity (production+import-export+stock changes- international maritime bunkers -international aviation) is the total energy supply (TES) and it represents the quantity of energy necessary to satisfy inland energy consumption of the territory. Final energy consumption (FEC) is the total energy consumed by end users (industry, services, households, domestic transport and agriculture).

3 Regional Economy-wide material flow accounts

Economy-wide material flow accounts (EW-MFA) give a representation of the overall material inputs into regional economies, the material accumulation within the economic system and the material outputs to other economies or to the environment. These accounts include all materials other than water and air, measured in tonnes per year. Primary resources from domestic extraction and all imported/exported products (raw, semi-finished and finished) have a classification consistent with the prevailing raw material of which they are made. All classifications are based on an international statistical standard for measuring the environment and its relationship with the economy, the SEEA Central Framework[2]. A regional database covering the main material flows was compiled following standardized guidelines [3]. Based on this framework a set of EW_MFA indicators can be calculated: the domestic extraction (DE), that represents the exploitation of natural resources from the regional environment, the physical trade balance (PTB) that means the quantity of goods imported from abroad and from other Italian regions, net of the quantities exported, and finally the domestic material consumption (DMC). This consumption (DMC) takes into account the resources extracted in the territory (DE) and imported net of exports (PTB) which are used in socio-economic processes, mainly accumulated in buildings, in controlled landfills or released as residues into the environment.
The material demand records significant differences between regions and its variability is linked to the main regional structural and socio-economic indicators.[4]

Table 1 - Extensive indicators - 2018

	TES [1000 toe]	***DMC [1000 t]***	***GDP [10^6 euro]***
North	81.042	252.681	969.467
Centre	29.635	85.473	370.055
South	43.937	149.673	379.280
Italy	154.613	487.828	1.718.802

4 Analysis of regional indicators' variability

Socioeconomic processes, described mainly through the indicators of GDP and GVA, are based on the transformation and consumption of natural resources and energy as shown in Figure 1.

Figure 1 - *TES per capita on x-axis, DMC per capita on y-axis, , GVA per capita on bubble surface, the colors represent the services share in GVA: the lighter color of the bubble denotes a greater share of services in GVA, the darker color of the bubble denotes a greater share of primary and secondary activities in GVA - 2018*

The regions where the consumption of materials (DMC) and energy (TES) per capita is higher are concentrated in the upper right quadrant, at the same time there is a greater share of primary and secondary activities on GVA. In these regions, the GVA per capita is generally higher.

Some regions are mainly suited to the extraction and processing of raw materials, in other regions goods and raw materials are mainly imported. In the second case, consumption is lower than the other regions since the finished products imported and consumed have a lower weight than the raw materials used in the production processes. Liguria and Valle d'Aosta[2] have the greater share (more than 65%) of imported goods on the material consumption.

Regarding energy it can be observed that some regions are characterized by relevant energy production sector: for example Puglia consumed in 2018 over 70% of primary fuels TES to produce energy. The energy transformation sector does not have a significant weight on the GVA but involves considerable consumption of energy and material (Fig.1).

The electricity, produced by fossil fuels and biomass, also involves consumption of raw materials. So the regions that satisfy their needs by importing materials,

[2] The per capita DMC of the Valle d'Aosta can be difficult to read as it refers to the smallest region, with the minimum values of per capita extraction, population and population density, it has a strong impact of the import of goods on consumption.

commodities and energy, for the same amount of wealth produced, have lower resource consumption than the producing regions. This aspect must be considered when comparing the indicators of energy and material consumption relating to territorial contexts in which primary and secondary activities prevail, which require high consumption of materials and energy, compared to other situations in which production is outsourced and the service sector prevails (lighter color of the bubble).
Analyzing TES per unit of surface and DMC per unit of surface (Tab3), a strong link arises between the two indicators. (Pearson coefficient = 0.9).

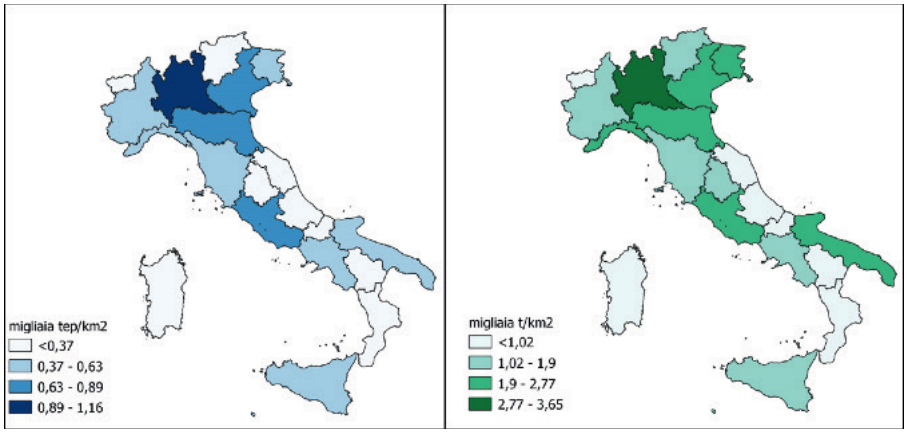


Figure 2 – Left. TES on surface ratio; Right. DMC on surface ratio

Looking at the energy dimension, the no industry FEC indicator (final energy consumption of services, households, agriculture and domestic transport) per capita has a low variability, coefficient of variation 0.25, while the per capita variability of final energy consumption in industry and construction has a medium-high variability, cv = 0.48. Furthermore, the total energy supply per capita of the regions is strongly correlated with the per capita energy consumption of industry and construction (Pearson coefficient 0.83). This suggests that the variability between regions of total energy supply per capita depends more on the industry and construction sector.
Material consumption is described applying the Eurostat EW-MFA methodology, which allows a detailed description of commodities but does not allow for a breakdown in the various sectors of the economy.
The extraction and consumption of non-metallic minerals predominate in Italy, mainly sand, gravel and other building materials directly attributable to the accumulation of infrastructures and buildings on the territory. Moreover, the per capita consumption of materials is meanly correlated with the per capita energy consumption of industry and construction (Pearson coefficient 0.61).
Finally, the efficiency measures linked to the use of materials and energy per unit of GDP are also correlated. As can be noted on Tab.4, even the TES compared to the regional GDP has a strong correlation with the material intensity calculated as the DMC ratio on GDP (Pearson coefficient 0.71).

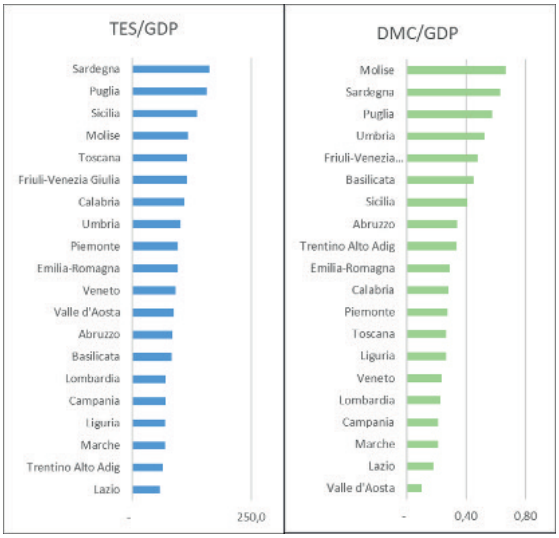


Figure 3 - Left. TES intensity (ktoe/1000 euro); Right. Material intensity (t/1000 di euro)

5 Conclusion

The consumption of energy and materials describes the pressures on the environment exerted by economic growth. The materials and energy demand of the Italian regions were compared in terms of per capita and unit area, and related to the main economic measures (GVA per capita and composition of the GVA) to highlight the dynamics of consumption and environmental pressures. This approach introduces an additional dimension to the analysis in term of sustenaible development on a regional basis, integrating the analysis based on socio-economic measures. Starting from different methodological references and sources, the analysis relating to the materials and energy demand of the Italian regions has also made it possible to find the consistency between the energy/material consumption data.

References

1. Eurostat, Energy balance guide. Methodology guide for the construction of energy balances. (2019)
2. United Nations - UN. 2021, System of Environmental-Economic 2012 - Central Framework
3. Eurostat, Economy-wide material flow accounts. Handbook. 2018 Edition. Manuals and guidelines. Luxembourg: Publications Office of the European Union (2018).
4. Istat, Ebook, Economia E Ambiente. Una Lettura Integrata (2021)

Health and survivorship

Increasing Inequalities in Mortality by Socioeconomic Position in Italy

Aumento delle diseguaglianze di mortalità per posizione socioeconomia in Italia

Chiara Ardito, Nicolás Zengarini, Roberto Leombruni, Angelo d'Errico, and Giuseppe Costa

Abstract This article assesses the evolution of socioeconomic inequality in life expectancy and mortality in Italy adopting measures of individual socioeconomic position based on income and occupational social class.
The analysis uses a large social security administrative dataset covering the population of private sector employees in Italy for the years 1990-2019 and the population of private and public sector workers from the census of the city of Turin (Italy) for the years 1981-2019. Life table techniques are used to estimate life expectancy at 65 years by different income quantiles and occupational classes. Cox regression analyses are performed to calculate the effect of socio-economic position on mortality controlling for several individual and contextual factors.
We find that inequalities in longevity have increased in recent decades. The redistributive implications are considerable since pension rules based on the average life expectancy in the population implicitly induce a regressive redistribution of pension resources.
Abstract *L'articolo stima l'evoluzione nel tempo delle disuguaglianze sociali nella longevità in Italia adottando misure di posizione socioeconomica basate su reddito e classe sociale professionale.*
L'analisi utilizza dati amministrativi relativi alla popolazione dei dipendenti del settore privato per gli anni 1990-2019 e alla popolazione dei lavoratori del settore

Chiara Ardito, University of Torino, Epidemiology Unit ASL TO3, NETSPAR – Network for Studies on Pensions, Aging and Retirement, Laboratorio R. Revelli, chiara.ardito@unito.it
Nicolás Zengarini, Epidemiology Unit ASL TO3, nicolas.zengarini@epi.piemonte.it
Roberto Leombruni, University of Torino, Laboratorio R. Revelli, roberto.leombruni@unito.it
Angelo d'Errico, Epidemiology Unit ASL TO3, angelo.derrico@epi.piemonte.it
Giuseppe Costa, University of Torino & Epidemiology Unit ASL TO3, giuseppe.costa@unito.it

privato e pubblico dal censimento della città di Torino (Italia) per gli anni 1981-2019. L'aspettativa di vita a 65 anni per diversi quartili di reddito e classi occupazionali è stimata tramite le tavole di mortalità. Analisi di regressione di Cox vengono eseguite per calcolare l'effetto della posizione socioeconomica sulla mortalità controllando per numerosi fattori individuali e contestuali.
Troviamo che negli ultimi decenni le disuguaglianze di longevità sono aumentate. Le implicazioni redistributive sono rilevanti in quanto la presenza di regole pensionistiche basate sull'aspettativa di vita media nella popolazione induce implicitamente una redistribuzione regressiva delle risorse pensionistiche.

Key words: Mortality, Inequality, Socio-economic position, Income, Occupational class, Pension system.

1 Introduction

Life expectancy inequalities do not only rise public health concerns, but they also have important consequences in terms of redistribution and equity of pension systems (e.g. [1,2,3,4,5]). Life expectancy deterministically translates in the number of years of retirement, and it is a key parameter used in pension rules. In countries adopting automatic adjustment mechanisms, life expectancy determines the pension benefits as in notional defined contributions-based systems, or the eligibility conditions, in countries like Italy, where raises in average life expectancy automatically translate into equal rises in the eligibility conditions. However, while pension rules focus on the average life expectancy in the population, important differences exist between social groups as more disadvantages categories display systematically lower life expectancy. Hence, pension systems may induce a regressive redistributive mechanism by ignoring the longevity differences across socio-economic groups.

In this work, we present updated evidence on the evolution over the last 30 years of socio-economic differentials in mortality inequality and in life expectancy at 65 in Italy, controlling for individual and contextual factors. Moreover, we exploit comprehensive administrative data from INPS, the Italian National Institute of Social Security, which covers the population of private sector employees, offering highly precise and nationally representative estimates based on the entire population rather than on samples limited to specific subpopulation or geographical areas. Our estimates are based on individual rather than aggregated data, providing very accurate measurements of socio-economic position (SEP) whereas most of previous studies on the evolution of longevity inequalities adopted measures of deprivation based on aggregated indicators at the level of areas. These are important for delivering evaluations on the longevity divide that can inform policy makers around the revision of retirement eligibility conditions for more disadvantaged categories.

2 Data and Methods

2.1 Data and variables

The empirical analysis exploits the administrative archives of the INPS. This data represents the most complete and up-to-date statistical source of information to study socio-economic longevity differentials among workers. We analyse data on the population of private sector employees registered in the INPS archives for the years 1990-2019. Furthermore, the analysis is replicated and tested on an independent information source, the Turin Longitudinal Study (TLS), a census-based database built on the censuses of the municipality of Turin, one of the four largest cities for population in Italy. For both data sources, it is possible to conduct mortality follow-up by linking administrative records with administrative mortality records up to very recent years, i.e., up to 2019. It should be noted that while INPS data allow obtaining results valid for private sector employees in Italy, the TLS database includes all workers residents in the municipality of Turin, thus offering the opportunity to extend the analysis to autonomous work and public sector too, at the cost of lower national representativeness.

For the analysis on INPS, we observe and select private sector employees in three distinct 5-year periods: 1990-1994; 1995-1999 and 2000-2004 (henceforth, we will refer to them as the "1990", "1995" and "2000" cohorts). For each cohort, we selected only individuals born in Italy, aged 15-95 at the start of the period and with at least one job spell lasting for one month or more. Then, individuals were followed until death or end of the follow-up (after 20 years from the start). By piling up all job spells observed during the 5-year periods, for each separate cohort of workers we were able to construct variables describing their work such as: average weekly wage, prevalent geographical area of work, main sector of activity, main occupational class, main firm size, average labour market attachment (% of weeks worked over the period). For TLS data we sample workers from three different censuses, i.e., 1981, 1991 and 2001.

We use two different SEP indicators. The first is occupational social class, categorized according to the European Socio-economic Classification (ESeC) into three categories: executives, white collars, and blue collars. The second is based on average weekly wage and ranks individuals according to income quartile, calculated separately for men and women. To define weekly wage, we took the sum of reported employment inflation-adjusted earnings divided by the total number of weeks worked and constructed an average weekly wage over the 5-year window for every given cohort.

2.2 Methods

Specific mortality rates were calculated for five-year age classes, sex, income, occupational class, and cohort as the ratio between the number of individuals who died in the age interval and the total population-years at risk in that age interval. Subsequently, we constructed abridged life tables using 5-year age intervals with a final age interval of 85+ to estimate life expectancy and confidence intervals (CIs) using the method described by [6], with standard errors formulas proposed in [7]. In the analysis, we focus on life expectancy at an age approaching statutory retirement age, i.e., at 65 years. The differential is computed as the difference in life expectancy between the highest-SEP and lowest-SEP group.

We complement the life expectancy analysis by estimating a set of Cox regression models on the INPS data to assess how the hazard rate (HR) of mortality for the lowest-SEP compared to the highest-SEP group changed over time controlling for work and individual characteristics. The models include two separate measures of socio-economic position simultaneously (occupational class and income quartiles) while adjusting for the following covariates: age (linear and transformed in logarithmic), macro-region of birth (5 categories), macro-region of work (5 categories), economic activity (4 ISIC categories), an indicator of work intensity constructed as the proportion of weeks worked over the 5-year window (divided in tertiles) and firm size (3 categories).

3 Results

3.1 Life expectancy in Italy by socio-economic position

Figure 1 plots the gap in life expectancy at 65 measured between the top and bottom SEP categories in the three cohorts of 1990, 1995 and 2000 using two measures of SEP: occupational social class (panel A) and income quartiles (panel B). The results show that the advantage of life expectancy in favor of workers in high socio-economic groups has increased significantly with time, regardless the indicator adopted, in both genders, although with more pronounced life expectancy gaps found among men. Being more precise, comparing individuals of different occupational class (panel A), we observe that 65-years old male executives could expect to live 1.33 years more than blue collars in 1990 (95% CI 1.23, 1.43) and this gap raised to 2.0 years in 2000 (95% CI 1.90, 2.11). Among women too it is possible to detect a significant increase with time, as the gap was null in the 1990 cohort but positive and significant in the 2000 cohort.

The divide in life expectancy at 65 widened also among individuals of different income levels. Focusing on income quartile (panel B), among men in the lowest 25% of the income distribution, the life expectancy disadvantage increased over the last 30 years, raising from 1.09 years in 1990 to 1.79 years in 2000. Women of different income groups had equal life expectancy at 65 in 1990 while a small but significant advantage of about 2 and 4 months for the richest quartile emerged in the 1995 and in 2000 cohort, respectively.

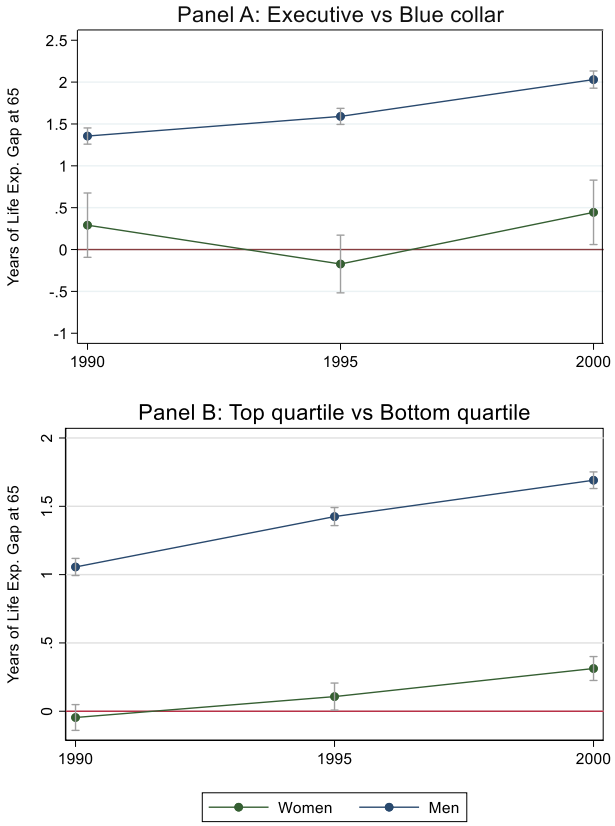


Figure 1: Evolution of the difference in life expectancy at 65 by SEP (highest SEP - lowest SEP), by sex and different SEP measures. Source: INPS data

3.2 Regression analyses on mortality

Regression results are substantially consistent with the findings from the life expectancy analyses, indicating that the socio-economic gradient in mortality

became stronger in the most recent cohorts also when adjusted for several individual and work-related characteristics. Table 1 presents gender-specific HRs and 95% CIs from Cox regression models predicting mortality by SEP in the 1990, 1995 and 2000 cohorts separately, controlling for age, region of birth and several work characteristics. The analysis is mutually adjusted for the two SEP indicators used in the previous section, i.e., occupational social class and income quartile. Mortality follow up is 20 years for all the cohorts.

Focusing on occupational social class, results show a clear time trend in inequality in mortality for both men and women: the HR of mortality for blue collars compared to executives, increased with time, raising from 1.128 (n.s.) to 1.663 ($p<.001$) for men (panel A) and from 1.051 (n.s.) to 1.175 ($p<.001$) for women (panel B). Income inequality seems to widen mortality inequality too, as the HR for the lowest income quartile raised for both men and women comparing the 1990 and 2000 cohorts.

Table 1: Cox model of hazard of death in relation to SEP and other individual characteristics, stratified by sex and cohort

	Panel A: Men		
	1990	***1995***	***2000***
	HR/ci95	*HR/ci95*	*HR/ci95*
Occupational Class:			
Blue collars	1.128	1.307*	1.663***
	[0.880,1.447]	[0.993,1.720]	[1.203,2.298]
White collars	1.069	1.147	1.327*
	[0.839,1.363]	[0.879,1.498]	[0.967,1.821]
Executives (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
Income Quartile:			
First	1.806***	2.040***	1.984***
	[1.545,2.110]	[1.740,2.392]	[1.691,2.328]
Second	1.343***	1.488***	1.471***
	[1.162,1.552]	[1.275,1.737]	[1.261,1.716]
Third	1.180***	1.297***	1.159**
	[1.042,1.336]	[1.132,1.485]	[1.007,1.335]
Fourth (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
Age	0.963**	0.954***	0.927***
	[0.932,0.996]	[0.924,0.985]	[0.898,0.958]
Ln (Age)	4.178**	7.561***	21.174***
	[1.066,16.373]	[2.059,27.770]	[5.788,77.464]
Work Intensity:			
Low	1.593***	1.734***	1.626***
	[1.403,1.810]	[1.523,1.974]	[1.425,1.855]
Mid	1.422***	1.464***	1.384***
	[1.295,1.560]	[1.324,1.620]	[1.249,1.534]
High (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
N subjects	3,987,425	4,050,416	4,294,761
N fails	321,452	269,094	238,787
N persons years	77,356,993	78,973,444	84,111,693

(Cont.)

(Cont.)

	Panel B: Women		
	1990	***1995***	***2000***
	HR/ci95	*HR/ci95*	*HR/ci95*
Occupational Class:			
Blue collars	1.051	1.174***	1.175***
	[0.967,1.143]	[1.099,1.253]	[1.086,1.271]
White collars	1.028	1.104***	1.085**
	[0.946,1.117]	[1.035,1.178]	[1.003,1.173]
Executives (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
Income Quartile:			
First	1.095***	1.156***	1.203***
	[1.071,1.119]	[1.131,1.180]	[1.181,1.226]
Second	1.022*	1.019*	1.039***
	[1.000,1.044]	[0.998,1.039]	[1.021,1.058]
Third	0.999	0.996	0.984*
	[0.980,1.019]	[0.977,1.014]	[0.968,1.000]
Fourth (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
Age	0.951***	0.960***	0.965***
	[0.946,0.955]	[0.956,0.965]	[0.961,0.969]
Ln (Age)	4.346***	3.455***	3.349***
	[3.633,5.198]	[2.896,4.122]	[2.822,3.975]
Work Intensity:			
Low	1.324***	1.342***	1.415***
	[1.300,1.350]	[1.318,1.366]	[1.392,1.438]
Mid	1.246***	1.216***	1.234***
	[1.227,1.265]	[1.199,1.234]	[1.218,1.250]
High (ref.)	1	1	1
	[1.000,1.000]	[1.000,1.000]	[1.000,1.000]
N subjects	1,984,973	2,406,683	2,808,114
N fails	45,543	52,683	66,317
N persons years	39,378,389	47,760,857	55,700,000

Notes: The table displays HR separately for gender and cohort, adjusted for the displayed covariates. Additional controls not displayed are macro region of birth; prevalent sector of activity; firm size and area of work as described in section 2.2. Source: INPS data. *p < .05; **p < .01; ***p < .001

4 Robustness

To test the robustness of results we performed several checks. First, we replicated the life expectancy analysis on a different administrative dataset, i.e., the Turin Longitudinal Study (TLS). TLS is a prospective study of mortality among persons residents in Turin (the fourth largest city of Italy) censused at the 1981, 1991 and 2001 national population censuses. The results from the analysis of the social differential in life expectancy at 65 on TLS confirm that social inequalities increased substantially among men, where the gap in life expectancy at 65 more than doubled from the 1981 census to the 2001 one. For women, no significant difference in life expectancy emerges in TLS data in all the census cohorts, suggesting that when the reference population encompasses the public sector too, longevity inequality among female workers vanishes.

Furthermore, we assessed the robustness of our life expectancy analysis by modifying the follow up duration, using more cohorts, and adopting a dynamic definition of income quartile. All these robustness tests have confirmed our main analysis showing that longevity inequalities are increasing in Italy (more details are available in [8]).

5 Conclusions

This work revealed that inequalities in life expectancy at 65 and mortality rates between income quartiles and occupational classes are widening among Italian workers.

These results have distributional implications for the pension system, too. As several studies have documented, life expectancy inequality translates into inequality of pension entitlements. In fact, individuals with a lower life expectancy spend less time in retirement and receive less than actuarially fair pension treatments. Hence, it is important for pension policy to establish compensating mechanisms that explicitly consider differential longevity. This is especially important in a pension system in which, as the Italian one, there is an automatic link of pension eligibility to the average life expectancy, an adjustment that in this moment applies also to social groups whose life expectancy is potentially stagnating.

References

1. Ardito, C., Leombruni, R., & Costa, G.: Social differences in health and equity in the Italian pension system, the Journal of Social Policy, vol. 3 (2019)
2. Caselli G. and Lipsi R. M.: Survival inequalities and redistribution in the Italian pension system, Vienna Yearbook of Population Research, vol. 16 (2018)

3. de Tavernier, W., & Boulhol, H.: Automatic adjustment mechanisms in pension systems. Chapter 2, Pensions at a Glance 2021, OECD and G20 Indicators (2021) https://www.oecd-ilibrary.org/finance-and-investment/pensions-at-a-glance-2021_d9c5d58d-en
4. Mazzaferro C., Morciano M. and Savegnago M.: Differential mortality and redistribution in the Italian notional defined contribution system, Journal of Pension Economics and Finance, 11(04) (2012)
5. Oecd, 2018 Oecd Pensions Outlook, Oecd Publishing, Paris (2018) https://doi.org/10.1787/pens_outlook-2018-en
6. Chiang, C.L.: The Life Table and its Construction. In: Introduction to Stochastic Processes in Biostatistics, pp. 189-214. John Wiley & Sons, New York (1968)
7. Eayres, D., & Williams, E. S.: Evaluation of methodologies for small area life expectancy estimation. Journal of Epidemiology & Community Health, 58(3) (2004)
8. Ardito, C., Zangarini, N., Leombruni, R., d'Errico, A. & Costa, G.: Increasing Inequalities in Mortality by Socioeconomic Position in Italy. Netspar Discussion Paper, No. 01/2022-004 (2022)

The role of health conditions in the relationship between socio-economic status and well-being: the counterfactual approach in mediation models

Il ruolo delle condizioni di salute nella relazione fra stato socio–economico e benessere: l'approccio controfattuale nei modelli di mediazione

Sara Manzella and Margherita Silan

Abstract In this work, the relationship between the socioeconomic status and the well-being of older people has been explored by investigating the mediator role played by physical health. A counterfactual approach in mediation models is applied to SHARE data, considering five waves from 2004 to 2015. The three key concepts, socioeconomic status (SES), well-being and physical health, can be seen as latent concepts and measured through factor analysis. The use of Marginal Structural Models also allowed to consider a longitudinal approach, discovering that the mediation effect of the individual's health history in the relationship between SES and well-being has a far greater impact than the one only considering the cross-sectional role of health on well-being.

Abstract *In questo lavoro è stata esplorata la relazione tra lo status socio-economico e il benessere degli anziani, indagando il ruolo di mediatore giocato dalla salute fisica. E' stato utilizzato un approccio controfattuale con modelli di mediazione applicati ai dati SHARE, considerando cinque wave dal 2004 al 2015. I tre concetti chiave, status socio-economico (SES), benessere e salute fisica, sono considerati concetti latenti e misurati tramite analisi fattoriale. L'uso di modelli strutturali marginali permette di considerare anche un approccio longitudinale, scoprendo che il ruolo di mediatore della storia di salute dell'individuo nella relazione tra SES e benessere ha un impatto molto maggiore rispetto al ruolo trasversale della salute sul benessere.*

[1] Sara Manzella, Department of Statistical Sciences, University of Padua; email: sara.manzella@studenti.unipd.it

Margherita Silan, Department of Statistical Sciences, University of Padua; email: silan@stat.unipd.it

Key words: Mediation models, Counterfactual, Well-being, Time-varying mediators, Marginal Structural Models

Introduction

The last decade has seen an increase in life expectancy in Italy and most European countries. This growth has also been accompanied by a significant increase in life expectancy in good perceived health and disability-free [1], therefore, a high proportion of older people are in good health. In this framework, the interest in analysing not only the general health of older people, but also their well-being has grown; questioning the extent of the direct and indirect impact through health, of the socioeconomic level on well-being.
Well-being is a very broad concept that encompasses different aspects of human beings characterising their quality of life in a society. It does not only include purely biological aspects, but it is important to be aware that, in reality it also combines psychological, mental, and social aspects of any individual's life.

In this work, the relationship between socioeconomic status and well-being of older people is explored by investigating the mediator role played by physical health. The method used is the counterfactual approach in mediation models [2].

Data

Data from the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and multinational database that collects panel data on health, socioeconomic status, households, and lifestyles of respondents, are used to explore ageing at the European level. The target population of the survey is people aged 50+ living in 27 European countries and Israel. From the first interview in 2004 to the sixth in 2015, more than 120000 respondents were involved. After the first wave, in addition to the longitudinal sample, a refreshment sample was introduced to cope with the dropout and exit of some subjects from the survey. Nine countries participated in all surveys (Austria, Belgium, Denmark, France, Germany, Italy, Spain, Sweden, and Switzerland), allowing both a cross-sectional and a longitudinal analysis of the data, considering the evolution of the sample over the period of time from 2004 to 2015.

Socioeconomic status, well-being, and physical health

The three key concepts, socioeconomic status, well-being, and physical health, are seen as latent concepts and measured through factor analysis.

Socioeconomic status (SES) is measured by wealth and it is based on the following one-dimensional measures collected by the questionnaire: value of main residence, value of other possessions, bank accounts and bonds, shares, and mutual funds.

To represent the concept of well-being, two indicators have been used: CASP-12, a recognised measure to assess the well-being of individuals, together with EURO-D, which allows taking into consideration purely psychological aspects of the subjects' under analysis lives [3].

The third key concept is physical health. To define it, the limitations in activities of daily living and mobility-related difficulties reported by the respondents have been used. Even if physical health is a broader concept that involves many different dimension, in this work a simplification is needed to represent it with the available data.

The literature and an initial exploratory analysis of the correlations highlighted the expected relationships between the three key concepts: improved health and well-being as the socioeconomic status of the individuals increases and increased well-being as the physical condition of the subjects improves (Figure 1).

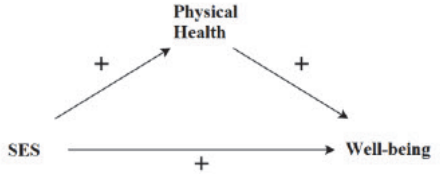


Figure 1: Graphical representation of the hypothesized relationships (and associations observed in the exploratory correlation analysis) between exposure (SES), mediator (physical health), and response variable (well being).

Methods

To trace the possible causal relationships between these variables, mediation models have been used following the counterfactual approach. This method allowed the decomposition of the total effect of socioeconomic status (treatment) on well-being (response variable) by assessing what portion of this effect is mediated by physical health (mediator).

In other words, it is hypothesized that well-being can change as a result of changes in socioeconomic status or physical health, but that physical health can also be influenced by socioeconomic status. In particular, four quantities are defined and estimated [2]:

- The Controlled Direct Effect (CDE), which quantifies whether there is a change on average in well-being if socioeconomic status varies but physical health does not change.
- The Natural Direct Effect (NDE), which expresses the average effect on well-being given a change in socioeconomic status, assuming that physical health is fixed at the value it would have been if the socioeconomic status

had not changed (i.e. the mediator does not change as a result of the change in socioeconomic status).

- The Natural Indirect Effect (NIE), which represents the average change in well-being assuming that socioeconomic status remains constant but physical health varies.
- The total effect (TE), which is simply the sum of NDE and NIE.

Several assumptions are needed to estimate these quantities [2] [4] [7] [8]. However, considering the longitudinal nature of the data coming from five waves of a panel, it is also possible to take into account time-varying exposure and mediators, despite the violation of some assumptions.

In order to handle these scenarios, Marginal Structural Models are combined with a system of weights, and the aforementioned assumptions may be adapted into a longitudinal framework [4]. Thus, it is possible to include in the analysis some important variables such as physical activity, whose inclusion would be otherwise tricky due to reverse causality. Estimation of the direct and indirect effect of physical health on well-being results from an adaptation of the VanderWeele and Tchetgen Tchetgen procedure [5]. Since in the longitudinal procedure exposition histories are considered, some weights may become particularly high due to less likely situations. Hence, the weights are trimmed in order to obtain more stable estimates [6].

Results

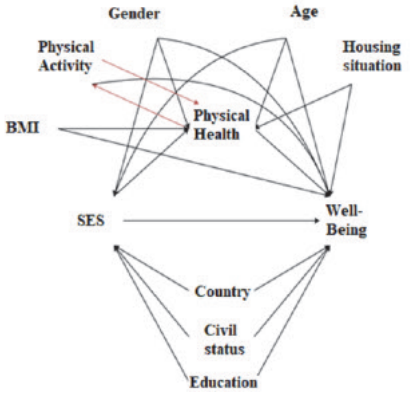
As a first step, the analysis was carried out without correcting for the presence of so-called confounding variables, estimating the necessary models to calculate the aforementioned effects without adding appropriate covariates.

This step of the analysis allows to estimate how much of the total effect is due to the mediating process of physical health. This percentage varies between 20% and 30% of the total effect, depending on the wave considered.

The effects estimated with the models without covariates cannot be interpreted as causal effects as some necessary assumptions are not respected; however, they represent a first confirmation of the mediating role played by physical health.

The term confounding refers to the bias and inaccuracy of association measures that occur when both the exposure of interest and the outcome are associated with a third variable. When conducting mediation analysis, identifying and cleaning effect estimates from this phenomenon is crucial. The method used to recognise confounding variables is called the change-in-estimate criterion, and it is based on the assessment of the proportion by which estimates of association measure change.

The relationships assumed in this second step of the analysis are considerably more complex and involve a larger number of variables (Figure 2).



Exposure Levels	TE	CDE ($m - mean\|min\|max$)	NDE	NIE
		0.0734 (0.06613: 0.08069)		
$a = \bar{a}$	0.0919	0.1116	0.0736	0.0183
$a^* = min$	(0.08404: 0.09979)	(0.04852: 0.1747)	(0.06626: 0.08097)	(0.01533: 0.02127)
		0.0705 (0.06315: 0.07781)		
		2.0673 (1.8636: 2.2711)		
$a = max$	2.3849	3.1457	2.0673	0.3175
$a^* = \bar{a}$	(2.0678: 2.7120)	(1.3672: 4.9243)	(1.8636: 2.2711)	(0.00485: 0.6399)
		1.9861 (1.7796: 2.1926)		
		2.1407 (1.9297: 2.3517)		
$a = max$	2.4768	3.2574	2.1480	0.3288
$a^* = min$	(2.1469: 2.8067)	(1.4157: 5.0990)	(1.9333: 2.3626)	(0.00502: 0.6626)
		2.0566 (1.8428: 2.2704)		

Figure 2: Graphical representation of the hypothesised relationships between socioeconomic level (exposure factor), physical health (mediator), and well-being (outcome), considering possible confounders (on the left). Effects estimates corresponding to different levels of exposure and mediator (on the right).

A part of the results from the analysis on the sixth wave is provided below.

The outcome is well-being, a quantitative variable derived from factor analysis, standardised, and with a range of about 6. The CDE, that is, the effect of socioeconomic level (SES) on well-being for fixed values of health, is 3.3 when comparing the minimum and maximum levels of SES and setting the value of health at its minimum. Hence, the level of well-being doubles when moving SES from the minimum to the maximum, keeping health levels set at the minimum. For health levels set at the mean or maximum value, well-being increases by just over two points. That is, whenever health is at its worst, the role of SES is greater, and the gap between minimum and maximum SES increases. Similar results are obtained if the comparison is between the average and the maximum SES. If, on the other hand, the comparison is between average and minimum SES, the effect of SES on well-being is greatly reduced, regardless of the value of health. This shows that a large gap is generated between very high SES levels and low to medium levels. The NIE, that is the average change in the outcome (well-being) when the exposure (SES) is fixed and the mediator (health) takes on values corresponding to different levels of SES, can account for 10% to 20% of the total effect, depending on the wave. In absence of confounders, this share was on the order of 30%. The estimates of the indirect effects cannot be considered high and in some waves are not significant. This is particularly noticeable when comparing very different levels of SES. Complete results are available at [4].

Cross-sectional analyses were performed both in the complete sample and in the one including only subjects who participated in all survey waves (panel sample), observing the majority of differences in the significance of estimates rather then in the magnitude of the effects. Thus, a negligible selection bias is expected using this sample.

Nevertheless, these analyses do not consider the longitudinal nature of the data, excluding the possible confounding effect that the variables observed at different time instants may have on each other and on themselves. To carry out this analysis, only

the panel sample has been used (6707). The longitudinal approach assesses the cumulative influence of socio-economic status and health over time identifying a significant, positive effect of both on the well-being of individuals. The estimated total effect indicates that a change in exposure history of one standard deviation across all time periods is followed by a change in well-being of 3.18%. This analysis shows that about one third of the total effect of socioeconomic status on individual well-being is mediated by the effect of physical health. This is a far greater effect than the one emerged from the cross-sectional analyses.

Conclusions

People's history, the accumulation of risk factors, and adverse health conditions strongly influence well-being in old age. Overall, some important findings arise from the analysis: first of all, the strong impact of socioeconomic level on well-being and the strong differentials in well-being generated by comparing high and low socioeconomic levels. Furthermore, the mediation effect of the individual's health history in the relationship between socioeconomic level and well-being, accounting for 33% of the total effect, has a far greater impact than the one obtained considering the cross-sectional role of health on well-being (10-20%).

References

1. G. Boccuzzo, L. Gargiulo, L. Iannucci, M. Silan, G. Costa (2021), “La salute degli anziani tra prospettive di resilienza e fragilità”. In: F. C. Billari, C. Tomassini, Rapporto sulla popolazione. L’italia e le sfide della demografia, pp. 213-237, Il Mulino, Bologna.
2. VanderWeele, T. (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford Univ. Pr.
3. VonDemKnesebeck, O., Hyde, M., Higgs, P., Kupfer, A., & Siegrist, J. (2005). Quality of Life and Well-being. In A. Börsch-Supan, A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist, & G. Weber, Health, ageing and retirement in Europe – First results from the Survey of Health, Ageing and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).
4. Manzella, S. (2019). L'approccio controfattuale nei modelli di mediazione: studio del ruolo delle condizioni di salute nella relazione fra Stato socio-economico e benessere., Università degli Studi di Padova (http://tesi.cab.unipd.it/62422/1/Manzella_Sara.pdf).
5. VanderWeele, T., & Tchetgen Tchetgen, E. (2016). Mediation analysis with time varying exposures and mediators. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3), 917-938.
6. Cole, S., & Hernán, M. (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology(168), 656-664.
7. Robins, J., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology(3), 143-155.
8. Pearl, J. (2001). Direct and indirect effects. Seventeenth Conference on Uncertainty in Artificial Intelligence (p. 411-420). San Francisco: Morgan Kaufmann.

Excess economic burden of multimorbidity: a population-based study in Italy

Eccesso di costo economico della multimorbilità: uno studio di popolazione in Italia

Chiara Seghieri, Niccolò Borri, Gaia Bertarelli and Sabina Nuti

Abstract In Italy the increasing incidence of chronic disease and multimorbidity are major challenges for health systems. When a patient suffers from more than one chronic condition, the conditions can interact causing a significant increase in patients' care needs. Using healthcare administrative databases of Tuscany region to identify cohorts of chronic prevalent patients and their total direct healthcare expenses, in this paper we aim to study the economic burden of multiple chronic conditions and calculate the excess cost when comorbidities occur in order to assess how combinations of chronic conditions in adults affect total direct health expenditure.

Abstract *In Italia la multi morbidità è una sfida importante per i sistemi sanitari. Quando un paziente soffre di più di una condizione cronica, le condizioni possono interagire causando un aumento dei costi sanitari del paziente. Tuttavia, resta da determinare la misura in cui le co-morbidità influiscono sui costi. Utilizzando i database amministrativi sanitari della regione Toscana per identificare le coorti di pazienti prevalenti e le loro spese sanitarie dirette totali, in questo lavoro vogliamo studiare l'onere economico a carico del Sistema Sanitario Nazionale di molteplici condizioni croniche e calcolarne l'eccesso di costo quando si verificano co-morbidità al fine di valutare come le combinazioni di condizioni croniche negli adulti incidono sulla spesa sanitaria diretta totale.*

Chiara Seghieri
Management and Healthcare Lab, Institute of Management, EMbeDS department, Sant'Anna School of Advanced Studies, 56127, Pisa, Italy, e-mail: chiara.seghieri@santannapisa.it

Niccolò Borri
Central European University e-mail: Borri_Niccolo@phd.ceu.edu

Gaia Bertarelli
Management and Healthcare Lab, Institute of Management, EMbeDS department, Sant'Anna School of Advanced Studies, 56127, Pisa, Italy, e-mail: gaia.bertarelli@santannapisa.it

Sabina Nuti
Management and Healthcare Lab, Institute of Management, Sant'Anna School of Advanced Studies, 56127, Pisa, Italy,e-mail: sabina.nuti@santannapisa.it

Key words: Health Care Costs, Multimorbidity, Comorbidity, Expenditures, Administrative Data.

1 Introduction and background

The Italian National Health System is going through a period of strong criticality and reflection, with reference to its sustainability at an economic-financial level. This occurs for several reasons, including the gradual increase in subjects suffering from Noncommunicable diseases (NCDs) that require adequate continuum of care. NCDs, also known as chronic diseases, are recognised as the leading causes of death and disability worldwide. According to WHO, 70% of the deaths worldwide are due to NCDs. Multimorbidity, which is the presence of two or more chronic conditions in the person, is particularly relevant in Italy where more than 7 million people - 11.7% of the total population - are more than 75 years old and 42.3% of them have multiple chronic conditions (Istat, 2020). As consequence, the increasing of multimorbidity is a major health-related challenge for patients, health professionals, and society more broadly: it is associated with poor health outcomes, including increased mortality (Nunes et al., 2016), worse quality of life (Makovski et al., 2019), functional decline and increasing use of healthcare services (Ryan et al., 2018) thus imposing considerable economic burden on health systems (Wang et al., 2018). Nonetheless, many people with chronic conditions, and in particular those with multiple conditions, are still failing to receive appropriate care. Despite the guidelines that lead to suggesting bringing the patient to the center of the health care process (Starfield, 2009), health care systems are still organised around a single disease model with professionals working in a fragmented way, focusing on the intervention for the treatment of a specific disease rather than on the patient overall. Prevention and treatment services should shift from a single disease focus to a person-centred approach, so that morbidity gains match that of longevity; a health, rather than simply illness, approach to the prevention and care of patients is warranted to protect and promote maintenance of good health and to contribute to the sustainability of the health systems (Pearson-Studdard, 2019).
Knowing the absorption of resources related to the chronic population in charge by the Health System is very important and can help policy and decision makers in the health planning phase, in order to allocate adequate resources with respect to the entire path of diagnosis and treatment of these patients. In this sense, the present study aims at identifying, quantifying, and characterizing in terms of demographic characteristics and direct costs the combinations of chronic disease in a population of multimorbid patients in Tuscany Region (Italy). We will then estimate the economic burden of multiple chronic conditions by identifying selected conditions and calculating the excess costs of the specific condition when other conditions co-occur in order to know more about how combinations of chronic conditions in adults affect total direct health care expenditures and increase awareness among health profes-

sionals on how the healthcare delivery should move to a value-based health service, based on patient perspective and needs.

2 Data and Methods

We conducted a population-based study using administrative healthcare data of the Tuscany region at the individual level. This study utilizes a database consisting of all residents enrolled in the Tuscan health care system, aged 16 or more, alived in 2019 and who were classified as affected or not affected by one or more specific chronic diseases according to administrative data algorithms.
The Tuscany healthcare administrative databases contain information on all public and private accredited healthcare providers. The individual-level databases used in the present study include: (i) hospital inpatient data; (ii) emergency care data; (iii) outpatient care data; (iv) drug prescription data (v) exemption data; and (vi) registered person database which contains demographic information on all residents enrolled in the Tuscan health care system including, sex, date of birth and date of death. The different administrative databases were linked at the individual (patient) level through a unique identifier. Aggregate direct care costs of the year 2019 (1 January 2019 and 31 December 2019) were calculated for each selected chronic patient by considering cost information from the following administrative databases: inpatient care (DRG tariffs), drugs (net costs) and outpatient services (outpatient tariffs), emergency department care (tariffs).
In order to conduct our analysis, we selected six non-communicable diseases that contribute to the bulk of deaths, preventable disabilities and morbidity worldwide, namely heart disease, heart failure, stroke, cancer, diabetes, and chronic respiratory disease (World Health Organization, 2020).

Direct care costs for the selected prevalent population were estimated using Generalized Linear Models (GLMs) given the low number of zero costs (1.6%). Three different GMLs were tested, specifying several families (Gamma, Poisson, Inverse Gaussian) and a link log. To decide on the correct specification of the model, we employed the Modified Park test and Pregibon's Link test which indicated that the Poisson-Log link model was adequate compared to other types of distributions. More specifically, we run a log-poisson model separately for each of the six selected pathologies using as independent variables: gender, age classes (16-34, 35-54, 55-69, 70-84 and over 85), the 5 binary chronic disease variables and the interaction of each pathology with each of the other selected pathologies. Additionally, in order to control for the presence of a chronic condition outside of the list of the other 5 selected chronic conditions, we also added a variable for the number of any other chronic condition, and interaction between the number of other chronicities and each of the selected disease. The estimated excess costs of one condition when other co-occur were then estimated through the recycled predictions methods (Glick et al., 2014). The increase in healthcare expenditure attributable to each selected disease

was estimated by subtracting average predicted expenditure for sick people from average predicted expenditure for the individuals with the other disease variables set to 0 or 1 in case of comorbidity for a given disease. This allows us to analyzed the excess cost of chronic combinations using all the permutations for every possible couple of diseases. Therefore we were able to simulate the impact of a specific chronic condition on an already existing one among the six considered, and controlling for the presence of any other co-occurring chronic condition. With this method, as an example, we were able to estimate the excess cost attributable to diabetes in a cohort of patients with cancer and vice versa. All considered models were also run separately by gender to analyse potential gender differences in the excess costs for each combination of conditions.

3 Preliminary Results

Overall, in 2019, 1,257,413 (approx. 40%) of the 3,203,190 registered residents in Tuscany aged 16 years old or more, had one or more of the 36 chronic diseases identified by the algorithms and were included for analysis. Among these patients, 55% were female and 45% men; as expected, the percentage of patients with at least one chronic condition increased with increasing age (Table 1).
The economic burden of chronicity in 2019 was 30% of the total healthcare expenditures (23,5% only considering the selected six diseases) and, on average, the yearly direct costs per patient with at least one chronic condition were 2436€(SD 6913€). Multimorbidity affected about half of the chronic patients (39% of all Tuscan population aged 16 years or more having two or more chronic conditions) and the average number of conditions per person was equal to 1.97 (SD 1.28).
The six conditions identified for the study counted for 50.25% of the chronic cases (69% if excluding hypertension and hypercholesterolemia) and 77% of the healthcare expenditures of chronic patients. Of these identified diseases, heart disease accounted for the highest number of cases (21.1%) followed by cancer (15.7%), COPD (7.7%), heart failure (6.7%), diabetes (6.4%) and stroke (6.3%) (Table 2). Among these conditions, the most prevalent combinations, as expected, were heart disease with heart failure (5.1%) followed by diabetes with heart diseases (4.7%). Table 3 shows that for all the six selected conditions, healthcare expenditures increased with each additional chronic condition (each different morbidity counts one for this analysis), with, approximately an exponential trend, except for cancer which is likely to show a linear behaviour on costs as the number of chronic conditions increases.

4 Strengths, limitations and future analyzes

The proposed short paper present preliminary results of the analysis of costs of the healthcare system in multichronic patients for residents in Tuscany through adminis-

	N	%	Mean of chronicities	Std. Dev.
16-34	50884	4.05%	1.14	0.44
35-54	219942	17.49%	1.35	0.74
55-69	368943	29.34%	1.78	1.1
70-84	451184	35.88%	2.3	1.38
85+	166460	13.24%	2.6	1.5
Male	563629	44.8%	2.08	1.35
Female	693784	55.2%	1.88	1.2
1 condition	633066	50.35%		
2 conditions	297368	23.64%		
3 conditions	162302	12.91%		
4 conditions	87960	7.00%		
5 conditions	44077	3.51%		
6 + conditions	32640	2.59%		

Table 1 Number and percentages of patients with chronicities in patients group according to age, sex and number of chronicities. Percentages sum up to 100 since we considered the share of each class over the total number of chronic and multichronic patients in Tuscany, t-test for gender gives us a significant difference between the two group in the number of chronicities, a oneway ANOVA test is performed for the groups of age range giving again significant difference in the number of chronicities.

	Prevalence rate	Diabetes	Heart failure	Heart disease	Stroke	COPD
Diabetes	16.4%					
Heart failure	21.1%	4.7%				
Heart disease	6.7%	2.0%	5.1%			
Stroke	6.3%	1.5%	3.2%	1.1%		
COPD	7.7%	1.4%	2.8%	1.5%	0.9%	
Cancer	15.7%	2.1%	3.1%	1.0%	1.0%	1.3%

Table 2 Prevalence of the combination of the six selected index chronic conditions in the study population.

Index disease	One disease	Two diseases	Three diseases	Four diseases	Five diseases	Six or more diseases
Diabetes	995(±29)	1680(±33)	2615(±53)	4030(±90)	5757(±127)	9175(±165)
Heart failure	2434(±308)	2166(±148)	3472(±102)	4805(±126)	6361(±136)	9514(±167)
Heart disease	1254(±47)	2874(±105)	4283(±118)	5692(±141)	7344(±194)	9988(±199)
Stroke	3295(±305)	3763(±169)	4282(±218)	5068(±141)	6110(±168)	8941(±186)
COPD	1700(±59)	2408(±47)	3445(±48)	4744(±77)	6390(±102)	9326(±140)
Cancer	3918(±78)	4720(±88)	5607(±101)	6816(±167)	8159(±210)	10445(±435)

Table 3 Costs in euros for the six selected index chronic conditions according to the number of comorbidities. Each different morbidity counts one for this analysis.

trative data. Estimates of the excess costs of a specific chronic condition when other chronic conditions co-occur, also stratified by gender, are an ongoing research.
There are limitations to this study. It utilizes only administrative data sources, which although widely available at reasonable cost, refer to health problems for which people seek medical care, therefore they might not provide correct prevalence for specific diseases. Additionally, they do not provide information on the gravity of the diseases and socio-economic details of the patients. Moreover, the type of data only allows the identification and estimation of direct medical costs thus underrepresenting the economic burden of the care of multimorbidity. It is also important to underline that in our analysis only 6 chronic diseases are considered and they only concern the 23.5% of healthcare costs in Tuscany, even if they contribute to most of deaths and preventable disabilities (World Health Organization, 2020). As to the strengths, this is a population study covering a large and well-characterized population with chronic conditions, and to the best of our knowledge it is the first study that shows the extent to which combinations of selected conditions contribute to the excess cost burden associated with multicronicity in Italy.

References

1. Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. Economic evaluation in clinical trials. OUP Oxford (2014)
2. Istat. Aspetti di vita degli over 75 Condizioni di salute, vicinanza ai figli, disponibilità di spazi esterni all'abitazione, cani in casa (2020) Available at: https://www.istat.it/it/files/2020/04/statisticatoday_ANZIANI.pdf (Accessed: 23 June 2021)
3. Makovski, T. T., Schmitz, S., Zeegers, M. P., Stranges, S., & van den Akker, M. Multimorbidity and quality of life: systematic literature review and meta-analysis. Ageing research reviews, 53, 100903 (2019)
4. Nunes, B. P., Flores, T. R., Mielke, G. I., Thumé, E., & Facchini, L. A. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. Archives of gerontology and geriatrics, 67, 130-138(2016)
5. Pearson-Stuttard, J., Ezzati, M., & Gregg, E. W. Multimorbidity—a defining challenge for health systems. The Lancet Public Health, 4(12), e599-e600 (2019)
6. Ryan, A., Murphy, C., Boland, F., Galvin, R., & Smith, S. M. What is the impact of physical activity and physical function on the development of multimorbidity in older adults over time? A population-based cohort study. The Journals of Gerontology: Series A, 73(11), 1538-1544 (2018)
7. Starfield, B. Primary care and equity in health: the importance to effectiveness and equity of responsiveness to peoples' needs. Humanity & Society, 33(1-2), 56-73 (2009)
8. Wang, L., Si, L., Cocker, F., Palmer, A. J., & Sanderson, K. A systematic review of cost-of-illness studies of multimorbidity. Applied health economics and health policy, 16(1), 15-29 (2018)
9. World Health Organization. (2021). Noncommunicable diseases. Available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (Accessed: 23 June 2021)

Depression-free life expectancy among 50 and older Americans by gender, race/ethnicity and education: the effect of marital disruption

Aspettativa di vita libera dalla depressione tra gli americani over 50 per genere, etnia e istruzione: l'effetto della rottura del legame matrimoniale

Alessandro Feraldi and Cristina Giudici

Abstract Depression is a common mental health disorder, positively associated with mortality and morbidity, especially in the elderly. The study examines differences in Depression-Free Life Expectancy by gender, race/ethnicity, education and marital status between 2012 and 2018, using a cohort of 50 and older Americans of the Health and Retirement Study**.** On average, people not in union anymore could expect to live less years in total as well as less years free of depression than people in union. Women, who were also more likely to be not in union anymore, could expect to live more years with depressive symptoms in late life than men. Estimates of depression-free life expectancies are important as they may contribute to the definition of current and future social and medical service needs and policies.

Abstract *La depressione è un disturbo mentale comune, associato positivamente a mortalità e morbilità, soprattutto negli anziani. Lo studio esamina le differenze nell'aspettativa di vita libera da depressione per genere, etnia, istruzione e stato civile tra il 2012 e il 2018, utilizzando una coorte di americani over 50 intervistati attraverso l'indagine Health and Retirement Study. In media, gli individui non più coniugati possono aspettarsi di vivere meno anni in totale e meno anni liberi dalla depressione rispetto a quelli sposati. Le donne, le quali hanno maggiore probabilità di non essere più sposate, possono aspettarsi di vivere più anni con sintomi depressivi in tarda età rispetto agli uomini. Le stime dell'aspettativa di vita libera dalla depressione sono importanti in quanto possono contribuire alla definizione delle esigenze e delle politiche dei servizi sociali e medici attuali e futuri.*

Key words: depression-free life expectancy, multistate life tables, marital disruption, ethnic groups, educational differences, gender differences

---

Alessandro Feraldi, Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro, 5 (00185), Rome; email: alessandro.feraldi@uniroma1.it

Cristina Giudici, MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy, e-mail: cristina.giudici@uniroma1.it

1. Introduction

Depression is among the most common mental disorder worldwide that can adversely affect both mental and physical health [12]. In 2015, the prevalence of depression and the lifetime prevalence in the global population were 4.4% and 10.0%, respectively [12]. Additionally, it has been observed a positive associations of depression with mortality and morbidity [7, 12]. Although this increased mortality risk may be indirectly caused by chronic conditions and risky behaviours (e.g. excess in alcohol and drug abuse), depression is likely to be responsible for a proportion of the increase in mortality risk [7].

Important studies have shown a higher prevalence of depressive symptoms among women compared to men at various stages of the life-course in the U.S. In their meta-analysis of 24 studies on people aged 75 and older, Luppa and colleagues [10] found that the men-to-women depression prevalence ratio of was 1:1.4–2.2. Yet, several studies have shown that gender differences in life expectancy are partially explained by social and economic status which differ not only by gender but also by education, ethnicity and marital status [1, 10, 11]. For example, low educational attainment is associated with higher depressive symptom burden in older adults [11]. Other evidences in the U.S. suggested that Black older adults have higher depressive symptoms and psychological distress than White, whereas Hispanics have lower lifetime prevalence of major depressive episodes [2].

2. Marital disruption and depression

It is generally well recognised that marital status and transitions have important implications for health. On the one hand, marriage can be protective for health and reduce morbidity and mortality: compared to unmarried people, separated, divorced and widowed generally have greater social and financial support, overall healthier behavioural patterns. On the other hand, marital disruption, such as separation, divorce and widowhood, are stressful life events that have been associated with poor health and survival outcomes [5]. Although divorces that occur after age 50 (grey divorce) has doubled between 1990 and 2010, marital disruption in late life mostly occurs through spousal death. Recent marital disruption has been associated with increased alcohol intake and decreased body mass index in men, whereas it has been associated with higher risk of smoking initiation/relapse in women [5]. Some studies found that men are more likely to experience widowhood-related depression than women [8]. In a recent study on a sample of middle-aged and older Australian adults, after adjusting for socio-demographic characteristics, the authors found that people who recently divorced had much higher odds depression (OR = 2.9) and divorce had a much stronger impact on depression in men than women [5]. Finally**,** to date, there has been limited longitudinal research on how divorce/widowhood affects mental health expectancies

As shown, several studies have provided important estimates of the prevalence of depressive symptoms and major depressive symptoms among older adults in U.S. as well as the factors associated with this condition. However, little is known about the dynamics of length of life with and without depression among recent cohorts of U.S. older adults and the effect of marital disruption, particularly by gender, race/ethnicity, and educational attainment. Although early reports provided estimates of mental ill-health and depression-free life expectancies in the mid-to-late 1990s, studies on this field have slowdown.

The aim of this study is to estimate the life expectancy with depression (DepLE), the depression-free life expectancy (Dep-FLE) as well as the total life expectancy (TLE) by

gender in U.S. adults and examine differences in race/ethnicity and education over a 7-year period (2012–2018). Additionally, we investigate differences in Dep-FLE due to distinctive marital status: in union and not in union.

3. Data and methods

3.1 Sample

The study uses data from the RAND version of the Health and Retirement Study (HRS), an ongoing nationally representative longitudinal survey of health characteristics of U.S. men and women aged 50 and older, with oversampling of minority ethnic groups [4]. Participants have been interviewed approximately every two years from 1992 to 2018 and several other cohorts have been added at each wave. Data on vital status and month and year of death are obtained through the mortality register and exit interviews. RAND HRS version is a user-friendly longitudinal data file, which is cleaned and compiled by the RAND Corporation [4]. After excluding 533 HRS respondents aged 50 years or older in 2012 who were never married (2.7%), the analytic sample consisted of 19,315 individuals.

3.2 Measurements

Depressive symptoms were measured using the eight-item Center for Epidemiologic Studies Depression scale (CESD-8). The eight-item CESD is a commonly-used and validated depressive symptom measure in older adults and it is computed as the sum of eight indicators: depression, everything is an effort, sleep is restless, felt alone, felt sad, could not get going (negative indicators) and felt happy and enjoyed life (positive indicators). A cut-off score of three was used as suggested by previous validation studies to indicate clinically relevant depressive symptoms: in each survey wave, a participant was classified as having a depression if one scored three or above on the CESD-8 in that wave. Information on marital status and race/ethnicity and educational attainment were included in the analysis: marital status was categorized as currently in union (i.e. married/partnered), hear after “in union”, and not in union anymore (i.e. divorced, separated and widowed), hereafter “not in union”; ethnicity was categorized as “White” and “non-White” (Black, Hispanic and others); education was categorized as “lower educated” (less than high school degree, high school degree or General Education Development, GED) and “higher educated” (some college, and college or more).

3.3 Statistical analysis

In order to estimate the age-specific hazard rates of transitions to depression, recovery and death, multistate life tables approach (MSLTs) were used. MSLTs is a Markov modelling of stochastic processes that involves individuals moving between a finite number of states over time, including exit and re-entry into the same state [6]. It allows to incorporate covariates into the models to relate individual characteristics to intensity rates and probabilities to better

explain the heterogeneity in the changes of depression status over time. Additionally, MSLTs applies the estimated rates to a synthetic cohort and summarizes the age-specific transition rates into durations [6], in our case DepLE and Dep-FLE. To assess the effect of marital disruption on the Dep-FLE, we used the Interpolated Markov Chain (IMaCh) software [3, 9] version 0.99r19. This method partitions the time intervals between successive interviews into shorter steps and then models the resulting transition probabilities by multinomial logistic regression on age and covariates (in our case race/ethnicity and education). Afterwards, estimated transition probabilities are used as inputs in the multistate life table. Ethnicity and education were modelled as two dummy variables enabling calculation of Dep-FLE for four groups: White, higher educated; White, lower educated; Non-White, higher educated; Non-White, lower educated. To study the effect of marital disruption to the transition rates and length of life with and without depression, marital status was included and it was measured as a time-varying covariate reflecting the current marital status at the time of the survey. Finally, the analyses were performed separated for men and women.

4. Preliminary results

Total sample comprised 11,123 women (42.4%) with mean age of 67.8 years (SD = 11.2 years) and 8,192 men (57.6%) with mean age of 67.4 years (SD = 10.6 years). Between 2012 and 2018, more women than men were not in union: 51.4% and 26.4%, respectively ($p < .05$). Total years of life, years without depressive symptoms, and years with depressive symptoms at age 65 for people in union and not in union are shown in Table 1. Estimates are presented by gender, ethnicity and education. On average, women could expect to live more years with depressive symptoms in late life than men. White women aged 65, who were higher educated could expect to live around 3.7 years more than their men counterpart, 2.3 years were free of depression and around 1.4 with depression ($p < .05$). Non-White women and men, who were lower educated and not in union, showed the largest gender difference in Dep-FLE. In this group, women aged 65 could expect to live around 4.1 years more than men. Nevertheless, 36.9% of women's remaining life expectancy was with depression whereas in men it was only 24.8% ($p < .05$). Concerning education attainment, higher educated women and men could expect to live more years than lower educated (i.e. total life years). Additionally, compared to higher educated people aged 65 who were in union, their lowed educated counterpart could expect to live around 4.2 fewer years free of depression ($p < .05$) and almost one year more with depression (men: 0.6 years; women: 1.0 year).

Table 1. Total life expectancy (TLE), years free of depression (Dep-FLE), and years with depression (DepLE) at age 65, by gender, ethnicity, education and marital status (SEs in parentheses).

		Life expectancies (years)						Proportion (%)	
Women		**TLE**	**(SE)**	**Dep-FLE**	**(SE)**	**DepLE**	**(SE)**	**DepLE**	**Dep-FLE**
White, higher educated	In union	26.0	(0.7)	22.1	(0.6)	3.9	(0.2)	15.1	84.9
	Not in union	24.3	(0.6)	20.0	(0.5)	4.3	(0.2)	17.6	82.4
White, lower educated	In union	23.0	(0.5)	18.1	(0.5)	4.9	(0.2)	21.3	78.7
	Not in union	21.5	(0.4)	16.1	(0.4)	5.4	(0.2)	25.0	75.0
Non-White, higher educated	In union	24.9	(1.1)	19.3	(0.9)	5.6	(0.4)	22.6	77.4
	Not in union	23.4	(1.0)	17.2	(0.8)	6.2	(0.5)	26.4	73.6
Non-White, lower educated	In union	22.5	(0.9)	15.3	(0.7)	7.2	(0.5)	31.8	68.2
	Not in union	21.3	(0.8)	13.4	(0.7)	7.9	(0.5)	36.9	63.1
Men									
White, higher educated	In union	22.3	(0.5)	19.8	(0.5)	2.5	(0.2)	11.4	88.6
	Not in union	20.7	(0.7)	18.0	(0.6)	2.7	(0.2)	13.0	87.0
White, lower educated	In union	18.5	(0.4)	15.4	(0.4)	3.1	(0.2)	17.0	83.0
	Not in union	17.2	(0.6)	13.9	(0.5)	3.3	(0.2)	19.1	80.9
Non-White, higher educated	In union	21.9	(1.0)	18.5	(1.0)	3.4	(0.3)	15.6	84.4
	Not in union	20.5	(1.1)	16.9	(1.0)	3.6	(0.4)	17.5	82.5
Non-White, lower educated	In union	18.2	(0.9)	14.1	(0.8)	4.1	(0.4)	22.4	77.6
	Not in union	17.2	(0.9)	12.9	(0.8)	4.3	(0.4)	24.8	75.2

In contrast to education, the stronger effect of marital disruption was reflected in years with no depressive symptoms than for total life years. This resulted in large differences in the proportion of remaining life in depression between people who were in union and who were not in union. The study revealed the double burden of lower life expectancy and higher proportion of time spent with depression for people who had a marital disruption than people who were still in union, especially in non-White women who were lower educated.

Comparisons of differences in Dep-FLE, DepLE and total life expectancy over ages are shown in Figure 1. For the sake of brevity, results are presented only for lower educated White Americans, however estimates for both White and non-White differences were very similar by marital status. Differences in the length of life with depressive symptoms between people in union and not in union were not statistically significant. Yet the latter could expect to have less total years of life expectancy as well as lower life expectancy free of depression than people in union.

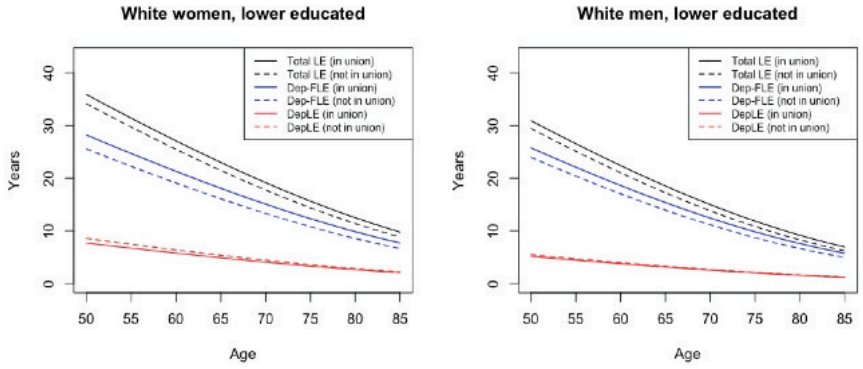


Figure 1. Total life expectancy (Total LE), years free of depression (Dep-FLE), and years with depression (DepLE) over ages, by gender and marital status.

5. Conclusion

This work gives a deeper insight of the Depression-Free Life Expectancy among older Americans, examining differences by gender, ethnicity and education. The study also aims at illustrating differences in Depression-Free Life Expectancy due to distinctive marital status (in union and not in union, i.e. marital disruption). The analysis revealed that, on average, people not in union could expect to live less years in total as well as less years free of depression. Additionally, women, who were less likely to be in union, could expect to live more years with depressive symptoms in late life than men. Our findings emphasize how estimates of depression-free life expectancies are helpful for researchers and policy makers as they may contribute to the definition of current and future social and medical service needs and policies. Furthermore, understanding the implications of marital disruption on health is relevant to the life of many around the world [5].

References

1. Abrams, L. R., & Mehta, N. K. (2019). Changes in depressive symptoms over age among older Americans: differences by gender, race/ethnicity, education, and birth cohort. SSM-population health, 7, 100399.
2. Assari, S., Moazen-Zadeh, E., Lankarani, M. M., & Micol-Foster, V. (2016). Cause mortality in the United States. Front Public Health. 2016; 4 (40).
3. Brouard N. (2019) Theory and applications of backward probabilities and prevalences in crosslongitudinal surveys. Handbook of Statistics, 40, Elsevier, pp.435-486, 9780444641526. ff10.1016/bs.host.2018.11.009ff. ffhal-02081863f
4. Bugliari, D., Carroll, J., Hayden, O., Hayes, J., Hurd, M., Karabatakis, A., Main, R., Marks, J., McCullough, C., Meijer, E., Moldoff, M., Pantoja, P., Rohwedder, S., St. Clair, P. (2021). RAND HRS longitudinal file 2018 (V1) documentation. RAND Center for the Study of Aging, Santa Monica, CA.
5. Ding, D., Gale, J., Bauman, A., Phongsavan, P., & Nguyen, B. (2021). Effects of divorce and widowhood on subsequent health behaviours and outcomes in a sample of middle-aged and older Australian adults. Scientific Reports, 11(1), 1-10.
6. Hougaard, P. (1999). Multi-state models: a review. Lifetime data analysis, 5(3), 239-264.
7. Jia, H., Zack, M. M., Thompson, W. W., Crosby, A. E., & Gottesman, I. I. (2015). Impact of depression on quality-adjusted life expectancy (QALE) directly as well as indirectly through suicide. Social psychiatry and psychiatric epidemiology, 50(6), 939-949.
8. Kristiansen, C. B., Kjær, J. N., Hjorth, P., Andersen, K., & Prina, A. M. (2019). The association of time since spousal loss and depression in widowhood: a systematic review and meta-analysis. Social Psychiatry and Psychiatric Epidemiology, 54(7), 781-792.
9. Lievre, A., Brouard, N., & Heathcote, C. (2003). The estimation of health expectancies from cross-longitudinal surveys. Mathematical population studies, 10(4), 211-248.
10. Luppa, M., Sikorski, C., Luck, T., Ehreke, L., Konnopka, A., Wiese, B., ... & Riedel-Heller, S. G. (2012). Age-and gender-specific prevalence of depression in latest-life–systematic review and meta-analysis. Journal of affective disorders, 136(3), 212-221.
11. Musliner, K. L., Munk-Olsen, T., Eaton, W. W., & Zandi, P. P. (2016). Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. Journal of affective disorders, 192, 199-211.
12. World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates; World Health Organization: Geneva, Switzerland, 2017.

Disability-free grandparenthood in Italy. Trends and gender differences

Vita da nonni libera da disabilità in Italia. Tendenze e differenze di genere

Margherita Moretti, Elisa Cisotto, Alessandra De Rose [1]

Abstract Longer life expectancies potentially translate into more years of intergenerational overlap, but this effect can be counterbalanced by the delay in the timing of grandparenthood. Grandparents are older today that in the past and it is therefore crucial to study their health, since it could affect the quality of intergenerational exchange. Applying the Sullivan method on Italian data from 1998 to 2016, disability-free grandparenthood at age 65 (period life expectancy as a grandparent free from disability) by gender is estimated, as well as the contribution of changes in mortality and grandparenthood-disability prevalence on the evolution on disability-free grandparenthood.

Abstract *Speranze di vita più lunghe si traducono potenzialmente in più anni di coesistenza fra generazioni diverse, ma questo effetto può essere controbilanciato dalla posticipazione al momento in cui si diventa nonni. Oggi i nonni sono mediamente più anziani che in passato ed è quindi cruciale studiarne lo stato di salute, da cui dipende fortemente la qualità degli scambi intergenerazionali. Utilizzando il metodo Sullivan, viene stimata la speranza di vita da nonni senza disabilità a 65 anni per genere, e viene valutato l'effetto dei cambiamenti avvenuti nella mortalità e nella prevalenza della disabilità tra i nonni dal 1998 al 2016.*

Key words: Disability, Grandparenthood, Life expectancy, Ageing

[1] Margherita Moretti, Sapienza University of Rome; email: margherita.moretti@uniroma1.it
Elisa Cisotto, Free University of Bolzano-Bozen; email: elisa.cisotto@unibz.it
Alessandra De Rose, Sapienza University of Rome; email: alessandra.derose@uniroma1.it

1 Introduction and background

Increasing longevity and changes in fertility have been two key demographic features of European countries during the twentieth century. From the one side, longer life expectancies potentially translate into more years of intergenerational overlap than in the past. On the other side, fertility delay and reduction can counterbalance the effect of the longevity revolution on kin networks. One of the intergenerational relationships most affected by these processes is grandparenthood [1]–[3]. Italian grandparents are older today that in the past [4] and whether the years of life gained in later life are of good or bad quality is a major concern [5]–[8]. Indeed, the health status of grandparents can strongly affect the generational overlap with grandchildren, both in terms of duration [9]–[11] and of quality [12]. In particular, grandparents' health is a crucial issue in understanding the direction of intergenerational transfers, since it could impact whether grandparents are providers or recipients of care: on the one hand, when grandparents are healthy, they can eventually generate intergenerational transfers, so as caring for grandchildren, or supporting (financially, functionally, or emotionally) adult children; on the other hand, when grandparents are unhealthy, they are more likely to be recipients of care, turning intergenerational relations into a caring burden for adult children or grandchildren. This is particularly relevant if we also value the beneficial effect of the time spent together, for both grandparents [13] and grandchildren [14], and how informal grandparental care of grandchildren impacts on adult children's outcomes, including labour force participation [15], [16], and fertility decisions [17]. Moreover, Italy is clearly one of the countries in which grandparents constitute a fundamental resource for the provision of informal childcare [18], [19], and families remain the most tenacious and preferred sources of support for individuals [16], [20]. The dynamics of mortality, health and fertility differ between men and women in several regards: women generally marry at a younger age (with older men), have children (thus grandchildren) earlier, live longer but in poorer health [3], [21], [22]. This implies that women can expect to live more grandparent years than men; at the same time, older women live in worst health conditions than men, and this may lead to equal or shorter period as healthy grandparents. Within this context, our study has three main objectives: first, to shed light on the evolution of the disability-free grandparenthood (DFGP) (i.e. the length of time that one can expect to live as a grandparent free from disability) at age 65 between 1998 and 2016 in Italy; second, to disentangle the DFGP evolution according to changes due to longevity revolution and due to grandparenthood-disability prevalence; third, to analyse gender differences in DFGP in Italy and its evolution.

2 Data and methods

Data are drawn from two sources. First, we use two waves of the *Family and Social Subjects* (FSS) survey of the Italian National Institute of Statistics (Istat) for 1998 and 2016[1]. Our analytical sample considers all respondents aged 65 and over, living in Italy and reporting their grandparenthood and disability status (i.e., individuals reporting having severe disability or not, and either being grandparent or grandchild-less). The second data source are Italian life tables of 1998 and 2016, by age and gender, released by Istat. Our research builds on the healthy grandparenthood measure recently introduced by Margolis and Wright [12], as we apply the Sullivan method [23] to calculate the disability-free grandparenthood (DFGP) for older women and men living in Italy in the two target years. We first compute the prevalence of individuals in each status (disability-free grandparent, grandparent with disability, disability-free grandchild-less and grandchild-less with disability). Second, we apply the age-specific prevalence of the four statuses of interest to the age-specific person-years lived from the life tables. In this way, we specify the person-years lived in each status. Finally, we sum these quantities for the ages above 65 and divided them by the lifetable survivors at age 65, resulting in residual life expectancy in the four statuses. The sum of life expectancy at age 65 in each of the four statuses equals the total life expectancy at the same age. Thus, life expectancy is partitioned into years spent being (i) disability-free grandparent, (ii) grandparent with disability, (iii) disability-free grandchild-less and (iii) grandchild-less with disability. The key outcome is the DFGP estimate (i.e., the period life expectancy as disability-free grandparent), which measures the average number of years that a hypothetical cohort of individuals can expect to live as grandparents free from disability, if they experience the mortality, disability and grandparenthood conditions observed in the studied year. Finally, by implementing the Horiuchi decomposition method [24] we analyse the age-specific contribution that the changes in mortality and grandparenthood-disability prevalence have on the evolution of DFGP from 1998 to 2016.

3 Results

Figure 1 shows the partition of overall life expectancy at age 65 (represented by the overall length of the bars and noted in black next to them) into the different states of grandparenthood and health status, by gender, for 1998 and 2016. In 1998, at the age of 65, Italian men have approximately a life expectancy of 16 years and women of almost 20. Of these, the years spent as grandparent (purple and pink part of the bars) outweigh those as grandchild-less, and only half of life expectancy is of

[1] The data processing for the year 2016 was conducted at Istat's Laboratory for the Analysis of Elementary Data (ADELE) and in compliance with the regulations on the protection of statistical confidentiality and personal data protection. The results and opinions expressed are the sole responsibility of the author and do not constitute official statistics.

DFGP (almost 9 and 10 years for men and women, respectively). Women can expect to live more years as grandmothers than men as grandfathers, but the share of years lived as grandparents over total life expectancy is similar for both gender (more than 70%). However, although women having one year more of DFGP than men, the share of DFGP on total life expectancy and on overall years as grandparents, are higher for men (respectively, 55% and 78%) than for women (50% and 70%). In fact, Italian women aged 65 in 1998 have more years with disability than men and, particularly, of those to live as grandmothers. From 1998 to 2016, as life expectancy at age 65 increases (by more than 3 years for men and 2.5 for women), DFGP at the same age increases for both genders, reaching more than 10 years for men and more than 12 for women. Over the same years, the increase in DFGP of men (of around 2 years) is at a slower pace than that of life expectancy, while for women the rise (more than 2.5 years) is faster. As a result, despite the gender difference in life expectancy decreases between 1998 and 2016, the gender difference in DFGP increases and reaches almost 2 years in 2016. Interestingly, in 2016, women and men share the same portion of DFGP life expectancy (around the 56%). However, compared to men, women display a higher portion of life expectancy as grandmothers (almost 70% vs 65% for men), and a lower share of DFGP over the total grandparent years (80% vs 85% for men), implying that the health quality of grandmothers is poorer than that of grandfathers.

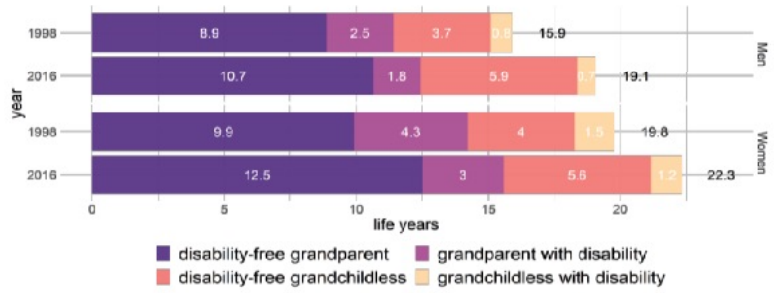


Figure 1: Life expectancy at age 65 by grandparent-disability status for Italian men and women in 1998 and 2016

Figure 2 displays the contribution of changing mortality rates and prevalence of disability-free grandparents to the change in DFGP between 1998 and 2016, by gender. For women, the contribution of mortality and prevalence of disability-free grandparents increase with ages (except for mortality contribution at age 85+). Up to the age 80, the contribution of the improved survival is more important in determining the evolution (increase) in DFGP, while for the oldest old women (aged 80+) there is a greater contribution of the prevalence of disability-free grandparents. Overall, from 1998 to 2016, there is an improvement in women's survival and an increase in the prevalence of healthy grandmothers for each age, which contributes to 2.6 years increase in DFGP. For men, mortality always contributes positively to the increase in DFGP, however, after the age of 70 its relevance decreases with age. In determining the evolution of DFGP, the prevalence of disability-free grandparents contributes less than mortality and, moreover, negatively between the ages of 65 and

74. Therefore, from 1998 to 2016, improvement in survival reduce its contribution to the change in DFGP as men gets older, while there is a reduction in the prevalence of disability-free grandparents for men, which lead to a decrease of 4 months (0.19 + 0.16 years) in the average number of years as DFGP. Knowing that overall population health improves from 1998 to 2016, this indicates that there is a noticeable reduction of the prevalence of grandparents during this period, resulting in a slowdown in the increase of DFGP, offsetting the positive effect of the reduction in disability and mortality along the observed period.

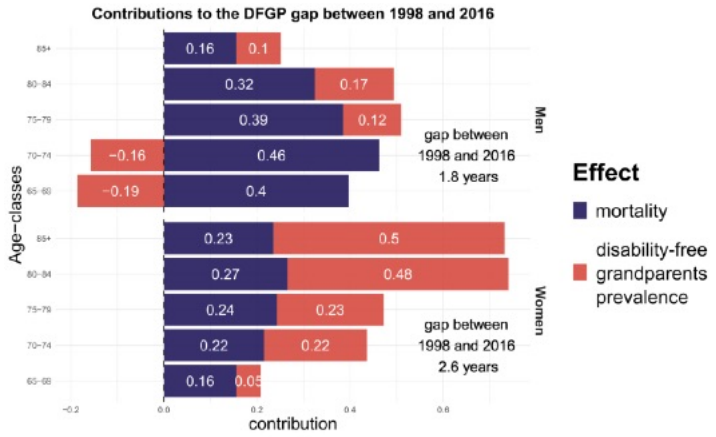


Figure 2: Mortality and disability-free grandparents prevalence contributions to the DFGP gap between 1998 and 2016 for Italian older men and women, by age classes

4 Conclusions

The average number of years individuals spend in DFGP is influenced by patterns of grandparenthood, disability, and mortality observed in the years under consideration. Because these rates and prevalence vary over time and across several dimensions, demographic methods such as Sullivan's can help to explain the effects of these changes on the quality and duration of intergenerational overlaps. Our study shows that, from 1998 to 2016, absolute DFGP and its share on the overall years spent as grandparents increases for both women and men. This reveals that, on average, Italian grandparents have more disability-free years of overlap with their grandchildren. Moreover, our results suggest that the speed of increase of DFGP is primarily led by the sharp improvement of health and survival conditions; hence, it has been slowed down by the postponement in the transition to grandparenthood. Overall, this study provides the first evidence on DFGP evolution in Italy and its gender differences. It also contributes to reflections on the roles of mortality, health, and family dynamics as measures to be considered simultaneously in a synthetic indicator, as life expectancy by disability and grandparenthood status.

References

[1] J. Skopek, “Grandparent status and multigenerational relationships,” *Research Handbook on the*

Sociology of the Family, no. June, pp. 278–299, 2021, doi: 10.4337/9781788975544.00028.
[2] M. Szinovacz, "Grandparents Today: A Demographic Profile," *Gerontologist*, vol. 38, no. 1, pp. 37–52, Jan. 1998, doi: 10.1093/geront/38.1.37.
[3] G. Di Gessa, V. Bordone, and B. Arpino, "The Role of Fertility in the Demography of Grandparenthood: Evidence from Italy," *Journal of Population Ageing*, 2020, doi: 10.1007/s12062-020-09310-6.
[4] E. Cisotto, E. Meli, and G. Cavrini, "Grandparents in Italy: trends and changes in the demography of grandparenthood from 1998 to 2016," doi: 10.31235/OSF.IO/4NH5E.
[5] J. Fries, "The Compression of Morbidity," *The Milbank Quarterly*, 1983.
[6] E. M. Gruenberg, "The failures of success," *Milbank Quarterly*, vol. 83, no. 4, pp. 779–800, 1977, doi: 10.1111/j.1468-0009.2005.00400.x.
[7] M. Kramer, "The rising pandemic of mental disorders and associated chronic diseases and disabilities," *Acta Psychiatrica Scandinavica*, vol. 62, no. 285 S, pp. 382–397, Sep. 1980, doi: 10.1111/j.1600-0447.1980.tb07714.x.
[8] K. G. Manton, "Changing concepts of morbidity and mortality in the elderly population.," *Milbank Memorial Fund Quarterly, Health and Society*, vol. 60, no. 2, pp. 183–244, 1982, doi: 10.2307/3349767.
[9] T. Leopold and J. Skopek, "The delay of grandparenthood: A Cohort comparison in East and West Germany," *Journal of Marriage and Family*, vol. 77, no. 2, pp. 441–460, Apr. 2015, doi: 10.1111/JOMF.12169.
[10] R. Margolis, "The Changing Demography of Grandparenthood," *Journal of Marriage and Family*, vol. 78, no. 3, pp. 610–622, Jun. 2016, doi: 10.1111/JOMF.12286.
[11] R. Margolis and A. M. Verdery, "A Cohort Perspective on the Demography of Grandparenthood: Past, Present, and Future Changes in Race and Sex Disparities in the United States," *Demography 2019 56:4*, vol. 56, no. 4, pp. 1495–1518, Jul. 2019, doi: 10.1007/S13524-019-00795-1.
[12] R. Margolis and L. Wright, "Healthy Grandparenthood: How Long Is It, and How Has It Changed?," *Demography 2017 54:6*, vol. 54, no. 6, pp. 2073–2099, Oct. 2017, doi: 10.1007/S13524-017-0620-0.
[13] G. Di Gessa, K. Glaser, and A. Tinker, "The impact of caring for grandchildren on the health of grandparents in Europe: A lifecourse approach," *Social Science & Medicine*, vol. 152, pp. 166–175, Mar. 2016, doi: 10.1016/J.SOCSCIMED.2016.01.041.
[14] C. A. Fruhauf and N. A. Orel, "Developmental issues of grandchildren who provide care to grandparents," *International Journal of Aging and Human Development*, vol. 67, no. 3, pp. 209–230, 2008, doi: 10.2190/AG.67.3.b.
[15] B. Arpino and V. Bordone, "Does Grandparenting Pay Off? The Effect of Child Care on Grandparents' Cognitive Functioning," *Journal of Marriage and Family*, vol. 76, no. 2, pp. 337–351, Apr. 2014, doi: 10.1111/JOMF.12096.
[16] C. Tomassini, J. Zamberletti, C. Lallo, and G. Cavrini, "Associations of family and social contact with health among Italian grandparents," *Genus*, vol. 76, no. 1, Dec. 2020, doi: 10.1186/S41118-020-00089-0.
[17] R. Rutigliano, "Counting on Potential Grandparents? Adult Children's Entry Into Parenthood Across European Countries," *Demography*, vol. 57, no. 4, pp. 1393–1414, Aug. 2020, doi: 10.1007/S13524-020-00890-8.
[18] K. Glaser and K. Hank, "Grandparenthood in Europe," *European Journal of Ageing 2018 15:3*, vol. 15, no. 3, pp. 221–223, May 2018, doi: 10.1007/S10433-018-0476-1.
[19] J. Zamberletti, G. Cavrini, and C. Tomassini, "Grandparents providing childcare in Italy," *European Journal of Ageing*, vol. 15, no. 3, pp. 265–275, Sep. 2018, doi: 10.1007/S10433-018-0479-Y.
[20] M. Kalmijn and C. Saraceno, "A comparative perspective on intergenerational support: Responsiveness to parental needs in individualistic and familialistic countries," *European Societies*, vol. 10, no. 3, pp. 479–508, 2008, doi: 10.1080/14616690701744364.
[21] A. Case and C. Paxson, "Sex differences in morbidity and mortality," *Demography*, vol. 42, no. 2, pp. 189–214, 2005, doi: 10.1353/dem.2005.0011.
[22] T. Leopold and J. Skopek, "The Demography of Grandparenthood: An International Profile," *Social Forces*, vol. 94, no. 2, pp. 801–832, Dec. 2015, doi: 10.1093/sf/sov066.
[23] D. F. Sullivan, "A single index of mortality and morbidity.," *HSMHA health reports*, vol. 86, no. 4, pp. 347–354, 1971, doi: 10.2307/4594169.
[24] S. Horiuchi, W. John R., and P. Scott D., "A decomposition method based on a model of continuous change," *Demography*, vol. 45, no. 4, pp. 785–801, 2008, doi: 10.1353/dem.0.0033.

Advances in regression models

Semiparametric M-quantile regression for modelling georeferenced housing price data

Regressione M-quantile semiparametrica per la modellizzazione di dati georeferenziati relativi ai prezzi delle abitazioni

Riccardo Borgoni, Antonella Carcagnì, Alessandra Michelangeli, Nicola Salvati, Francesco Schirripa Spagnolo

Abstract This paper applies the M-quantile regression approach to address the heterogeneity of the housing market of a modern European city, Milan (Italy). Our findings show that several housing attributes differ significantly across the response distribution suggesting that buyers of lower-priced properties behave differently from buyers of higher-priced properties.

Abstract *L'obiettivo di questo lavoro è quello di applicare l'approccio di regressione M-quantile per studiare l'eterogenità dei prezzi della abitazioni nella città di Milano, una moderna città europea. I risultati ottenuti, mostrano come alcune caratteristiche abbiano un effetto diverso sulla distribuzione condizionata dei prezzi, suggerendo che gli acquirenti delle proprietà a basso valore si comportano in modo diverso rispetto a quelli delle proprietà ad alto valore.*

Key words: Robust methods; Hedonic approach; Real Estate Observatory

1 Introduction

In this paper we use the M-quantile approach to model the distribution of housing prices conditional to housing attributes at different points of the housing price distribution. We focus on the Milan apartment market, characterized by a complex market dynamics reflecting the heterogeneity of such a big city in terms of neighbourhood, building and population features. Milan residential market has already

Nicola Salvati; Francesco Schirripa Spagnolo
Dipartimento di Economia e Management Università di Pisa, Via Cosimo Ridolfi, 10, Pisa, Italy, e-mail: nicola.salvati@unpi.it; francesco.schirripa@unipi.it

Riccardo Borgoni; Antonella Carcagnì; Alessandra Michelangeli
Dipartimento di Economia, Metodi Quantitativi e Strategia di Impresa Università degli Studi di Milano Bicocca, Via Bicocca degli Arcimboldi, 8, Milano, Italy, e-mail: riccardo.borgoni@unimib.it; antonella.carcagni@unimib.it;alessndra.michelangeli@unimib.it

been investigated in a number of quite recent studies [1, 2, 4, 8]. In particular, one of these studies, [1], uses the hedonic approach to estimate the effect of culture, public transport, education and environmental conditions on the housing market value in the city of Milan. In this paper, we maintain the hedonic framework and we propose an statistical model based on the M-quantile regression in order to obtain a more robust and efficient estimation of the hedonic price function accounting for the heterogeneities of the price market mentioned above.

For our purposes, we consider a semiparametric M-quantile regression, proposed by [10, 11] to account for potential non-linear effects of the predictors and for the spatial dependence in the response variable.

2 The Semiparametric M-quantile regression

M-quantile regression [5] is a 'quantile-like' generalization of regression based on influence functions (M-regression) able to grasp differential effect of a covariate at different levels of the conditional distribution of the response variable. Given $\boldsymbol{x}$, the linear M-quantile regression model is defined by $MQ_y(q \mid x, \psi)) = x'\beta$ where β represents a vector of unknown parameters and ψ_q denotes an asymmetric influence function. In our paper, the set $\boldsymbol{x}$ includes a range of variables representing housing-specific characteristics and urban amenities described in details in Section 3. Throughout the paper, the influence function is Huber proposal-2 influence function: $\psi(u) = uI(-c \leq u \leq c) + c\text{sgn}(u)I(|u| > c)$, with a tuning constant $c = 1.345$ that guarantees 95% of efficiency of the estimates under normality [6].

The semiparametric model for the conditional for M-quantile q considered in this paper assumes the following form:

$$MQ_Y(\mathbf{x}, t, \mathbf{s}_i; \psi) = \mathbf{x}^T\beta_q + f_{1q}(t) + f_{2q}(\mathbf{s}_i), \tag{1}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ represents two unknown arbitrary smooth functions; $\mathbf{s}_i \in \mathbb{R}^2$ represents the geographical coordinates of unit i; t is the scalar covariate expected to impact on Y nonlinearly. In this paper $f_{1q}(\cdot)$ and $f_{2q}(\cdot)$ are represented by linear combination of thin plate spline basis functions.

The regression parameters of the semiparametric M-quantile model (S-MQ) are estimated via penalised least squares by solving the following estimation equations [11]:

$$\sum_{i=1}^{n} \psi_q(y_i - \mathbf{z}_i^T\eta_q)\mathbf{z}_i^T + \lambda_{1q}\mathbf{D}_1\eta_q + \lambda_{2q}\mathbf{D}_2\eta_q = \mathbf{0}, \tag{2}$$

where $\eta_q = (\beta_q^T, \theta_{1q}^T, \theta_{2q}^T)^T$, $\mathbf{z}_i^T = (\mathbf{x}_i, t_i, t_i^2, \mathbf{b}_{1i}, \mathbf{b}_{2i})$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are two penalty matrices and λ_{1q} and λ_{2q} are the smoothing parameters estimated via external Generalised Cross Validation (GCV). In particular, the GCV to be minimised in order to

obtain $\Lambda_q = (\lambda_{1q}, \lambda_{2q})$ is $\text{GCV}(\Lambda_q) = \dfrac{||(\mathbf{I} - S_{\Lambda_q})\mathbf{y}||^2}{(1 - n^{-1}\delta \text{tr}(S_{\Lambda_q}))^2}$, where S_{Λ_q} is a smoother-type matrix associated with $MQ_Y(\mathbf{x}, t, \mathbf{s}_i; \psi)$ and and δ is a penalization term for the additional degrees of freedom given by the trace of the smoother matrix.

3 Modelling housing price in Milan using semiparametric M-quantile regression

The data come from different sources and are combined into a single data set. Housing market data are from the Real Estate Observatory (Osservatorio del Mercato Immobiliare). These data provide information about 4,000 individual housing transactions in Milan between 2004 and 2010. In addition to housing market values, the data set provides information and detailed description of housing-specific attributes of the sample units. Housing units in the sample are spatially identified by their civic address. Each civic address is geocoded by its UTM coordinates using a Java script that retrieves this information from Google Maps geographical databases. This allows us to add geocoded data about urban amenities to the housing transaction dataset. Urban amenity variables are taken from the open data portal of the municipality of Milan and the Regional Environmental Protection Agency (ARPA) of the Lombardy region. In particular, we consider the availability of public transport, education, air quality, cultural activities and related infrastructures. Finally, to control for the effect of the financial crisis of 2008 on the housing market, a binary variable that identifies all the transactions occurred before and after this year has been also defined. The list, description and definition of the variables used in the empirical analysis is given in Table 1.

Table 1 Variables description

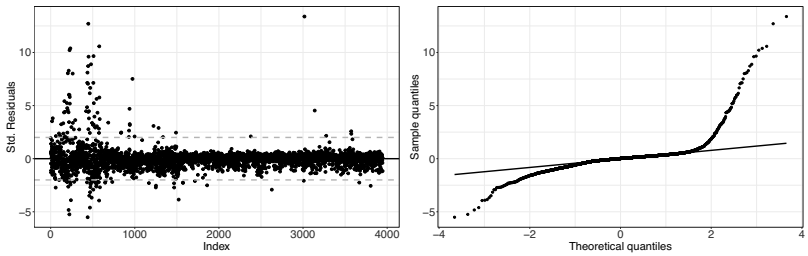
Variables	*Values*
Housing specific characteristics	
Annual market value	annual market value in Euro
Total floor area	coded on three levels < 63; (63,115]; (115,491]
Elevator	1: at least one elevator; 0 otherwise
Parking area	1: the house has a parking place or garage; 0 otherwise
Bathroom	1: two or more bathrooms; 0 otherwise
Floor	1: the house is located at the second floor or above; 0 otherwise
Heating system	1: autonomous heating system; 0 otherwise
Urban Amenities	
Cultural catalyst	Positive real values (rescaled in 0-1)
Metro	1: far from houses; 0 otherwise
University	1: far from houses ; 0 otherwise
Air quality index	1: bad air quality; 0 otherwise
Abandoned area	1 distance from houses $\leq$200; 0 otherwise
Year	1: sold in the post-crises sample period (2009-2010); 0 otherwise

To assess the need for using a robust approach a standard semiparametric linear model has been estimated:

$$\mathrm{E}(y|\mathbf{x},t,\mathbf{s}_i;\psi) = \mathbf{x}^T\beta_q + f_{1q}(t) + f_{2q}(\mathbf{s}_i), \tag{3}$$

In this paper, we include the univariate spline component, $f_{1q}(\cdot)$, to take into account the potential nonliner effect of the Cultural Catalyst (t) on the target variable (see [1]). This variable, is a composite indicator of cultural amenities available in our data set: theatres, museums, libraries and auditoria. Moreover, a bivariate spline component, $f_{2q}(\cdot)$, has been included to account for the spatial dependence. The nodes are located in the plane by the cluster separation method `clara` implemented in `R`. More in detail, $f_{2q}(\cdot)$ is a regular function in the spatial coordinates approximated by a bivariate spline while $f_{1q}(\cdot)$

We report in Fig. 1 the plot and the normal probability plot of standardized residuals. The plots indicate that the normality assumption adopted in standard semiparametric modelling does not hold and the presence of outliers. Therefore, the application of a robust model is justified.



(a) (b)

Fig. 1 Plot of standardised residuals (a) and Normal probability plots of standardized residuals (b) derived by fitting a semiparametric linear model.

Table 2 reports the results of the SMQ model. The model has been fitted by using the tuning constant of the Huber influence function equal to 1.345 and $\delta = 3$.

This results, suggest that several variables tend to have a different effect at different points of the conditional distribution of the house market values and also the statistical significance tends to change. Regarding the housing-specific characteristics, "lift", "bathroom" and the two dummies of the floor area, have a positive and statistically significant effect at all M-quantiles, moreover, the magnitude of the effect is higher at the top of the distribution than it is elsewhere. "Parking area" is not statistically significant at $q = 0.10$, while "floor" and "heating system" tend to do not affect the price of houses in Milan. Focusing on amenities variables, no variable significantly different from zero across the entire conditional distribution. More in detail, "University" is significant at every M-quantile except the most extreme ($q = 0.9$) and it slightly increases as the price increases. As expected, the neighbouring degradation (variable "Abandoned area") contributes negatively to the

Table 2 Results of SPMQ ($q = 0.10, 0.25, 0.50, 0.75, 0.90$) estimated using 20 knots for f_{1q} and 40 knots for f_{2q} - $\delta = 3$†.

Variable	$\beta_{0.10}$	$\beta_{0.25}$	$\beta_{0.50}$	$\beta_{0.75}$	$\beta_{0.90}$
Intercept	4935.782***	5209.595***	5671.031***	5523.785***	6462.095***
	(556.182)	(443.818)	(427.635)	(445.742)	(923.167)
Floor	74.654	44.52	-38.158	126.041	103.107
	(142.661)	(105.638)	(103.286)	(151.521)	(285.607)
Lift	318.348*	387.703***	438.9***	510.638***	628.974*
	(175.271)	(128.867)	(126.145)	(183.432)	(338.666)
Heating System	104.459	116.445	-117.099	-137.465	317.89
	(205.973)	(150.936)	(150.23)	(220.826)	(433.194)
Parking area	1228.605	2490.858***	2532.513***	2934.755***	2139.779*
	(798.288)	(522.463)	(511.556)	(775.048)	(1208.882)
Bathroom	1694.72***	2099.488***	2644.737***	3024.309***	3743.777***
	(188.409)	(138.653)	(138.311)	(216.687)	(455.585)
Housing Area (63-115 mq)	1384.897***	1666.37***	1813.707***	1982.92***	2068.723***
	(161.32)	(118.616)	(114.747)	(166.447)	(309.87)
Housing Area (>115mq)	4870.209***	5950.868***	8136.386***	12163.097***	17699.38***
	(230.836)	(169.916)	(172.043)	(273.719)	(586.557)
Abandoned area	-168.165	-392.739***	-625.692***	-892.985***	-1158.447***
	(203.575)	(150.37)	(145.79)	(211.22)	(379.229)
Air quality index	-209.882	-149.989	-133.472	-623.268***	-540.749
	(221.333)	(164.655)	(160.783)	(215.902)	(417.582)
University	755.285***	889.743***	904.51***	937.331***	397.294
	(160.234)	(120.549)	(122.185)	(187.219)	(367.485)
Metro	42.901	-88.366	-217.969*	-108.494	37.003
	(154.617)	(117.743)	(116.45)	(172.836)	(333.708)
Year	-439.852***	-380.757***	-305.147***	-93.855	-2.894
	(143.262)	(106.007)	(103.865)	(155.027)	(292.648)

†Point estimates with standard errors in parentheses and the associated p-value: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

house price and in particular it affects most the the price of high valued houses while it is not significant for very low valued hoses ($q = 0.1$).

4 Conclusions

The aim of this paper has been to apply the M-quantile approach to examine how the effect of housing characteristics may vary across the conditional distribution of house prices in Milan. By employing a a semiparametric M-quantile regression, the findings suggest that several housing attributes differ significantly across the response distribution, suggesting the usefulness of estimating the conditional M-quantile functions in addition to the conditional mean.

References

[1] Borgoni, R., Michelangeli, A., Pontarollo, N.: The value of culture to urban housing markets. Regional Studies **52**, 1672–1683 (2018)
[2] Borgoni, R., Degli Antoni, G., Faillo, M., Michelangeli, A.: Natives, immigrants and social cohesion: intra-city analysis combining the hedonic approach and a framed field experiment. International Review of Applied Economics **33**, 697–711 (2019)
[3] Boudreaux, D.: Globalization. Greenwood guides to business and economics, Greenwood Press, https://books.google.it/books?id=6wX3ZwCjn_IC (2008)
[4] Brambilla, M., Michelangeli, A., Peluso, E.: Equity in the city: On measuring urban (ine) quality of life. Urban Studies **50**, 3205–3224 (2013)
[5] Breckling, J., Chambers, R.: M-quantiles. Biometrika **75**, 761–771 (1988)
[6] Huber, P.J.: Robust Statistics. New York: John Wiley & Sons (1981)
[7] Koenker, R., Bassett, Jr G.: Regression quantiles. Econometrica, 33–50 (1978)
[8] Michelangeli, A., Zanardi, A.: Hedonic-based price indexes for the housing market in italian cities: theory and estimation. Politica Economica **25**, 109–146 (2009)
[9] Newey, W.K., Powell, J.L.: Asymmetric least squares estimation and testing. Econometrica, 819–847 (1987)
[10] Pratesi, M., Ranalli, M.G., Salvati, N.: Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US., Environmetrics **19**, 687–701 (2008)
[11] Pratesi, M., Ranalli, M.G., Salvati, N. Nonparametric m-quantile regression using penalised splines. Journal of Nonparametric Statistics **21**, 287–304 (2009)

Resampling-Based Inference for High-Dimensional Regression

Inferenza Tramite Ricampionamento per Regressioni ad Alta Dimensionalità

Anna Vesely and Jelle J. Goeman and Angela Andreella and Livio Finos

Abstract We propose a novel procedure for resampling-based multiple testing in high-dimensional regression. First, we construct permutation test statistics for each individual hypothesis by means of repeated random splits of the data. In each split, half of the observations is used to perform variable selection, and half to build test statistics for the selected variables. Then we define an asymptotically exact test for any subset of hypotheses by aggregating the individual statistics through a suitable function, e.g., maximum or weighted sums. The procedure is flexible, allowing different selection techniques and combining functions. It can be embedded into closed testing methods to make simultaneous confidence statements on the proportion of true discoveries (TDP) of all subsets, valid even under post-hoc selection.

Abstract *Si propone un metodo basato sul ricampionamento per test multipli in regressioni ad alta dimensionalità. Si definiscono statistiche test per ogni singola ipotesi tramite partizioni casuali dei dati, in cui metà delle osservazioni sono usate per selezionare le variabili, e metà per costruire statistiche per le variabili selezionate. Aggregando tali statistiche, ad es. con il massimo o somme pesate, si ottiene un test asintoticamente esatto per ogni sottoinsieme di ipotesi. La procedura è flessibile, poiché ammette diverse tecniche di selezione e diverse funzioni di combinazione. Può essere utilizzata all'interno di metodi di closed testing per ottenere*

Anna Vesely
Department of Developmental Psychology and Socialization, University of Padua, Italy, e-mail: anna.vesely@unipd.it

Jelle J. Goeman
Biomedical Data Sciences, Leiden University Medical Center, The Netherlands, e-mail: j.j.geoman@lumc.nl

Angela Andreella
Department of Economics, University of Venice, Italy, e-mail: angela.andreella@unive.it

Livio Finos
Department of Developmental Psychology and Socialization, University of Padua, Italy, e-mail: livio.finos@unipd.it

limiti di confidenza per la proporzione di variabili attive (TDP), simultaneamente su tutti i sottoinsiemi di ipotesi.

Key words: high-dimensional regression, multiple testing, Multisplit, resampling-based test, true discovery proportion

1 Introduction

In linear regression, interest usually lies in discovering relevant predictor variables and assessing statistical significance; however, many challenges arise in high-dimensional settings. Researchers are often interested in studying subsets of variables with an exploratory approach, quantifying activation inside. Moreover, when they do not know a priori which subsets they are interested in, they may want to study many and make the selection post hoc.

We propose a multiple testing method for high-dimensional linear regression that defines a test for any subset of variables and ensures asymptotic error control. We use the permutation framework, which is often more powerful than the parametric approach, especially when considering multiple hypotheses [2]. The method relies on two building blocks: the Multisplit proposed by [4], that computes adjusted p-values for all variables using variable selection techniques and repeated splits of the data, and the sign-flipping test given in [3].

First, we construct permutation test statistics for all variables. Then we aggregate these individual statistics to define an asymptotically exact test for any subset of variables. Different combining functions are possible, including the maximum and weighted sums. As we can test any subset, the procedure can be embedded into closed testing methods that give simultaneous confidence sets for the true discovery proportion (TDP), such as [1] and [6]. This way we are able to provide confidence statements on the TDP of all subsets, valid even under post-hoc selection.

The structure of the paper is the following. We introduce the model and its assumptions in Sect. 2, then we define the method in Sect. 3. Finally, in Sect. 4 we compare the proposed method and the Multisplit through simulations.

2 High-Dimensional Linear Regression

We consider a linear regression framework with n observations and m variables, potentially high-dimensional ($n < m$). The model is

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I) \tag{1}$$

where $\mathcal{N}_n$ denotes the multivariate normal distribution, and $I \in \mathbb{R}^{n\times n}$ is the identity matrix. Here $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times m}$ is a fixed design matrix,

$\beta \in \mathbb{R}^m$ is the vector of coefficients and $\varepsilon \in \mathbb{R}^n$ is a random error vector. We assume that X has rank m, and $X^\top X/n$ converges to a finite positive semi-definite matrix as $n \to \infty$.

We are interested in exploring which variables in X are active, meaning that they have non-null coefficients. Let $M = \{1, \ldots, m\}$ be the set of variable indices, and $N = \{j \in M : \beta_j = 0\}$ the unknown subset corresponding to inactive variables. For any $j \in M$, we may define the null hypothesis $H_j : \beta_j = 0$, that is true when $j \in N$, regardless of the value of other variables' coefficients. We want to study more variables taken together, i.e., test intersection hypotheses of the form

$$H_S = \bigcap_{j \in S} H_j : \beta_j = 0 \text{ for all } j \in S \qquad (S \subseteq M,\ S \neq \emptyset)$$

with significance level $\alpha \in [0, 1)$. H_S is true if all variables in S are inactive ($S \subseteq N$).

To test any H_S, we will rely on a variable selection procedure that estimates the set of active variables with $A \subseteq M$. As in [4], we assume the following properties.

Sparsity $\quad |A| \leq n/2$.
Screening property $\quad \lim_{n\to\infty} P(M \setminus N \subseteq A) = 1$.

The ideal selection procedure, for which the screening property always holds, is an oracle method that selects all truly active variables, plus eventually some others. Even though this is not available in practice, it can be used in simulations to show the performance of the proposed method when the assumptions are ensured. When studying real data, we suggest using the Lasso [5] with a suitable calibration of the λ parameter, so that it selects enough variables for the screening property to be likely. If m_1 is an estimate of the expected number of active variables, we recommend choosing λ so that the Lasso selects $\min(2m_1, n/2)$ variables.

3 Resampling-Based Multisplit

We propose an asymptotically exact test for any intersection hypothesis H_S corresponding to a non-empty set $S = \{j_1, \ldots, j_s\} \subseteq M$. The method will efficiently construct permutation test statistics for H_S by combining statistics for the individual hypotheses H_j with $j \in S$. We take as combining function any $g : \mathbb{R}^s \longrightarrow \mathbb{R}$ which is increasing in each argument, such as the maximum or (weighted) sums.

To define permutation test statistics for all individual hypotheses H_j, we use Q random splits of the data and B random sign-flipping transformations. The values of Q and B do not need to grow with m; larger values of B tend to give more power, but to have non-zero power we only need $B \geq 1/\alpha$ [2]. Hence fix B diagonal sign-flipping matrices $F_1, \ldots, F_B \in \mathbb{R}^{n\times n}$, where $F_1 = I$ is the identity, while the diagonal elements of the other matrices are independently and uniformly drawn from $\{-1, 1\}$. As in [4], for each split q we randomly partition observations into two equally-sized subsets $\mathscr{D}_0^q$ and $\mathscr{D}^q$, and we use $\mathscr{D}_0^q$ to estimate the set of active vari-

ables with $A^q \subseteq M$. Then we use $\mathscr{D}^q$ and A^q to compute test statistics similarly to [3], as follows.

For each split q, we restrict the design matrix X to observations in $\mathscr{D}^q$ and variables in A^q, obtaining $X^q = X_{\mathscr{D}^q, A^q}$. For each selected variable $j \in A^q$, we define X^q_{-j} as X^q without the j-th column, then we construct the split's residual maker matrix

$$R^q_{-j} = 0 \in \mathbb{R}^{n\times n} \qquad \text{except} \qquad R^q_{-j;\mathscr{D}^q,\mathscr{D}^q} = I - X^q_{-j}(X^{q\top}_{-j}X_{-j})^{-1}X^{q\top}_{-j}$$

where all elements are zero except those corresponding to observations in $\mathscr{D}^q$.

We give statistics for all variables by aggregating information over the Q splits:

$$C_{j,b} = \sum_{q:\, j\in A^q} R^q_{-j} F_b R^q_{-j} \in \mathbb{R}^{n\times n}$$

$$T^b_j = \begin{cases} 0 \text{ if } \|X_j^\top C_{j,b}\| = 0 \\ \|X_j^\top C_{j,b}\|^{-1} |X_j^\top C_{j,b} Y| \text{ otherwise} \end{cases} \qquad (j \in M,\, b \in \{1,\dots,B\})$$

where X_j is the j-th column of X.

To test H_S, it is sufficient to combine the individual statistics T^b_j as

$$T^b_S = g\left(T^b_{j_1}, \dots, T^b_{j_s}\right) \qquad (b \in \{1,\dots,B\}).$$

As a critical value we take $T_S^{(\lceil (1-\alpha)B \rceil)}$, where $T_S^{(1)} \le \dots \le T_S^{(B)}$ are the sorted statistics, and $\lceil \cdot \rceil$ denotes the ceiling function.

Theorem 1. *The test that rejects H_S when $T^1_S > T_S^{(\lceil (1-\alpha)B \rceil)}$ is asymptotically an α-level test.*

Proof. Assume that H_S is true, and consider any couple of variables $j,h \in S$, any transformation b, and any split q. Suppose that the variable selection procedure selects all active variables; by the screening property, this is true at least asymptotically, so this assumption does not affect asymptotic results. As H_j is true and A^q contains all active variables, we can write

$$Y_{\mathscr{D}^q} = X_{-j;\mathscr{D}^q,A^q}\beta_{-j;A^q} + \varepsilon_{\mathscr{D}^q}, \qquad \varepsilon_{\mathscr{D}^q} \sim \mathscr{N}_{n/2}(0, \sigma^2 I).$$

For any selected variable $j \in A^q$, the matrix R^q_{-j} has non-null elements only corresponding to observations in $\mathscr{D}^q$, and so the effective score for this model is

$$V^{qb}_j = \frac{1}{\sqrt{n}} X_j^\top R^q_{-j} F_b R^q_{-j} Y = V^{*qb}_j + o_P(1), \qquad V^{*qb}_j = \frac{1}{\sqrt{n}} X_j^\top R^q_{-j} F_b \varepsilon$$

(see Theorem 2 in [3]). Hence the sB-dimensional vectors

$$\mathbf{V}_S = (V^1_{j_1}, \dots, V^B_{j_1}, \dots, V^1_{j_s}, \dots, V^B_{j_s})^\top, \qquad V^b_j = \sum_{q:\, j\in A^q} V^{qb}_j$$

$$\mathbf{V}^*_S = (V^{*1}_{j_1}, \dots, V^{*B}_{j_1}, \dots, V^{*1}_{j_s}, \dots, V^{*B}_{j_s})^\top, \qquad V^{*b}_j = \sum_{q:\, j\in A^q} V^{*qb}_j$$

are asymptotically equivalent. Notice that T_j^b is the standardization of $|V_j^b|$. Since

$$\mathbf{V}_S^* \xrightarrow[n\to\infty]{\mathrm{d}} Z \sim \mathcal{N}_{sB}(0, \Xi \otimes I)$$

$$I \in \mathbb{R}^{B\times B}, \qquad \Xi = (\xi_{k\ell}) \in \mathbb{R}^{s\times s}, \qquad \xi_{k\ell} = \sigma^2 \lim_{n\to\infty} \frac{1}{n} X_{j_k}^\top C_{j_k,1} C_{j_\ell,1} X_{j_\ell}$$

the B vectors $(V_{j_1}^1,\ldots,V_{j_s}^1),\ldots,(V_{j_1}^B,\ldots,V_{j_s}^B)$ converge to i.i.d. random vectors, and so do the vectors $(T_{j_1}^1,\ldots,T_{j_s}^1),\ldots,(T_{j_1}^B,\ldots,T_{j_s}^B)$. Therefore the combinations of their elements $T_S^1,\ldots,T_S^B$ converge to i.i.d. random variables. Moreover, high values of T_S^1 correspond to evidence against H_S. From Lemma 1 in [3],

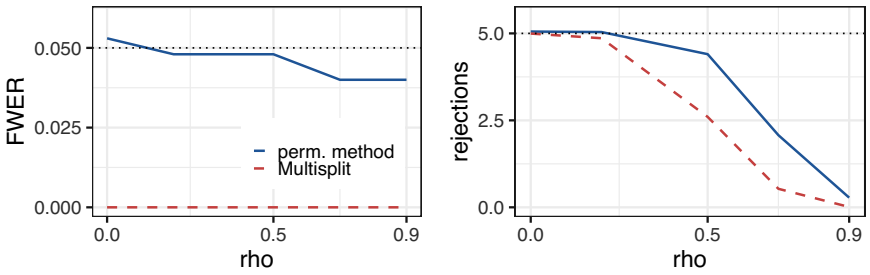
$$\lim_{n\to\infty} P\left(T_S^1 > T_S^{(\lceil(1-\alpha)B\rceil)}\right) = \frac{\lfloor \alpha B \rfloor}{B} \leq \alpha.$$

□

To summarize, we have constructed permutation test statistics for all variables in high-dimensional regression, that are sufficient to define an asymptotically exact test for any intersection hypothesis H_S. The test is obtained combining the statistics for variables in S through any function g that is increasing in each argument.

4 Simulations

We compare the proposed method with the Multisplit, using simulation settings similar to those in [4], with $n = m = 100$. We simulate X from a centered multivariate normal distribution with $\mathrm{cov}(X_j, X_h) = \rho^{|j-h|}$ for $j,h \in M$, taking $\rho \in \{0, 0.2, 0.5, 0.7, 0.9\}$. We compute Y as in (1), where $\beta = (1,\ldots,1,0,\ldots,0)$ has 5 non-null elements, and σ is such that the signal-to-noise ratio is 4. We take $\alpha = 0.05$,



(a) FWER. The dotted line corresponds to the significance level α.

(b) Number of rejections. The dotted line denotes the number of active variables.

Fig. 1: Results by covariance parameter ρ.

$Q = 50$, $B = 200$, and an oracle selection that returns 10 variables. We simulate data 1000 times and study the set of all variables. For the proposed method, we correct for multiplicity with the maxT-method [8], corresponding to $g = \max$.

Results are shown in Fig. 1. Both methods control the FWER, computed as the proportion of simulations where at least one true null hypothesis is rejected. In terms of rejections, the proposed method is equivalent to the Multisplit when the covariance parameter ρ is small, and more powerful in all the other scenarios.

5 Discussion

We have considered the problem of testing multiple hypotheses in high-dimensional linear regression. Our proposed approach provides asymptotically valid resampling-based tests for any subset of hypotheses, which can be employed within closed testing procedures to make confidence statements on the number of active predictor variables (TDP) within any set. These confidence statements are valid even when the subsets of interest are chosen post hoc, after seeing the data.

First, we have provided a procedure that repeatedly splits the data into two random subsets, using the first to select variables and the second to build permutation test statistics for each variable. Then we have shown that statistics for any intersection hypothesis can be defined by aggregating individual statistics with any function which is increasing in each argument, including the maximum and weighted sums. Our method is extremely flexible, allowing different selection procedures and combining functions. Preliminary simulations show a considerable increase in power over the Multisplit [4]. An implementation of the procedure is available in [7].

References

1. Goeman, J.J., Solari, A.: Multiple testing for exploratory research. Stat. Sci. **26**(4), 584–597 (2011)
2. Hemerik, J., Goeman, J.J.: False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. J. R. Stat. Soc. Series B Stat. Metodol. **80**(1), 137–155 (2018)
3. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. R. Stat. Soc. Series B Stat. Metodol. **82**(3), 841–864 (2020)
4. Meinshausen, N., Meier, L., Bühlmann, P.: p-values for high-dimensional regression. J. Am. Stat. Assoc. **104**(488), 1671–1681 (2009)
5. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Series B Stat. Metodol. **58**(1), 267–288 (1996)
6. Vesely, A., Finos, L., Goeman, J.J.: Permutation-based true discovery guarantee by sum tests. Pre-print arXiv:2102.11759 (2021)
7. Vesely, A.: splitFlip: Resampling-based Multisplit. https://github.com/annavesely/splitFlip. R package version 1.1.0 (2021)
8. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York (1993)

Quantile regression coefficient modeling for counts to evaluate the productivity of university students

Modellazione dei coefficienti di regressione quantile per dati di conteggio per valutare la produttività degli studenti universitari

Viviana Carcaiso and Leonardo Grilli

Abstract The extension of quantile regression to count data raises several issues. In this research we compare a solution which exploits jittering to obtain a continuous working variable to a more recent approach, in which the coefficients of quantile regression are modeled by parametric functions. Both methods are applied for evaluating the effect of remote teaching on university students' productivity. In this context, the latter approach is found to be advantageous.

Abstract *L'estensione della regressione quantile ai dati di conteggio provoca diversi problemi. In questa ricerca confrontiamo una soluzione che sfrutta il jittering per ottenere una variabile continua con cui lavorare con un approccio più recente, in cui i coefficienti di regressione quantile sono modellati da funzioni parametriche. Entrambi i metodi sono applicati per valutare l'effetto della didattica a distanza sulla produttività degli studenti universitari. In questo contesto, l'approccio più recente risulta vantaggioso.*

Key words: COVID-19, jittering, parametric modeling, quantiles, university credits

1 Introduction

Quantile regression has the main advantage of being completely distribution-free and avoiding some of the restrictive assumptions of conventional regression: it does not require homoscedasticity or a specific type of distribution for the response and it

Viviana Carcaiso
Department of Statistical Sciences, University of Padova, Padova, Italy. e-mail: viviana.carcaiso@phd.unipd.it

Leonardo Grilli
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy. e-mail: leonardo.grilli@unifi.it

is very flexible since regression equations at different quantiles are fitted separately. Most of its theoretical developments and empirical applications concern continuous outcomes. Indeed, the response variable is generally assumed to be sampled from an absolutely continuous population, which is not true if the outcome is a count. The extension of quantile regression to count data raises several issues, mainly related to the fact that the conditional quantile function of a discrete random variable cannot be a continuous function of the regression parameters [6]. The solution suggested by Machado and Santos Silva (2005) [6] is to generate an artificial continuous variable by adding a uniform random variable to the original counts, a procedure referred to as jittering. This allows to obtain a continuous working variable whose quantiles have a one-to-one relation with those of the response (see [4] for an application to the productivity of university students). A recent alternative approach considered in this research is the one by Frumento and Salvati (2021) [3], who suggest applying the quantile regression coefficients modeling (QRCM) paradigm described in [2] to model a discrete response. As Machado and Santos Silva, they want to impose some degree of smoothing to the assumed distribution, without altering the response itself. Their solution permits to avoid jittering and was shown to provide numerous advantages, including parsimony, efficiency and simplicity in terms of interpretation. Here, we apply both methods to evaluate the effect of remote teaching due to COVID-19 pandemic on university students' productivity of freshmen from the academic year 2019/2020 of the University of Florence. The rest of the paper is organised as follows. Section 2 outlines the theory of the two methods, Section 3 illustrates the case study and the applied statistical models, and in Section 4 we discuss the obtained results.

2 Methods

In 2005, Machado and Santos Silva [6] proposed analyzing count data using quantile regression, which permits avoiding strong parametric assumptions and enables investigating every aspect of the conditional distribution of the response variable, not just its mean. However, there are some complications, related to the non-smoothness of the objective function combined with the discreteness of the response variable. The solution of Machado and Santos Silva [6], that will be referred to as QRCJ (Quantile Regression for Counts based on Jittering), is based on the artificial smoothing of the data using jittering.

More specifically, after defining $Z = Y + U$, where U is a uniform random variable independent of Y and $\boldsymbol{x}$, and noting that $Q_Z(p|\boldsymbol{x})$ is bounded from below by p, the authors specified a model of the form $Q_Z(p|\boldsymbol{x}) = p + \exp(\boldsymbol{x}'\boldsymbol{\beta}(p))$. In other words, they propose to use a monotone transformation of the conditional quantile function of Z, namely $T(Z;p) = \log(Z-p)I(Z > p) + \log(\zeta)I(Z \leq p)$, where ζ is an arbitrary small positive number, and then $\boldsymbol{\beta}(p)$ can be estimated by running a linear quantile regression of $T(Z;p)$ on $\boldsymbol{x}$. The value of $\hat{\boldsymbol{\beta}}(p)$ depends not only on the sample information but also on the specific realization of U, thus the authors

propose to generate m "jittered" samples and to average the estimates (they call this procedure "average-jittering"). The average jittering estimator is proved to be more efficient than the one based on a single sample and, under mild conditions on the covariates, to be consistent and asymptotically normal, therefore standard asymptotic theory can be used to perform inference. The simulation study in [6] shows that the proposed estimator has good properties in finite samples of size $n = 500$.

The fact that quantiles are estimated one at a time and that no parametric structure is assigned to the coefficient functions has relevant drawbacks [3]. The quantile regression coefficients modeling (QRCM) approach, first developed by [2], and then applied to count data by [3], consists in describing the regression coefficients by smooth parametric functions with closed-form mathematical expressions and should provide a gain in terms of efficiency and simplicity of interpretation. In the general QRCM framework presented by [2] the q regression parameters are defined as $\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$, where $\boldsymbol{b}$ is a vector of k known functions of p that are assumed to be continuous and differentiable, and $\boldsymbol{\theta}$ is a $q \times k$ matrix. In the specific case of quantile regression for count data, the model proposed by Frumento and Salvati is driven by the empirical evidence that the estimators obtained by applying QRCM to the jittered response, $Z = Y + U$, and directly to $Y^{\circ} = Y + E[U]$ are almost identical, due to imposed parametric structure [3]. Therefore, after assuming, without loss of generality, $E[U] = 0.5$, the model is applied to a transformation $T(\cdot)$ of $Y^{\circ} = Y + 0.5$, that is

$$Q_{T(Y^{\circ})}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\theta}\boldsymbol{b}(p).$$

The estimation is carried out by minimizing the integrated objective function, which is the integral, with respect to the order of the quantile, of the loss function of standard quantile regression. This estimation approach allows to estimate the entire quantile process instead of a discrete set of quantiles. Furthermore, the integrated loss function is a smooth function of its arguments, and this allows to carry out minimization using standard algorithms, like Newton–Raphson or gradient search, and to apply the standard theory of M-estimators to investigate the asymptotic behaviour of the estimator. Fitting linear quantiles is the most common practice because of the simplicity of the interpretation [2], but in [3] it is also proposed a logarithmic transformation.

The idea is that the imposed parametric structure should yield some gain in terms of efficiency, as shown by the simulation results reported by the authors in [3]. However, specifying a "good" parametric model for a quantile function it is not easy, since there are many different possible specifications. The model is determined by the choice of $\boldsymbol{b}(p)$, which must be defined in advance, and by the restrictions that are imposed on $\boldsymbol{\theta}$. In practice, the most common choices are polynomials, splines, piecewise linear functions, roots, logarithms, trigonometric functions, quantile functions of known distribution (e.g., that of a Normal, Beta or Gamma distribution), and combinations of the above. In many situations, some of the quantile regression coefficients can be assumed to be linear functions of p, or not to depend on p at all. [2] propose a goodness-of-fit test procedure that is based on the idea of assessing the

model fit by comparing the distribution of $F(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})$, $i = 1,\ldots,n$, with a $U(0, 1)$ distribution, but it just represents an approximation in the framework of count data.

3 Application

The evaluation of the impact of the change to remote teaching on university students' careers represents an interesting application of quantile regression for count data. The data were collected in the administrative records and concern first-year students who enrolled in the Bachelor's degree courses of Psychological Sciences and Industrial Design in the academic years 2018/2019 and 2019/2020 at the University of Florence. Two different degree courses are selected to have an insight in different fields of study. The idea is to compare the productivity of freshmen in the second semester of 2019/2020, which is the one affected by online teaching, to that of first-year students in the second semester of 2018/2019, who received the usual frontal lectures. Both cohorts attended regular lectures in the first semester, but were exposed to different forms of teaching in the second one. In both academic years of interest the study plan remained the same. After excluding students who got no credits in the first semester, the data set consist of 946 first-year students, whose characteristics are displayed in Table 1. We can notice that within each degree course the two cohorts have similar characteristics.

Table 1 Summary of background characteristics of freshmen by degree course and year of enrollment (2018 or 2019), University of Florence.

	Psychological Sciences			Industrial Design		
	2018	2019	Total	2018	2019	Total
Nr. observations	313	336	649	139	158	297
Gender (%)						
Female	76.7	82.1	79.5	72.7	62.0	67.0
Male	23.3	17.2	20.5	27.3	34.0	33.0
Type of HS (%)						
Scientific	37.7	32.7	35.1	32.4	25.3	28.6
Humanities	16.3	17.0	16.6	5.04	6.96	6.06
Language	7.35	7.14	7.24	2.88	3.80	3.37
Human Sciences	24.6	19.9	22.2	12.2	9.49	10.8
Art School	1.28	2.98	2.16	22.3	29.1	25.9
Technical	9.27	15.2	12.3	19.4	18.4	18.9
Other	3.51	5.06	4.31	5.76	6.97	6.40
HS grade						
Average	82.1	80.3	81.17	76.9	78.3	77.63
Std. error	10.9	11.1	11.00	10.1	11.1	10.79

Students' productivity is measured by the number of gained credits (ECTS). According to the study plan, students should obtain approximately 30 credits per semester. Since the number of credits is always a multiple of 3, the response variable used in the models, denoted by Y, is defined as the number of credits gained in the second semester divided by 3. Since Y presents an irregular distribution in

both degree courses, quantile regression represents an appealing methodology. The covariates included in the models are:

- X_1: number of credits obtained during the first semester, centred around 15.
- X_2: dummy variable for cohort 2019 vs 2018.
- X_3: dummy variable for male vs female.
- X_4: high school grade, centred around 80.
- $X_5, ..., X_{10}$: dummy variables for the types of high school (except for the baseline category "Scientific").

The analysis is performed separately for each bachelor's degree course and the effect of interest is represented by the estimate of the coefficient of X_2, since this is the variable that distinguishes the students whose second semester was affected by remote teaching (cohort 2019) from the others (cohort 2018). Both the QRCJ and QRCM approaches are exploited and, more specifically, the applied models are

$$\text{QRCJ}: \quad Q_Z(p|\boldsymbol{x}) = p + \boldsymbol{x}'\boldsymbol{\beta}(p) = p + \beta_0(p) + \sum_{i=1}^{10} \beta_i(p) x_i,$$

$$\text{QRCM}: \quad Q_{Y^\circ}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \beta_0(p|\boldsymbol{\theta}) + \sum_{i=1}^{10} \beta_i(p|\boldsymbol{\theta}) x_i.$$

In both of them the transformation of the working variable is the identity function: in the QRCJ approach the logarithmic function is not applied, in order to ease the comparison with QRCM. The methods were computationally implemented by using the function `rq` of the R package `quantreg` [5] for QRCJ, and the function `iqr` of the `qrcm` package [1] for QRCM.

For both degree programs, a variety of parametric models were applied to the conditional quantile function of $Y^\circ = Y + 0.5$ and they were compared based on the number of free model parameters, the integrated loss function, the p-values of the Kolmogorov–Smirnov (KS) test for goodness-of-fit and the similarity of the estimates of the constant and $\beta_2(p|\boldsymbol{\theta})$ to the ones obtained from QRCJ. In both cases, in the selected model the intercept and the coefficient which represents the effect of interest, $\beta_2(p|\boldsymbol{\theta})$, are modelled in a more complicated and flexible way. In particular, they are formulated as combinations of shifted Legendre polynomials and logarithmic functions. The other coefficients are instead approximated by constant (for the ones related to the type of high school) and linear functions.

4 Discussion and final remarks

Quantile regression represents an appealing methodology for dealing with an outcome variable characterized by an irregular distribution, as in the case of the number of credits gained by university students. In this research two different methods for applying quantile regression to count data were employed: the jittering approach

of [6] and the parametric modelling one of [3]. The obtained results show that they lead to similar estimates of regression coefficients, especially for the variables which were modelled through flexible parametric functions (combinations of polynomial and logarithmic functions) in the QRCM setting. The QRCM approach in most cases entails a gain in efficiency, especially in the tails of the distribution of the response. This gain is generally more substantial for the control variables that are approximated by constant functions, whereas is lower or not present for the coefficients described by more flexible parametrizations. The parametric modeling approach has the advantage of providing estimates of the regression coefficients that are smooth functions of p, and makes the interpretation more straightforward. Moreover, it allows to estimate multiple quantiles at the same time, resulting in a much faster computation. However, model selection is not easy and requires time and expertise. The suggested goodness-of-fit procedure is not completely appropriate in the case of count data, and the integrated loss function is not very useful since it depends on the number of parameters. In our case study, the number of possible models was reduced by the fact that there is a specific coefficient of interest, but it was still quite large. A combination of QRCJ and QRCM was found to be a good method for model selection. Indeed, the estimates obtained with the jittering method can be used as a benchmark when choosing among a set of parametric specifications, in order to derive a satisfactory functional form for some specific covariates, keeping always under control the value of the loss function and the goodness of fit.

As far as the interpretation is concerned, the effect of remote teaching on Psychological Sciences freshmen is negative, but small, at most quantiles. It is stronger on the tails, namely for students with low or high productivity (but not for the ones with the greatest performance), however all the estimates are not significantly different from 0 (at a level of 5%). For freshmen in Industrial Design the effect is instead positive, but again small, and stronger for the group of the most productive students. Indeed, the QRCM estimates do not reach statistical significance at 5% except at a few quantiles around 0.95. In both courses the data size does not seem large enough for identifying such small effects.

References

1. Frumento, P. (2021). `qrcm`: quantile regression coefficients modeling. R package version 3.0. http://CRAN.R-project.org/package=qrcm.
2. Frumento, P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics*, 72: 74-84.
3. Frumento, P. and Salvati, N. (2021). Parametric modeling of quantile regression coefficient functions with count data. *Stat Methods Appt*.
4. Grilli, L., Rampichini, C., Varriale, R. (2016). Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: an approach based on quantile regression for counts. *Stat. Modelling* 16(1):47–66.
5. Koenker, R. (2021). `quantreg`: Quantile Regression. R package version 5.85. https://CRAN.R-project.org/package=quantreg.
6. Machado, J. A. F. and Santos Silva, J. M. C. (2005). Quantiles for counts, *J. Am. Stat. Assoc.*, 100, 1226–1237.

Adaptive smoothing spline using non-convex penalties

Spline con smoothing adattivo utilizzando penalizzazioni non convesse

Daniele Cuntrera, Vito M.R. Muggeo

Abstract We propose a new adaptive penalty for smoothing via penalized splines. The new form of adaptive penalization is based on penalizing the differences of the coefficients of adjacent bases, using non-convex penalties. This makes possible to estimate curves with varying amounts of smoothness. Comparisons with respect to some competitors are presented.

Abstract *Proponiamo una nuova penalizzazione adattiva per le spline lisciate. La nuova forma di penalizzazione adattiva si basa sulla penalizzazione delle differenze dei coefficienti delle basi adiacenti, utilizzando forme non convesse di penalizzazione. Questo rende possibile la stima di curve il cui grado di lisciatura è determinato in modo adattivo. In questo lavoro vengono presentati alcuni confronti.*

Key words: regression, B-spline, P-spline, adaptive smoothing, non-convex penalties, SCAD, MCP

1 Introduction

The use of splines in statistical modelling has undergone a significant increase in the analyses conducted: in just five years, the number of packages using splines has tripled (Perperoglou et al., 2019). The success lies in their flexibility in different settings, such as modelling non-linear effects in regression models and estimating baseline risk in survival models). However, when the underlying relationship to be estimated shows particular and somewhat complex shapes, the amount of smoothing

Daniele Cuntrera
Ph.D Student, Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche,
e-mail: daniele.cuntrera@unipa.it

Vito M.R. Muggeo
Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche,
e-mail: vito.muggeo@unipa.it

should not be kept constant. The resulting non-constant or varying smoothing is referred as adaptive smoothing. To date, there exist just few packages dealing with adaptive smoothing. This contribution is organized as follows: in the second section, we briefly present the rationale of B-splines and P-splines, and then we present our proposal in some detail. Some simulations and conclusions are reported in the last section.

2 Splines, B-splines and P-splines

Roughly speaking, a spline is a function consisting of a set of polynomials joined together, which interpolate a set of points called nodes. The function is continuous at least up to the specified order of derivatives depending on the degree of the piecewise polynomials. A set of properly defined spline functions represent the *B*-spline (Eilers and Marx, 1996, 2010), namely $B(x) = [B_1(x), B_2(x), \ldots, B_K(x)]$ where K is the number of bases depending on degree polynomial and number knots.

Any unspecified but smooth function $f(x)$ can be expressed via the B-spline basis. For the generic value x_i we write

$$f(x_i) = \sum_{k=1}^{K} b_k B_k(x_i) = B(x_i)^T b. \tag{1}$$

where the coefficients b_ks have to be estimated by data. For the response values $\{y_i\}_{1,\ldots,n}$ the objective to be minimized is the well known least square $S(b) = \sum_i (y_i - B(x_i)^T b)^2$ which yields K regression coefficient estimates. To by pass the issues related to the selection of the basis dimension K, namely in turn the number and location of knots, a generous basis with several knots is built up and a penalty is set on the dth coefficient differences $\Delta^d b$; for instance if $d = 2$ is chosen $\Delta^2 b_k = (b_k - b_{k-1}) - (b_{k-1} - b_{k-2})$. This leads to so-called penalized splines or simply *P*-splines [ref].

Therefore, the penalized least square to be optimized is

$$S_\lambda(b) = \sum_{i=1}^{n} (y_i - B(x_i)^T b)^2 + \lambda \sum_k (\Delta^2 b_k)^2, \tag{2}$$

where λ is the tuning parameter determining how the curve has to be penalized. The larger λ, the smoother the fitted curve, and the optimal value is usually selected by cross-validation. However, regardless of its value, a unique and constant λ implies that the amount of smoothing does not change across the covariate range. Unfortunately, in some particular cases, such *constant* smoothing leads to undesirable fits as illustrated in Figure 3 where the grey line portrays fitted curve corresponding to the unique λ selected by minimizing the BIC.

3 The proposal: using non-convex penalties

The idea proposed in this paper is to adopt non-convex penalties for P-Spline estimation adaptively and thus to have a λ that varies along the x-domain. In this way, it is possible to obtain a curve that modifies the jaggedness according to the dispersion of the points of x, allowing to have a more or less smooth curve according to the needs dictated by the data. In our context, we penalize for the differences between the coefficients estimated for each base for orders of difference greater than or equal to 2. So the minimize function becomes

$$S = \sum_{i=1}^{n} (y_i - B(x_i)^T b)^2 + \sum_k \lambda_k \, p_\lambda(|\Delta_k^d b|). \tag{3}$$

The term Δ_k^d denotes the j-th difference of order d that is penalized. The term $p_\lambda(\cdot)$ thus indicates one of the two non-convex penalties taken into account, namely SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). These are defined as

$$p_{\lambda_{SCAD}}(|\Delta_k^d|) = \begin{cases} \lambda|\Delta^d| & \text{if } |\Delta^d| \leq \lambda, \\ \frac{2\gamma\lambda|\Delta^d| - \Delta^{D^2} - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |\Delta^d| < \lambda\gamma, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{otherwise} \end{cases}$$

$$p_{\lambda_{MCP}}(|\Delta_k^d|) = \begin{cases} \lambda|\Delta^d| - \frac{\Delta^{D^2}}{2\gamma} & \text{if } |\Delta^d| \leq \lambda\gamma, \\ \frac{1}{2}\gamma\lambda & \text{otherwise} \end{cases}$$

where γ is the threshold that defines the shape of SCAD and MCP penalties. We will set $\gamma = 3.7$. The advantage of using these penalties is that the estimated parameters are less biased than lasso. In fact, for both, after the threshold value $\lambda\gamma$, the penalty does not increase, unlike the lasso. On the other hand, these penalties have unavoidable estimation difficulties due to their non-concave form for the parameters. This leads to substantial difficulties in maximizing the penalized likelihood function. Fan and Li (2001) propose a local quadratic approximation, which allows to write the penalty function as

$$p_\lambda(|\Delta_k^d|) \approx p_\lambda(|\Delta_k^{d,(0)}|) + \frac{p'_\lambda(|\Delta_k^{d,(0)}|)}{2|\Delta_k^{d,(0)}|}(\Delta_k^{d^2} - \Delta_k^{d,(0)^2}) \tag{4}$$

where $\Delta^{d,(0)}$ is the point in which the linear quadratic approximation is centred. The same linear quadratic approximation will also be used for the MCP penalty.

The definition of λ_k takes advantage of the lower bias of SCAD and MCP for coefficients with high estimates. We then decompose λ_k as λw_k, where

$$w_k = p'_\lambda(|\Delta_k^d|; \lambda, \gamma). \tag{5}$$

High coefficient differences will have a smaller penalty, so the estimated curve will be less smoothed and will be able to follow the data in a better way; on the other hand, similar coefficients will be penalized more, so a smoother part of the curve will be obtained. In this way, we allow the penalty to adjust according to the dispersion of the data. Defined the number of bases, the degree of polynomials, the order of differences, and the penalty to be used (SCAD or MCP), the estimate is made as follow:

1. Estimation of the non-penalized model, to initialize the algorithm
2. Calculation of d-order differences and weights
3. Construction of augmented matrices and estimation of the penalized model
4. Update of estimated b_j
5. Repeat steps 2 to 4 till convergence

For the optimal λ search, we proceed with a grid search to minimize the BIC.

4 Simulation study

In this section we show the simulation study conducted to assess the performance of our proposals empirically. These will be compared with B-splines, P-splines and spatially-adaptive P-splines (SOP) (Rodríguez-Álvarez et al., 2019). We present two different analyses: in the first one we consider as signal the Doppler function $\mu = \sqrt{x(1-x)}\sin\frac{2\pi(1+2^{(9-4*6)/5})}{x+2^{(9-4*6)/5}}$, while in the second one we define the signal as $\mu = cos(x) + cos(\pi x)$.

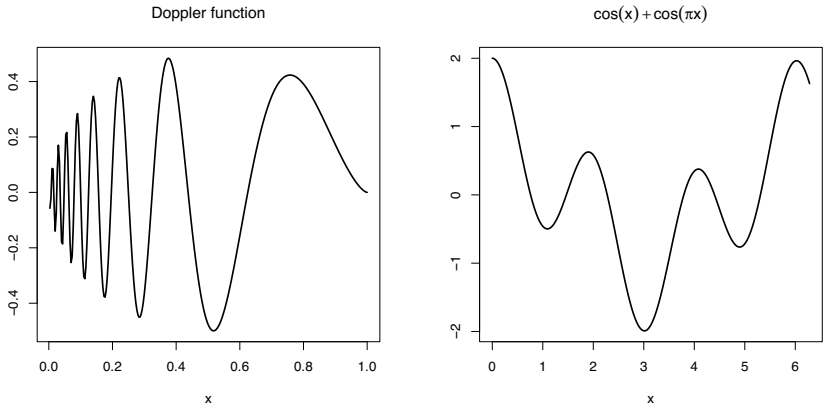


Fig. 1 The two signals used for simulation study

The Doppler signal is a typical example in the literature on adaptive smoothing. In contrast, the second signal is useful for assessing whether, in a context that does not require adaptive smoothing, the method has advantages over b-spline and p-spline. Once the signals have been determined, we then define

$$y_i = \mu_i + \sigma_j \varepsilon_i \quad , \quad i = 1,...,n; \quad j = 1,...,4$$

where x were equally spaced on [0,1], $\varepsilon_i \sim N(0,1)$, $n = 300$. We also consider different values for σ_j to assess how performance varies as noise increases. For all setting, we carried out 500 simulations. To measure the behaviour of the five methods tested, we define the measure $MISE = \sum_{i=1}^{n}(\mu(x_i) - \hat{\mu}(x_i))^2$.

For all 5 methods, 50 bases were used, with polynomials of order 3 and penalizing differences of order 3.

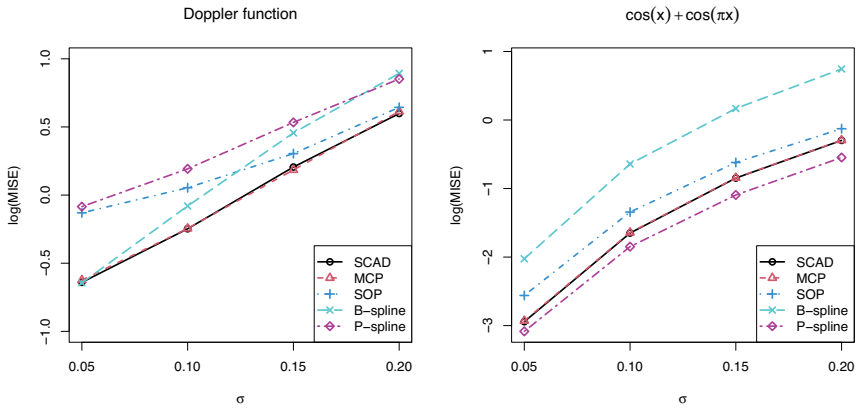


Fig. 2 log(MISE) for different method and σ

Figure 2 shows the results of the simulations, considering the log(MISE) (for graphical purpose). In the scenario using the Doppler function, it is observed that, for the proposed method, there is no substantial difference in the use between the SCAD or MCP penalties. Compared to the other methods, better performance is observed for all the σ values used. Compared to the SOP (the only adaptive alternative), better results are observed, which tend to converge for higher σ values. The scenario changes in the second simulation setting: when an adapting λ is not required from the signal, the best results are obtained using P-splines. The adaptive penalizations with SCAD and MCP return almost equal results, always better than those obtained with SOP.

Figure 3 shows a generic curve estimated by adaptive (SCAD) and non-adaptive method. From the graph, it is possible to observe how the adaptive method estimates a curve much closer to the true signal, since the adaptivity of λ makes it possible to obtain a smoother curve on the right-hand less smooth on the left-hand side.

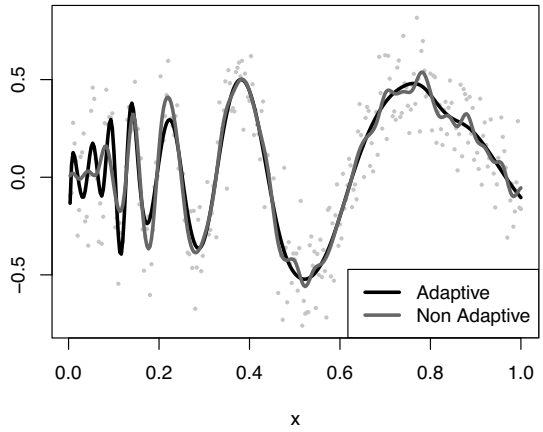


Fig. 3 Adaptive (SCAD) and non-adaptive estimated curve, using a Doppler signal with σ=0.15

This paper presents a new proposal to fit adaptive P-spline using the SCAD and MCP-based penalties. In the presented simulation experiments, the proposed method had good performance with respect to the traditional, constant smoothing, and the adaptive SOP competitor.

References

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121.

Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in R. *BMC medical research methodology*, 19(1):1–16.

Rodríguez-Álvarez, M. X., Durban, M., Lee, D.-J., and Eilers, P. H. (2019). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statistics and Computing*, 29(3):483–500.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Conditional tests for generalized linear models

Test condizionali per modelli lineari generalizzati

Riccardo De Santis, Jelle J. Goeman, Anna Vesely, Livio Finos

Abstract The relevance of any possibly significant covariate when using a generalized linear model should be further investigate by means of an appropriate statistical test. The most known parametric approach might fail when some assumptions are missing, losing the control of type I error. Permutation tests can often be useful in these situations, requiring fewer assumptions, and we will apply them in this general framework.

Abstract *Nell'ambito dei modelli lineari generalizzati, la rilevanza di qualsiasi covariata potenzialmente significativa deve essere confermata da una valida statistica test. Il più usato approccio parametrico può fallire quando alcune assunzioni vengono meno, causando una perdita del controllo dell'errore di I tipo. I test di permutazione offrono una valida alternativa in queste situazioni, richiedendo meno assunzioni, e li applicheremo in questo contesto.*

Key words: Generalized linear model, Permutation test, Score test, Sign flipping

Riccardo De Santis
Department of Statistical Sciences, University of Padova, Italy, e-mail: riccardo.desantis.1@phd.unipd.it

Jelle J. Goeman
Biomedical Data Sciences, Leiden University Medical Center, The Netherlands, e-mail: j.j.geoman@lumc.nl

Anna Vesely
Department of Developmental Psychology and Socialization, University of Padova, e-mail: anna.vesely@unipd.it

Livio Finos
Department of Developmental Psychology and Socialization, University of Padova, e-mail: livio.finos@unipd.it

1 Introduction

Generalized linear models (GLMs) are a flexible tool which permits to relate one dependent variable with some potentially significant covariates. They allow to model non-linear functions of the mean and require the dependent variable to belong to the exponential distribution family, which encompasses many of the most common distributions. However, this complexity leads to some drawbacks. It comes harder to properly check the reliability of the assumptions made and it becomes more difficult to build robust testing procedures.

In the ambit of hypothesis testing, permutation methods offer an alternative which requires fewer assumptions than the standard parametric approach. Permutations are encompassed in the general definition of groups of transformations, which might be also rotations or sign flipping contributions [6]. However, these methods are generally limited in the ambit of GLMs since they require the observations to be exchangeable, which prevent us to use a naive permutation approach.

In order to overcome this limit [3] proposed the effective flip score test. This test based on flipping score contributions defines a valid approach to the problem and results to be robust against some forms of overdispersion. Furthermore, the simulation study shows some promising results against generic variance misspecification. However, this method is not fully satisfactory. A slightly anti-conservative behavior remains for small-sample sizes which demands to look for a further step. In this work we answer this open question by proposing an improvement – called Standardized Score test. We will see that the proposed test turns out to provide a better type I Error control.

In Section 2 we will present some notation about generalized linear models, Section 3 includes a discussion of statistical tests available while Section 4 contains a simulation study. Finally, the conclusions are in Section 5.

2 Generalized Linear Models

Assume that we observe n independent observations $Y = \{y_1, \dots, y_n\}$ drawn from a particular distribution which belongs to the exponential dispersion family, whose i-th element has density of the form [1]

$$f(y_i; \theta_i, \phi_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\},$$

where θ_i and ϕ_i are respectively the canonical and the dispersion parameter.

According to the usual literature of GLMs [1] we have

$$\mu_i = E[y_i] = b'(\theta_i) \qquad v(\mu_i) = b''(\theta) = \frac{Var(y_i)}{a(\phi_i)}$$

which are respectively the mean and the variance function. We assume that the mean of y_i depends on some observed covariates $(x_i, \mathbf{z}_i)$ through a non-linear relation

$$g(\mu) = \eta = X\beta + \mathbf{Z}\gamma \tag{1}$$

where $g(\cdot)$ is the link function, $(X, \mathbf{Z})$ is the design matrix with $dim(X) = n \times 1$ and $dim(\mathbf{Z}) = n \times q$.

Let $l(\beta)$ denote the log-likelihood function, from which we can derive the score function with respect to β

$$\frac{\partial l(\beta)}{\partial \beta} = l_*(\beta) = X^T D V^{-1}(y - \mu),$$

where, in a compact matrix form, we have

$$D := diag\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\} \qquad V := diag\{Var(y_i)\}.$$

Furthermore, deriving the score function we can obtain the expected Fisher information

$$E\left[\frac{\partial^2 l(\beta)}{\partial \beta^2}\right] = J(\beta) = X^T W X$$

where $W := DV^{-1}D$. The maximum likelihood estimator does not have a closed form, but it is estimated through a Fisher scoring algorithm, which is an iterative method based on the second-order approximation of the log-likelihood function.

We will focus on univariate hypothesis testing of the form

$$H_0 : \beta = \beta_0 \qquad vs \qquad H_1 : \beta \neq \beta_0 \tag{2}$$

or, equivalently, to one-sided hypothesis testing, while the other parameters γ are treated as nuisance parameters. The difficulties of deriving a valid hypothesis testing are described in the following section.

3 Conditional resampling in GLMs

Type I error control is a crucial property of a statistical test in the ambit of hypothesis testing in order to ensure an adequate analysis. An exact control of the type I error is therefore a desiderable property for any test. Linear models with normal responses admit exact control within the well-known parametric approach. However, in the more general case of GLMs, this property is rarely met, even if the parametric approach to hypothesis testing still ensures accurate control. Indeed, it is based on a second-order Taylor approximation, and hence it focuses on the first and second moment of the distribution of the test statistic. This, together with the Central Limit Theorem, ensures a quick convergence to a common normal distribution, towards an

exact control of the type I error, whose rate of convergence is specific of any particular model. Nevertheless, this parametric approach relies on stringent assumptions, that are usually difficult to be checked properly.

Conditional tests have been developed to answer a general demand of less stringent assumptions. A general structure can be described as follows. Let $\mathbb{F}$ be a group of transformations. Conditional approaches rely on applying this group of transformations (or an appropriate sub-group [2]) to the observed data. For a given test statistic T, the conditional distribution of the test statistic under the null hypothesis is defined by the set $\{T \circ F; \forall F \in \mathbb{F}\}$ (the observed test statistic is usually defined by $F = \mathbf{I}$ – with $\mathbf{I}$ identity matrix). Depending on the way T and $\mathbb{F}$ are defined, the resulting test will reach different inferential properties. In order to have a valid test, the requirement is that they all share the same conditional distribution, *i.e.*, $T \circ F \overset{d}{=} T \circ \mathbf{I}\ \forall F \in \mathbb{F}$. This property is usually met only in linear model [6, 4] or in GLMs with simple experimental designs. However, a general approach is still lacking, and only some *ad hoc* solutions can be found in the literature [5]. The *second-order exactness* provided by GLMs – and mentioned above – has its conditional equivalent in the *second-moment null-invariance* property, that is, the test statistics have mean and variance independent of the flip: $S \circ F \overset{2M}{=} S \circ \mathbf{I}\ , \forall F \in \mathbb{F}$. Under this condition, the test statistics do not have the same distribution, but they have the first two moments in common, providing almost exact control for practical purposes.

Inference in the GLM framework can be made by the use of flipping score contributions. Unfortunately, the use of basic scores is unfeasible in presence of unknown nuisance parameters. Their estimation makes the basic scores no longer exchangeable, since they are no longer sharing the same expected value. This property is retrieved by [3] in an asymptotic framework by proposing the use of contributions based on the effective scores. The flipping score contributions are introduced in form of a diagonal matrix F of dimension equal to n, whose elements are independent random variables which can take values on $\{-1;+1\}$. Consequently we define the effective flipped score for any F as

$$S(F) = n^{-1/2} X^T W^{1/2} (I - H) V^{-1/2} F (y - \hat{\mu}) \tag{3}$$

where $\hat{\mu}$ are the fitted values under the null model, while

$$H := W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}. \tag{4}$$

is a particular matrix, which recovers the concept of a projection matrix in the GLM framework. It is symmetric and idempotent, and it is a proper projection matrix for studentized units as shown by the following relation [1]

$$V^{-1/2}(\hat{\mu} - \mu) = H V^{-1/2}(y - \mu)(1 + o_{\mathbb{P}}(1)) \tag{5}$$

where there is an asymptotically negligible error term.

It is trivial to show that all the flipped effective scores share the same expected value, while it is not true for the variances, which using the results above are equal to

$$Var[S(F)] = n^{-1}X^T W^{1/2}(I-H)F(I-H)F(I-H)W^{1/2}X + o_{\mathbb{P}}(1). \quad (6)$$

As proven in [3] the exactness of the test based on the flipped effective scores is only asymptotic. It turns out to be a weaker version of *second-moment null-invariance*, *i.e.*, an asymptotic one. The relevance of this result is related to the general derivation for any kind of GLMs. The quality of this control is often good-enough, but it has poor results for small sample sizes and demands for a further improvement.

In this work we propose a straightforward but efficient way to get finite *second-moment null-invariance*, that is, finite second-moment exactness. We have called it Standardized Score test where the standardized score is defined as

$$S_S(F) = S(F)/\sqrt{Var(S(F))}, \quad (7)$$

whose advantadges can be easily exploited:

Theorem 3.1 (Standardized Flip Score test) *Standardized Flip Score as defined in (7) is finite second-moment null-invariant. The test is exact in case of normal response with identity link and asymptotically exact in all other cases.*

The result is quite trivial. The expected value is left unchanged from that of the effective score, while the variance of each flipped statistic is somehow standardized and made equal to one. Hence the independence of the first two moments from any fixed flip F.

4 Simulation study

In this simulation we compare the parametric test with the two permutation tests when estimating a logistic regression model. The model is set as in (1) with $dim(\gamma) = 3$, $\gamma = c(1,1,1)$, $\beta = 0$, correlation of the non-relevant variable with the others equal to $(0.5, 0.1, 0.1)$. We carried on 5000 simulations testing (2), setting the sample size equal to 40.

In Figure 1 on the left is reported the type I error control for the most interesting values of α, *i.e.* $\alpha \in (0, 0.10)$. We observe a conservative behavior of the parametric test, while as expected the effective score test is anti-conservative. The standardized score test gives a more satisfying control, which is nearly exact. This simulation shows a clear preference of the new test over that based on the effective score, while it is surprisingly performing better than the parametric approach.

To complete the analysis we perform a comparison of the power of the tests, shown in the right plot of Figure 1. We set $\gamma = (1,1,0), \beta = 1.5$, again with 5000 simulations and sample size equal to 40. The standardized test is still satisfying, having slightly more power than the parametric counterpart for low values of α.

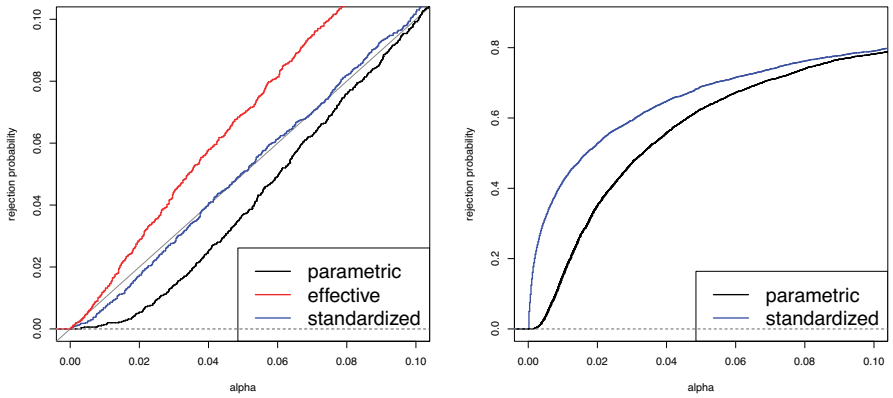


Fig. 1 Left: Type I error control for a logistic regression model. **Right:** Power comparison for a logistic regression model for the two tests with a non anticonservative control of type I error

5 Conclusion

We have seen that permutation testing is a valid alternative to the wide-spread standard parametric approach when conducting statistical tests in the ambit of GLMs. We propose an almost exact test which turns out to be satisfying for small sample sizes and inherits some properties of the effective score test, being competitive to the parametric testing in a well-specified logistic regression model. Further research is needed in order to exploit general robustness with respect to the usual assumptions (as general heteroscedasticity) and regarding the extension to a multivariate setting, following the way introduced by [3].

References

1. A. Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley, New York, 2015.
2. J. Hemerik and J.J. Goeman. Exact testing with random permutations. *TEST*, 27:811–825, 2018.
3. J. Hemerik, J.J. Goeman, and L. Finos. Robust testing in generalized linear models by sign flipping score contributions. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 82(3):841–864, 2020.
4. S. Kherad-Pajouh and O. Renaud. An exact permutation method for testing any effect in balanced and unbalanced fixed effect anova. *Computational Statistics & Data Analysis*, 54(7):1881 – 1893, 2010.
5. F. Pesarin. *Multivariate permutation tests: with applications in biostatistics.* Wiley, Chichester, 2001.
6. A. Solari, L. Finos, and J. J. Goeman. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014.

Methods and applications in economics and finance

Mixed models for anomaly detection in anti-money laundering aggregate reports

Modelli lineari misti per l'individuazione di anomalie nelle Segnalazioni Anti-Riciclaggio Aggregate (S.AR.A.)

Stefano Iezzi[1] and Marianna Siino[2]

Abstract Banks and other financial intermediaries operating in Italy are mandated to report monthly to Italy's Financial Intelligence Unit (UIF) all transactions after aggregating them by branch, customer sector and type of transaction. These non-nominal reports, the so-called Anti-money Laundering Aggregate Reports (S.AR.A.), are then analysed to the end of identifying money laundering- or terrorist financing-related phenomena. For the detection of anomalies in international financial flows we propose to apply linear mixed models on cross-border wire transactions between Italy and three counterpart countries in 2019 following a high-performance statistical procedure to deal with high-dimensional data. Several model specifications are compared and an in-sample validation through perturbation of data is performed providing satisfactory results.

Abstract *In base alla normativa antiriciclaggio, le banche e gli altri intermediari finanziari operanti in Italia inviano mensilmente all'Unità di Informazione Finanziaria per l'Italia (UIF) le cosiddette Segnalazioni Antiriciclaggio Aggregate (S.AR.A.), relative alle operazioni effettuate dalla propria clientela. Tali dati vengono analizzati al fine di individuare potenziali fattispecie di riciclaggio e finanziamento del terrorismo. In questo lavoro proponiamo una metodologia per la rilevazione delle anomalie nei dati S.AR.A. mediante l'utilizzo di modelli lineari misti focalizzando l'analisi sui bonifici esteri del 2019 tra l'Italia e tre paesi esteri, adottando una procedura che tenga conto dell'alta dimensionalità dei dati. Diverse modelli sono stati stimati e opportunamente confrontati, ed è stata effettuata una verifica in-sample che ha fornito risultati soddisfacenti.*

Key words: anti-money laundering, illicit financial flows, cross-border wire transactions, anomaly detection, linear mixed model, high-dimensional data.

[1] Stefano Iezzi, UIF - Banca d'Italia; email: stefano.iezzi@bancaditalia.it

[2] Marianna Siino, UIF - Banca d'Italia; email: marianna.siino@bancaditalia.it

1 Introduction and data description[1]

Money laundering (ML) is a complex phenomenon encompassing all the techniques deployed by criminals to deceptively introduce money, earned by illegitimate sources, into the financial system. In spite of the strict international standards enforced in most countries for the purpose of fighting it, criminal organisations and terrorists actively engage in attempting to use financial institutions as vehicles for ML, which presents the latter with challenges such as regulatory compliance, maintaining financial security, preserving goodwill and reputation, and avoiding operational risks like liquidity crunch and lawsuits. With its connections to organized crime as well as terrorist financing, ML has become a serious issue worldwide and has been receiving considerable attention from national governments and international bodies.

Under the Italian anti-money laundering (AML) law[2], banks and other financial intermediaries in addition to filing suspicious transaction reports (STRs) to Italy's central anti-money laundering authority, the so-called Financial Intelligence Unit (FIU), are also mandated to produce the Anti-Money Laundering Aggregate Reports (S.AR.A., from the Italian acronym) on a monthly basis. These are threshold-based reports referring to all transactions for 15,000 euros or more[3]. For each reporting intermediary, transactions are aggregated according to several information related to the type of transaction, some of the customer's features and, with reference to wire transfers, the counterpart's country of residence and that of his/her intermediary. The data are non-nominal as no personal information is provided on the customer or the counterpart thereof; hence massive data mining and statistical analysis can be performed with no confidentiality-related issues, as opposed to analyses using STRs which contain sensitive data.

The view taken in our analysis is that of AML/CFT authorities: the approach described here may support FIUs in discharging their institutional task to identify money laundering related trends and patterns at a national scale. The relevance of the threats that illicit international financial flows pose to the stability of national financial systems and countries can never be understated, so much that fighting them is included within the United Nations Sustainable Development Goals[4]. Cross border transfers may thus represent a crucial vulnerability for a country financial integrity and need to be systematically and efficiently monitored.

In order to achieve this goal, we offer a contribution to the class of techniques for anomaly detection [1,2,4,6,10] by proposing the application of linear mixed-effect models to the Italian data on monthly cross-border wire transfers coming from the SARA. In particular, the models are estimated on 2019 data of cross-border wire

[1] The views and the opinions expressed in this paper are those of the authors and do not necessarily represent those of the institution they are affiliated with.

[2] Legislative Decree no. 231/2007

[3] Since 2021 the threshold has been lowered to 5,000 euros.

[4] Target 16.4 of the Sustainable Development Goals (SDGs) calls to "*by 2030, significantly reduce illicit financial flows and arms flow, strengthen the recovery and return of stolen assets and combat all forms of organised crime*."

transactions between Italy and three foreign countries with different volumes of records and amounts.

2 Model specification

The presence of multiple levels and repeated measurements on S.AR.A. data implies the presence of a complex source of random variability and correlation patterns. The class of linear mixed models [8,12] provides a flexible framework for modelling this type of data. The basic specification of our model is:

$$y_{ijkpq\bar{p}\bar{q}t} = \alpha + \gamma_t + \beta Sector + \mu_j + \delta_k + \vartheta_{i|jk} + \\ + \eta_p^{ITA} CL\_ITA + \eta_q^{FR} CL\_FR + \varphi_{\bar{p}}^{ITA} CT\_ITA + \varphi_{\bar{q}}^{FR} CT\_FR + \varepsilon_{ijkpqt}$$

where $y_{ijkpq\bar{p}\bar{q}t}$ is the amount of wire transfers recorded in month t in branch i of municipality j and bank k, ordered by clients resident in municipality p or foreign country q vis-à-vis counterparts resident in municipality $\bar{p}$ or foreign country $\bar{q}$. Parameter α is the intercept, γ_t is a time component, and $\varepsilon_{ijkmpqt}$ is the error term.

We have a set of pure random effects μ_j, δ_k and $\vartheta_{i|jk}$ which represent, respectively, the effect of the bank level, the effect of the branch's municipality, and the effect of the branch level, that is nested in the combination of the two previous levels. Then we have a random coefficient part: CL_ITA is the dummy for clients resident in Italy and CL_FR is the dummy for clients resident in foreign countries; η_p^{ITA} and η_q^{FR} are the corresponding random coefficients across municipalities and across countries respectively. A mirror random coefficient part is defined for counterparts by dummies CT_ITA and CT_FR, and random coefficients $\varphi_{\bar{p}}^{ITA}$ and $\varphi_{\bar{q}}^{FR}$. Finally, we have a fixed effect part made by the variable Sector, which defines the economic sector of the clients, and its fixed effect coefficient β. Furthermore, we adopt four alternative specifications for the time component: a quadratic parametric trend, a cubic parametric trend, monthly dummy variables and monthly dummy variables interacted with the economic sector of the clients, meaning that for each sector we estimate a time dynamic.

In this multilevel model structure we may wish to know at what level(s) and with respect to which explanatory variable(s) a particular transaction is outlying, as more than one level of analysis might be of interest [9]. On the one hand, model residuals are used to detect individual outlier transactions [7]: accordingly, a record is identified as anomalous when its Studentized residual is higher than the 1 per cent right tail critical value of the standard-normal distribution. On the other hand, group outliers can be identified by analysing the random part of a mixed effect model. The realized values of random effects, that are estimated by the Best Linear Unbiased Prediction (BLUP) procedure [11,12], can be interpreted as group residuals since they reflect

how much the level of a particular random effect factor deviates from the overall fit across the groups. The application of a proper inferential test allows the identification of potential group anomalies.

3 Main results

The proposed methodology is applied to 2019 S.AR.A. data of monthly wire transfers between Italy and three countries for which we have different amounts of records in our dataset (large, average and small amount).[1] Results are shown only for the country with a medium size amount of observations.

Before estimating the models the response variable is properly transformed with the Box-Cox method [3,5] in order to adjust for positive skewness. The transformed response variable is the dependent variable of seven alternative specifications. Overall, the model that has a complete structure with monthly dummies for the temporal components seems to be the best choice among all specifications since it performs well for all the countries and directions of flows.

In Table 1 and Table 2 the main results of the identified individual anomalies (in terms of percentage of records and total amount) and group anomalies (in terms of percentage of random effects for each aggregation variable) are shown, respectively. For instance, for the inward transactions 0.42 % of records are identified as anomalous, corresponding to 15.64% of total amounts. In addition, for group anomalies, 3.47% of branches are identified as anomalous.

Table 1 – Number of transactions and amount in millions of euros for the outward and inward wire-transactions between Italy and the selected country in 2019. The percentage of identified anomalous transactions and the corresponding percentage of total amounts are reported.

	Outward		**Inward**	
	N	% of anomalies	N	% of anomalies
Number of transactions	101,721	0.47	192,040	0.42
Amounts in millions of euros	79,272	13.35	85,024	15.69

Table 2 – Descriptive results of the group anomalies identified for the selected country considering the test statistics on the estimated random parameters with a p-value less than 0.0001. The number of significant random effect (n) and their percentage (%) with respect to the total number of effect for the specific group are shown.

			Outward		**Inward**	
Variable description		**Model parameters**	**n**	**%**	**n**	**%**
Financial intermediary	identification code	δ_k	3	0.76	1	0.24
	municipality of the branch	μ_j	2	0.07	1	0.03

[1] The names of the countries under analysis are not disclosed for confidentiality reasons.

	branch identification code	$\vartheta_{i\|jk}$	235	1.87	509	3.47
Customer	country of residence	η_q^{FR}	13	12.50	13	11.82
	municipality of residence	η_p^{ITA}	178	4.89	224	4.95
Counterpart	country of residence	$\varphi_{\bar{q}}^{FR}$	1	1.23	7	6.93
	municipality of residence	$\varphi_{\bar{p}}^{ITA}$	-	-	1	1.11

4 In-sample performance validation

In order to assess the capacity of the models to detect anomalies, we implement an in-sample validation through perturbation of the data by inserting artificial anomalies. Despite such anomalies do not replicate real anomalies we simulate them to the best of our capabilities as follows:

1. the number of records of each k-combination of the aggregate variables, including the month, are computed where k = 1,…, M+1, and M+1 are the number of aggregate variables plus the month;
2. the combinations with less than 10 records are discarded;
3. we randomly select one combination;
4. within the selected combination, a record is randomly selected;
5. the selected record is perturbed by assigning a random value in the range of $[V_1, V_2]$

$$V_1 = Q_3 + C \cdot IQR$$
$$V_2 = Q_3 + 2C \cdot IQR$$

where Q_3 and IQR are the third quartile and the interquartile range of the records within the selected combination, and $C > 1$ is a predefined constant, determining the degree of anomaly.

We generate four perturbed datasets for inward cross-border wire transfers by using a degree of anomaly c equal to 1.5, 2, 2.5 and 3 and injecting r={100, 1000} anomalous data points.

The performance of the anomaly indicator based on the Student residual is assessed in terms of sensitivity and specificity by changing the number of injected points (r) and the degree of anomaly (c) in the simulation design. According to the simulation results in Table 3, after increasing the degree of anomaly, both the sensitivity and specificity increase. Moving from 100 to 1,000 injected points in the dataset, the two statistics slightly decrease. Furthermore, the ROC curves computed by changing the threshold value used to define the indicator are overall satisfactory.

Table 3 – Sensitivity and specificity for the anomaly indexes based on the Student residual. The c value indicates the degree of anomaly.

r	Sensitivity				Specificity			
	c				c			
	1.5	2.0	2.5	3.0	1.5	2.0	2.5	3.0
100	0.680	0.750	0.810	0.840	0.996	0.996	0.996	0.997
1,000	0.600	0.700	0.775	0.808	0.997	0.998	0.998	0.999

5 Final remarks

The proposed approach can play a fundamental role in the strategic analysis of cross-border wire flows carried out by FIUs. Although the analysis focuses on the cross-border wire transactions, the proposed methodology, which takes properly into account the complex multi-level structure of S.AR.A. data, can have a more general purpose for transaction monitoring and detection of other types of potential anomalies, such as the cash usage, and for banking AML supervision. Furthermore, the outcome of this type of analysis can give an overview of specific dynamics and trends that can be share to the national investigative bodies and competent judicial authorities for AML/CTF purposes.

Moreover, the large amount of records (about 100 million per year) which banks and other financial intermediaries deliver in compliance with the Italian anti-money laundering law is a challenge for the Italian FIU in terms of analytical techniques and computational tools, and the proposed procedure is a step forward in this direction.

References

1. Barnett, V., & Lewis, T. (1994). Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics.
2. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.
3. Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.
4. Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2013). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and data Engineering, 26(9), 2250-2267.
5. Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the Box–Cox transformation to the linear mixed model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(2), 273-288.
6. Han, J., Huang, Y., Liu, S., & Towey, K. (2020). Artificial intelligence for anti-money laundering: a review and extension. Digital Finance, 2(3), 211-239.
7. Hawkins, Douglas M. Identification of outliers. Vol. 11. London: Chapman and Hall, 1980.
8. Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 963-974.
9. Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 161(2), 121-160.
10. Mehrotra, K. G., Mohan, C. K., & Huang, H. (2017). Anomaly detection principles and algorithms (Vol. 1). New York, NY, USA:: Springer International Publishing.
11. Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. Statistical science, 15-32.
12. Verbeke, G., & Molenberghs, G. (2009). Linear Mixed Models for Longitudinal Data. Springer Science & Business Media.

On the drivers of Greenwashing risk: evidence from Eurostoxx600

Le determinanti del Greenwashing: evidenze empiriche dall'Eurostoxx600

Yana Kostiuk, Costanza Bosone, Paola Cerchiello

Abstract The aim of this article is analysing both the external and internal drivers of greenwashing, relying on FinScience[1] company's well-established methodology and on rigorous statistical methods. We will shed light on ESG components, carefully identifying all the indicators which may potentially impact on a company sustainable performance. As crucial insights, we will show that the companies' tendency to "greenwash" seems correlated to (I) the industry in which they operate (II) the level of female presence in the company itself and (III) the size of the Board. Similar results, although with different variables, are in line with the existing literature (e.g. Chen et al., 2018; Lenard et al., 2014; Romero et al., 2019; Wilson & Altanlar, 2009; Yu et al., 2020).

Abstract *Questo articolo mira ad analizzare le determinanti interne ed esterne del "greenwashing". Utilizzeremo la solida metodologia della società di Fintech "FinScience" e specifici metodi statistici. Studieremo nel dettaglio le componenti degli ESG e le modalità in cui ogni componente può potenzialmente influenzare la per-*

Yana Kostiuk
University of Pavia, San Felice al Monastero, Pavia 5, Italy
University School for Advanced Studies Pavia (IUSS) - Palazzo del Broletto, 27100 Pavia (Italy)
e-mail: yana.kostiuk@iusspavia.it

Costanza Bosone
University of Pavia, San Felice al Monastero, Pavia 5, Italy,
University School for Advanced Studies Pavia (IUSS) - Palazzo del Broletto, 27100 Pavia (Italy)
e-mail: costanza.bosone@iusspavia.it

Paola Cerchiello
University of Pavia, San Felice al Monastero, Pavia 5, Italy,
University School for Advanced Studies Pavia (IUSS) - Palazzo del Broletto, 27100 Pavia (Italy)
e-mail: paola.cerchiello@unipv.it

[1] FinScience is a data-driven fintech company founded in 2017 by Google's former senior managers and Alternative Data experts. The company originates from the merger of the world of Finance and the world of Data Science through Augmented Analytics solutions applied to Investing activities (FinScience, n.d.).

formance di un'impresa in termini di sostenibilità. Come osservazioni fondamentali proveremo che la tendenza di un'impresa ad ingaggiare in pratiche di "greenwashing" sembra correlata (I) al settore in cui l'impresa stessa opera (II) al livello di donne assunte dall'impresa ed (III) al numero di membri del Consiglio. Simili risultati, pur con variabili differenti, sono in linea con la letteratura precedente (e.g. Chen et al., 2018; Lenard et al., 2014; Romero et al., 2019; Wilson & Altanlar, 2009; Yu et al., 2020).

Key words: ESG; Sustainable Finance; Greenwashing; Risk Management, AI

1 Introduction

The implementation of ESG criteria in the investment decision-making process of a company is defined as "sustainable finance" (Forstater & Zhang, 2016) and it has been gaining importance ever since the 2008's Great Recession, which brought about a positive shift in capital markets perception towards corporate sustainability (Escrig-Olmedo et al., 2019).

Despite a similar growth, ESG criteria have experienced little or no regularisation in terms of common guidelines for their definition or interpretation (Escrig-Olmedo et al., 2019; Mooney, 2021; Strobel, 2020), with the most significant achievement being the "EU taxonomy" as defined by the Technical Experts Group (TEG) (2019; 2020).

In a similar context, the interpretation of ESG scores, as provided by major financial platforms such as Refinitiv or Bloomberg, becomes challenging and sometimes, deceiving. Indeed, more and more firms engage in "greenwashing" behaviours, thus misleading stakeholders about their environmental performance and impact (Delmas & Burbano, 2011; Furlow, 2010).

The aim of this article is to shed light on the drivers of greenwashing, furtherly expanding the literature on the topic.

2 Methodology

We will shed light on ESG components, carefully identifying all the indicators potentially impacting on a company's sustainable performance. With regard to our scientific approach, we will rely on FinScience company's well-established methodology and on rigorous statistical methods, which will be explained more in details in the following paragraphs.

As for FinScience methodology, we opted for Eurostoxx600 as a sample of data. Since the EU seems more sensible to sustainability problems than other institutions (i.d. TEG, 2019; 2020), this choice is consistent with the aim of conducting an analysis on ESG. Time span of extraction ranges from 2020 to 2021. The sustainability

performance of a company is based both on self-disclosed data (internal score), external data (external score) and a combination of the two (overall score). Such scores range from 0 to 100. The Internal sources include: Sustainability/CSR Reports, Corporate Websites, Sustainability Memberships/Affiliations, Certifications, Sustainability Rankings, Controversies and Reviews. External sources include: ONG, Vertical Website and mainstream NEWS. We can state that internal scores are company-generated, whereas external scores are stakeholder-generated.

By combining together such two scores, we can produce an indicator of greenwashing risk. More in details, we consider the difference between internal and external scores - hereafter referred as *delta* - and used it as a proxy for greenwashing. In absolute terms, a higher *delta* implies a higher risk of greenwashing, as a company's internal perception drifts away from its stakeholders' external one. On the contrary, low values of *delta* refer to a better perception from the external stakeholders' point of view relative to the declared internal status.

We widen the Eurostoxx600 dataset by downloading the following variables from Refinitiv: total assets, Revenue per Share, ROA, ROE (all used as control variables); a variable accounting for the size of the Board, another accounting for the number of women employed in a company (%) and finally, a variable accounting for the presence of the CSR Sustainability Committee (dummy variable). In addition, we included the country where the headquarter of each firm is located as provided by Refinitiv (Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Spain, Sweden, UK) and clustered firms into 9 main operating sectors, according to the Global Industry Classification Standard (GICS) (MSCI, 2018): Communication services & Info, Consumer discretionary, Consumer Staple, Energy, Financials & Real Estate, Healtcare, Industrial, Materials, Utilities.

In line with those studies claiming that a company tendency to greenwash can be mitigated by influential public interests via a less corrupted country system (e.g. Yu et al., 2020), we included a variable accounting for the level of corruption in a country, namely: the Corruption Perception Index (CPI) as computed by the Transparency International Agency (2020). In the end, we obtained a dataset of roughly 8000 observations.

Tab. 1 provides a summary of statistics for all the continuous variables included in our dataset.

Variable	**Mean**	**Std.Dev.**	**Min**	**Max**
Delta	-46.41	13.34	-48	38
Board size	11.29	3.5	4	23
Total Assets	$1.20e^{11}$	$3.23e^{11}$	$0.4e^{11}$	$2.99e^{12}$
Women Employees	0.37	0.15	0.11	0.76
Revenue Per Share	59.04	191.29	0.06	3229.74
Return on Assets	0.06	0.08	-0.17	0.87
Return on Equity	0.14	0.26	-3.29	1.37

Table 1 Summary Statistics of continuous variables (not standardized).Source: Authors' Elaboration based on Finscience Data and Refinitiv, 2020-21

3 Empirical Findings

We standardized the dataset and performed a standard linear regression model and a stepwise linear regression model, using delta as dependent variable. **Tab.2** reports results for the stepwise model.

The following crucial insights can be retained: (I) the presence of women in a company (Women Employees) is negatively and significantly correlated to *delta*, thus implying that higher female presence favours lower risks of greenwashing. This is line with current literature (e.g. Chen et al., 2018; Lenard et al., 2014).

(II) The risk of greenwashing depends on the industry in which the company operates; companies in Energy, Consumer Staples and Discretionary, Healthcare and Materials seem to suffer from higher risks of greenwashing than others. This is fairly in line with the literature enquiring the correlation between environmental sensibility and different operating sectors (e.g. Romero et al., 2019).

(III) The size of the Board impacts on the probability of greenwashing. Though this result is in line with the literature, it usually goes in the opposite direction (e.g. Yu et al.,2020). As an explanation, we argue that our sample only focuses on the biggest companies in Europe which are likely to present similarly-sized Boards, thus reducing the impact of the variable.

The presence of a Corporate Social-Responsibility Committee (CSR Sustainability Committee) shows a positive and significant correlation, which may sound counter-intuitive to some extent. In reality, CSR committees are designed to fight corruption, protect stakeholders and reduce a company's exposure to failures in contexts where management becomes more complicated (Gennari & Salvioni, 2019), hence companies are more likely to set up such Committees when facing some distresses as a way of conquering or maintaining investors' trust. This explains the positive correlations between the presence of a CSR committee and the probability of greenwashing.

The corruption CPI index does not make its appearance in the model as it is not statistically significant. As a reasonable explanation, European countries do not greatly differ in terms of perceived levels of corruption (Transparency International, 2020).

4 Conclusions

In conclusion, by relying on FinScience internal and external score for sustainability, we obtain a *delta* score, which is used as a proxy for greenwashing. Firms with a higher number of female employees are less likely to engage in greenwashing behaviours. Vice versa, firms operating in Retail, Transports, Energy & Utility and Healthcare & Chemical face a higher risk of greenwashing. On its part, the size of the Board impacts as well on the probability to greenwash.

All these results are in line with the existing literature (e.g. Chen et al., 2018; Lenard et al., 2014; Romero et al., 2019; Wilson & Altanlar, 2009; Yu et al., 2020).

	Estimate	Std.Error	T value	Pr $> \|t\|$
(Intercept)	-0.35	0.12	-5.46	0.004***
Board size	0.1	0.04	2.31	0.026*
Total Assets	0.09	0.05	1.44	0.061 .
Women Employees	-0.2	0.04	-3.14	0.00***
CSR Sustainability Committee	0.15	0.14	3.15	0.00***
Energy	0.89	0.25	3.55	0.00***
Industrials	0.29 .	0.17	1.9	0.06 .
Consumer Discretionary	0.35	0.17	2	0.05 *
Consumer Staples	0.63	0.19	3.4	0.00***
Financials	0.04	0.17	0.248	0.8
Healthcare	0.85	0.19	4.5	0.00***
Materials	0.9	0.18	4.861	0.00***
Utilities	-0.07	0.21	-0.325	0.75

Table 2 Source: Authors' elaboration based on FINTECH data and data downloaded from Refinitiv, 2020-2021. Notes: Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1; Residual standard error: 0.88 on 449 degrees of freedom Multiple R-squared: 0.21; Adjusted R-squared: 0.19 ; F-statistic: 10.12 on 12 and 449 DF, p-value: 2.2e-16

Further analysis are needed. Given the considerable variation across non-European countries in terms of sustainability and "green-actions", enlarging the dataset beyond the Eurostoxx600 horizon -by means of inclusion of American and Asian companies- will lead to a more complete analysis on the drivers of greenwashing. Moreover, it is worth considering a deeper analysis of the two components of the 'delta' score indicator.

References

1. Chen, J., Leung, W. S., Evans, K. P. (2018). Female board representation, corporate innovation and firm performance. *Journal of Empirical Finance, 48, 236-254*
2. Delmas, M. A., Burbano, V. C. (2011). The drivers of greenwashing. *California management review*, 54(1), 64-87.
3. Escrig-Olmedo, E., Fernández-Izquierdo, M. Á., Ferrero-Ferrero, I., Rivera-Lirio, J. M., & Muñoz-Torres, M. J. (2019). Rating the raters: Evaluating how ESG rating agencies integrate sustainability principles. *Sustainability*, 11(3), 915.
4. FinScience. (n.d.). *About US| FinScience.* FinScience: Investment AI. Retrieved February 12, 2022 from: `https://finscience.com/en/about-us/`
5. Forstater, M.; & Zhang, N.N. (2016). Definitions and Concepts: Background Note. *UNEP Inquiry. Nairobi, Kenya*
6. Furlow, N. E. (2010). Greenwashing in the new millennium. *The Journal of Applied Business and Economics*, 10(6), 22.
7. Gennari, F., & Salvioni, D. M. (2019). Csr committees on boards: The impact of the external country level factors. *Journal of management and Governance*,23(3), 759–78
8. Lenard, M. J., Yu, B., York, E. A., Wu, S. (2014). Impact of board gender diversity on firm risk. *Managerial Finance*.
9. Mooney, A. (2021, May 10). ESG benchmark divergence no barrier to investor demand. *Financial Times*. Retrieved from: `https://www.ft.com/content/df328c34-6d9b-4fe6-9074-74091ce23ac7`

10. MSCI. (2018, September). *Global Industry Classification Standard (GICS)*. MSCI S&P GLOBAL. Retrived from: https://www.msci.com/our-solutions/indexes/gics
11. Romero, S., Ruiz, S., & Fernandez-Feijoo, B. (2019). Sustainability reporting and stakeholder engagement in Spain: Different instruments, different quality. *Business Strategy and the Environment*, 28(1), 221-232.
12. Strobel, G. (2020, May 5). Making Sense Of ESG: A Primer On Social Corporate Responsibility. *Forbes*. Retrieved from: https://www.forbes.com/sites/forbesfinancecouncil/2020/03/05/making-sense-of-esg-a-primer-on-social-corporate-responsibility/?sh=43bf58e9d471
13. TEG. (2019). *Teg Final Report On Climate Benchmarks And Benchmarks' Esg Disclosures*. Retrieved from: https://ec.europa.eu/info/sites/default/files/business_economy_euro/banking_and_finance/documents/190930-sustainable-finance-teg-climate-benchmarks-and-disclosures_en.pdf
14. TEG. (2020). *Taxonomy: Final report of the Technical Expert Group on Sustainable Finance*. Retrieved from: https://ec.europa.eu/info/sites/default/files/business_economy_euro/banking_and_finance/documents/200309-sustainable-finance-teg-final-report-taxonomy_en.pdf
15. Transparency International. (2021). *Corruption Perceptions Index*. Transparency International: The Global Coalition Agaisnt Corruption. Retrieved February 25, 2022, from https://www.transparency.org/en/cpi/2021
16. Wilson, N., & Altanlar, A. (2009). Director characteristics, gender balance and insolvency risk: an empirical study. *Gender Balance and Insolvency Risk: An Empirical Study (September 22, 2009)*.
17. Yu, E. P. Y., Van Luu, B., Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance, 52, 101192*.

Modelling Financial Returns with Finite Mixtures of GED

Modellazione dei Rendimenti Finanziari con Misture Finite di GED

Pierdomenico Duttilo and Stefano Antonio Gattone

Abstract As widely identified by the empirical evidence, daily returns on financial assets are not Normally distributed, because they are characterized by excess kurtosis and different degrees of skewness. Finite mixtures of distributions have been proposed in literature to capture these features. In this work a finite mixture of two Generalized Error Distributions (GED) is applied to fit the distribution of the daily returns on the Dow Jones Industrial Average (DJIA) index for the period from January 4, 2016 to January 31, 2022. Moreover, in order to highlight the flexibility of the shape parameter over time, the entire analysis period was divided in three different sub-periods and for each one the mixture of GED was estimated.

Abstract *Come ampiamente identificato dall'evidenza empirica, i rendimenti giornalieri delle attività finanziarie non sono distribuiti normalmente, poiché sono caratterizzati da curtosi eccessiva e diversi gradi di asimmetria. Per catturare queste caratteristiche, in letteratura sono state proposte misture finite di distribuzioni. In questo lavoro viene applicata una mistura finita di due distribuzioni GED (Generalized Error Distribution) per stimare la distribuzione dei rendimenti giornalieri dell'indice Dow Jones Industrial Average (DJIA) per il periodo dal 4 gennaio 2016 al 31 gennaio 2022. Inoltre, al fine di evidenziare la flessibilità del parametro di forma nel tempo, l'intero periodo di analisi è stato suddiviso in tre diversi sottoperiodi e per ciascuno è stata stimata la mistura di GED.*

Key words: daily financial returns, excess kurtosis, skewness, finite mixtures of GED

Pierdomenico Duttilo
University "G. d'Annunzio" of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy e-mail: pierdomenico.duttilo@unich.it

Stefano Antonio Gattone
DISFIPEQ, University "G. d'Annunzio" of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy
e-mail: gattone@unich.it

1 Introduction

Commonly, financial time series such as daily returns on stocks, indices, currencies, cryptocurrencies and many others financial assets do not follow a Gaussian distribution because they are characterized by excess kurtosis and different degrees of skewness. Finite mixtures of distributions have been proposed in literature to capture these features.
Finite mixture of Normal distributions known also as mixtures of Gaussians are widely used in this field, for instance Kon (1984) [5] proposed a discrete mixture of Normal distributions to approximate the excess kurtosis and positive or negative skewness of daily returns distribution of common stocks and indices. More recent studies [1, 2, 4] argued that finite mixtures of Gaussians (with two or three components) are a good tool to fit the empirical distribution of financial returns. However, these mixtures impose a priori specific constraints on the form of the returns distribution since the components are Gaussians. The finite mixture of GED, known also as the finite mixture of generalized normal distribution can overcome this critical issue thanks to the flexibility provided by the additional shape parameter ν_k.
In this framework, the recent contribution of Wen et al. [8] (2020) is remarkable. They studied a univariate mixture of GED and proposed an expectation conditional maximization (ECM) algorithm for parameter estimation. Additionally, using data sets of the S&P 500 and Shanghai Stock Exchange Composite Index (SSEC), it was found that the mixture of GED better describes the excess kurtosis and skewness of daily returns compared to mixtures of Gaussians.
This work aims to enrich the existing literature on the use of mixtures of GED in finance by applying a finite mixture of two generalized error distributions to fit the distribution of the daily returns on the Dow Jones Industrial Average (DJIA) index. Besides, the likelihood-ratio test (LR Test) and information criteria were applied to compare the goodness of fit performance among the mixture of two GED, the mixture of two Gaussian distributions and the mixture of a Gaussian and a Laplace distribution [3].
Moreover, the entire analysis period was divided in three different sub-periods and for each one the mixture of GED was estimated in order to highlight the flexibility of the shape parameter over time.
The rest of the paper was organized as follows. Section 2 illustrates the methodological framework. Section 3 illustrates the empirical application. Finally, Section 4 provides the results discussion and some conclusions.

2 Methodological Framework

2.1 Generalized Error Distribution

A random variable X is said to have the generalized error distribution with parameters μ (location), σ (scale) and ν (shape) if its *probability density function* (p.d.f.) is given by

$$f(x|\mu,\sigma,\nu) = \frac{\nu}{2\sigma\Gamma(1/\nu)} \exp\left\{ -\left| \frac{x-\mu}{\sigma} \right|^{\nu} \right\}, \tag{1}$$

with $\Gamma(1/\nu) = \int_0^{\infty} t^{1/\nu-1} \exp^{-t} dt$, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$. Thanks to the shape parameter, the GED distribution is a flexible tool to capture a large class of statistical distributions [7, 8], for example with $\nu = 1$ and $\nu = 2$ GED becomes a Normal and Laplace distribution, respectively.

2.2 Finite Mixtures of GED

A finite mixture of GED with K components is given by the marginal distribution of the random variable X

$$\begin{aligned} f(x|\mu_k,\sigma_k,\nu_k) &= \sum_{k=1}^{K} \pi_k p(x|\mu_k,\sigma_k,\nu_k), \\ &\sum_{k=1}^{K} \pi_k \frac{\nu_k}{2\sigma_k\Gamma(1/\nu_k)} \exp\left\{ -\left| \frac{x-\mu_k}{\sigma_k} \right|^{\nu_k} \right\}, \end{aligned} \tag{2}$$

where $\nu_k > 0$, $\sigma_k > 0$, $\mu_k \in \mathrm{R}$, $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$. With $K = 2$ the mixture of two GED is given by:

$$\begin{aligned} f(x|\theta) &= \sum_{k=1}^{2} \pi_k p(x|\mu_k,\sigma_k,\nu_k), \\ &= \frac{\pi_1\nu_1}{2\sigma_1\Gamma(1/\nu_1)} \exp\left\{ -\left| \frac{x-\mu_1}{\sigma_1} \right|^{\nu_1} \right\} + \frac{\pi_2\nu_2}{2\sigma_2\Gamma(1/\nu_2)} \exp\left\{ -\left| \frac{x-\mu_2}{\sigma_2} \right|^{\nu_2} \right\}, \end{aligned} \tag{3}$$

where $\theta = (\pi_1,\pi_2,\mu_1,\mu_2,\sigma_1,\sigma_2,\nu_1,\nu_2)$. The mixture of two GED can be estimated via the ECM algorithm [8]. As shown by Wen et al. (2020) [8], this model has important properties because it nests several distributions as its sub-models. Especially, depending on the value of the shape parameter (ν_k), the mixture of two GED reduces to:

- the mixture of two Gaussians when $\nu_1 = \nu_2 = 2$;
- the mixture of two Laplace distributions when $\nu_1 = \nu_2 = 1$;
- the mixture of a Gaussian and a Laplace distribution when $\nu_1 = 2$ and $\nu_2 = 1$;
- the mixture of a Gaussian and a GED distribution when $\nu_1 = 2$ and $\nu_2 > 0$;
- the mixture of a Laplace and a GED distribution when $\nu_1 = 1$ and $\nu_2 > 0$.

As a result the mixture of GED does not impose a priori specific constraint on the shape of each component of the mixture [6].

3 Empirical Application

The daily closing prices of the DJIA index between January 4, 2016 to January 31, 2022 were collected for the analysis from https://finance.yahoo.com. Next, the daily return r_t in period t is defined as $r_t = (\ln P_t - \ln P_{t-1})100$ where P_t and P_{t-1} are the closing prices at time t and $t-1$, respectively. Panel (a) of Figure 1 shows the daily returns on the DJIA index.
Estimation results of the mixtures of distributions with two components for the daily returns on the DJIA index are summarized in Table 1. According to the LR Test and information criteria the mixture of two GED (Figure 1, panel b) is preferred over the mixture of two Gaussians and the mixture of a Gaussian and a Laplace distribution. Furthermore, the estimated shape parameters show that the first component has heavier tails than a Laplace distribution ($0.53 < 1$), while tail weights of the second component intermediate between the Gaussian and the Laplace distribution ($1 < 1.12 < 2$). These results are in line with those found by Wen et al. (2020) [8] who estimated the two-component mixture of GED on the daily returns of the S&P500 identifying a mixture component that allows a more extreme tail behaviour compared to the other.

In order to highlight the flexibility of the shape parameter over time, the entire analysis period was divided in three different sub-periods and for each one the two-component mixture of GED was estimated. As showed in panel (a) of Figure 1 each sub-period reflects different volatility levels: low (2016-2017), intermediate (2018-2019) and high (2020-2022). In addition, the latter two sub-periods are characterized by a higher number of large negative returns, i.e. negative skewness. Finally, Table 2 suggests that the third sub-period has a two-component mixture of GED with heavy tails compared to the other sub-periods.

It is important to note that the current version of the mixture model proposed in this work can be cast in the framework of unconditional (with respect to time) estimation which suggests a poor predictive ability. Indeed, previous studies [1, 2, 4, 5, 8] do not apply mixtures of distributions to make predictions. Nonetheless, this work provides an out-of-sample application of the Value at Risk (VaR) estimation. The first two sub-periods are taken as "in-sample" observations and the third sub-period as the "out-of-sample" observations. The estimated $\widehat{\text{VaR}}_{\alpha=0.01}$ in the "in-sample" observations is -2.736, -2.482 and -2.364 for the mixture of GED, mixture of Gaussians, and mixture of a Gaussian-Laplace, respectively. The empirical $\text{VaR}_{\alpha=0.01}$ of

the third sub-period (out-of-sample) is -5.702. The evidence suggests that the third sub-period is characterised by an extreme $\text{VaR}_{\alpha=0.01}$ value (due to the COVID-19 crisis) and the estimated $\widehat{\text{VaR}}_{\alpha=0.01}$ of the two-component mixture of GED is the closest compared to the other two-component mixtures.

4 Conclusion

It has been shown that the mixture of GED is a powerful and flexible tool to fit the empirical distribution of financial returns. Considering the results of the empirical application in Section 3 at least two interesting considerations arise. Firstly, the mixture of GED with two components can model the behaviour of daily returns more appropriately and steadily compared to benchmark models, i.e. the mixture of two Gaussians and the mixture of a Gaussian and a Laplace distribution. Secondly, the estimated shape parameters change over time, they are not constant. Thus, the shape parameter changes according to the behaviour of daily returns i.e. market conditions. Consequently, the overall volatility estimated by the mixture in each sub-period (0.63, 0.97, 1.73) reflects the corresponding volatility level showed by Figure 1 (low, intermediate and high).

Appendix

Table 1: Estimation results of the two-components mixtures of distributions.

Parameter	Mixture of distributions		
	Gaussian	Gaussian-Laplace	GED
π_1	0.1416	0.5415	0.3221
π_2	0.8583	0.4584	0.6778
μ_1	-0.4242	0.1353	-0.0721
μ_2	0.1245	-0.0095	0.1553
σ_1	3.8912	0.6652	0.2115
σ_2	0.8876	1.0912	0.6317
ν_1	2	2	0.5304
ν_2	2	1	1.1245
$Stdev_1$	2.7515	0.4703	1.7759
$Stdev_2$	0.6276	1.5433	0.7468
$Stdev$	1.2031	1.1031	1.1855
LL	-2047.13	-2017.19	-1997.13
LR Test	100.00*	40.12*	
AIC	4099.26	4039.37	4001.25
BIC	4130.93	4071.04	4045.59
HQIC	4114.18	4054.30	4022.15
EDC	4133.39	4073.20	4049.03

*p-value = 0.

Table 2: Estimation results of the two-components mixtures of GED in the three sub-periods.

Parameter	Sub-periods		
	2016-2017	2018-2019	2020-2022
π_1	0.4890	0.7942	0.1180
π_2	0.5110	0.2058	0.8820
μ_1	0.0481	0.1507	-0.5622
μ_2	0.1220	-0.4872	0.1198
σ_1	0.3940	0.7469	3.4994
σ_2	0.8773	2.2960	0.9137
ν_1	2.1787	1.5205	1.1075
ν_2	1.3384	1.9479	1.2465
$Stdev_1$	0.2674	0.6344	4.2266
$Stdev_2$	0.8402	1.6464	0.9469
$Stdev$	0.6302	0.9716	1.7267
LL	-421.34	-647.61	-857.05

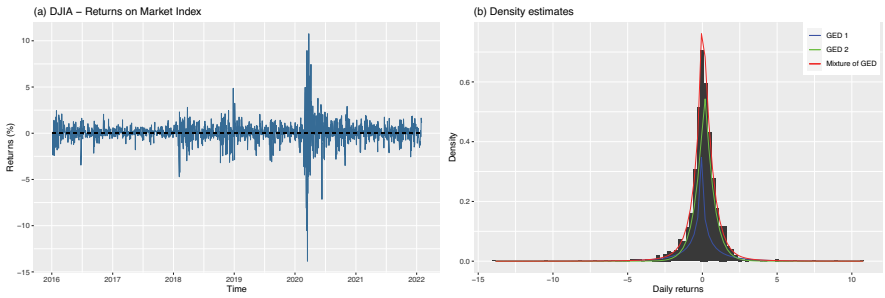


Fig. 1: Daily returns on the DJIA index and estimation of the density of the two-component mixture of GED.

References

1. Behr, A., Pötter, U.: Alternatives to the normal model of stock returns: Gaussian mixture, generalised logF and generalised hyperbolic models. Ann. Finance **5**, 4-9–68 (2009)
2. Bellalah, M., Lavielle, M.: A Decomposition of Empirical Distributions with Applications to the Valuation of Derivative Assets. Multinational Finance Journal **6**, 99–130 (2002)
3. Haas, M., Mittnik, S., Paolella, M.: Modelling and predicting market risk with Laplace-Gaussian mixture distributions. Applied Financial Economics. **16**, 1145–1162 (2006) doi:10.1080/09603100500438817.
4. Han, L., Yan, H., Zheng, C.: Normal mixture method for stock daily returns over different sub-periods. Communications in Statistics - Simulation and Computation **48**, 447–457 (2019) doi:10.1080/03610918.2017.1383423
5. Kon, S.J.: Models of Stock Returns–A Comparison. The Journal of Finance **39**, 147–165 (1984)
6. Mohamed, O.M., Jaïdane, M.: Generalized Gaussian mixture model. 17th European Signal Processing Conference, 2273–2277 (2009)
7. Nadarajah S.: A generalized normal distribution. Journal of Applied Statistics **32**, 685–694 (2005) doi:10.1080/02664760500079464
8. Wen, L., Qiu, Y., Wang, M., Yin, J., Chen, P.: Numerical characteristics and parameter estimation of finite mixed generalized normal distribution. Communications in Statistics - Simulation and Computation (2020) doi:10.1080/03610918.2020.1720733

Risk Parity strategy for portfolio construction: a kurtosis-based approach

Risk Parity strategy per la costruzione di un portafoglio: un approccio basato sulla curtosi

Maria Debora Braga, Consuelo Rubina Nava and Maria Grazia Zoia

Abstract Investors interested in the homogeneous distribution of responsibility for the dispersion of portfolio returns can pursue not only the equal constituents' contribution to portfolio volatility as, alternatively, they can focus on portfolio's extreme outcomes (either positive or negative). This leads to the proposal of a kurtosis-based risk parity strategy (KRP) that is a new risk parity strategy based on portfolio kurtosis as reference measure. In this paper, the portfolio weights, obtained with the KRP strategy. are compared with those estimated following the traditional standard deviation-based risk parity strategy. An analysis of KRP strategy performance is carried out through an out-of-sample study which takes advantage of real data from a global equity investment universe.

Abstract *Gli investitori, interessati a una distribuzione omogenea del rischio tra i vari asset di portafoglio, possono concentrarsi non solo sull'equa contribuzione dei costituenti alla volatilità di portafoglio poiché, in alternativa, essi si possono focalizzare sui risultati estremi di portafoglio (sia positivi che negativi) prendendo in considerazione la curtosi di portafoglio. In questo contesto, si sviluppa una strategia di parità di rischio basata sulla curtosi (strategia KRP) di portafoglio quale misura di rischio anziché sulla sua volatilità. I pesi per la costruzione del portafoglio ottenuti con la strategia KRP sono messi a confrontati con quelli ottenuti con la tradizionale strategia di parità di rischio basata sulla volatilità. Le performance della strategia KRP sono studiate con un approccio out-of-sample basato su dati reali di un universo di investimento azionario globale.*

Maria Debora Braga
SDA Bocconi School of Management, Via Roberto Sarfatti 10, Milano, e-mail: mariadebora.braga@sdabocconi.it

Consuelo Rubina Nava
Università degli Studi di Torino, Lungo Dora Siena 100, Torino, e-mail: consuelorubina.nava@unito.it

Maria Grazia Zoia
Università Cattolica del Sacro Cuore di Milano, Largo Gemelli 1, Milano e-mail: maria.zoia@unicatt.it

Key words: Kurtosis, Risk parity, Risk diversification, Asset allocation

1 Introduction

Since 2008, in finance, a new strategy for portfolio construction, known as risk parity (RP), has appeared [4]. Its main feature is the wealth allocation (accordingly to a suitable defined risk-measure) among asset classes in such a way that each asset contributes in the same manner to the selected portfolio risk-measure. Thus, risk contribution forms the basis for the development of the RP strategy [3]. So far, portfolio volatility has been the most used reference risk-measure.

Here we develop the new version of the RP strategy where portfolio volatility is replaced by the portfolio kurtosis as the reference risk-measure. The existing literature has already dealt with other risk measures but never with the kurtosis. This is also due to the absence of a close form expression of the portfolio kurtosis. This is also due to difficulty of deriving closed form expressions for the marginal risk contribution of each asset and, consequently, the set-up of the optimization problem, necessary to identify optimal portfolio weights.

In this paper, a risk parity strategy based on portfolio kurtosis as reference measure (KRP – kurtosis-based risk parity) is investigated, applied, and its performance is compared with the traditional standard deviation-based risk parity (SRP). In such a way a novel strategy able to better accommodate the needs of investors interested in the homogeneous distribution of responsibility for portfolio returns' huge dispersion, as that measured by portfolio kurtosis, is developed. A novel and effective expression for portfolio kurtosis is proposed, which integrates the results proposed in [1]. The proposed methodology is applied on real data and compared with the classical RP allocation strategy based on volatility in order to explore its peculiar features.

The paper is organized as follow. Section 2 proposes an original formula for portfolio kurtosis. Section 3 describes the dataset used for the empirical application and provides main results. Section 4 concludes the paper.

2 Methodology

Let consider a portfolio composed by N assets with returns $\boldsymbol{R} = [R_i]$ and associated weights $\boldsymbol{w} = [w_i]$ with $i = 1, \ldots, N$. The weights w_i are the percentage of asset class i in the portfolio and satisfy the following normalizing constraint $\sum_{i=1}^{N} w_i = 1$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ be the $N \times 1$ vectors of the expected returns and standard deviations for the selected asset classes in the investment universe: μ_i is the expected return of asset class i while σ_i is its risk. Furthermore, let $\boldsymbol{C} = [\rho_{ij}]$ and $\boldsymbol{\Sigma} = [\sigma_{ij}]$ be the $N \times N$ correlation matrix and covariance matrix. Accordingly, the expected return of the portfolio P and its variance are:

$$\mu_P = \boldsymbol{w}'\boldsymbol{\mu},\ \sigma_P^2 = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}.$$

Let's now recall for the purpose of this contribution some results concerning the kurtosis of two generic random variables R_i and R_j

$$K_i = K(R_i) = \mathbb{E}\left[\left(\frac{R_i-\mu_i}{\sigma_i}\right)^4\right] = \frac{\mathbb{E}[(R_i-\mu_i)]^4}{[\mathbb{E}[(R_i-\mu_i)]^2]^2} = \frac{\mu_{i,4}}{\sigma_i^4}$$

with $\frac{R_i-\mu_i}{\sigma_i}$ representing the standardized returns. Thus,

$$K(w_iR_i) = \mathbb{E}\left[\left(\frac{w_iR_i-w_i\mu_i}{w_i\sigma_i}\right)^4\right] = \frac{\mu_{i,4}}{\sigma_i^4} = K(R_i)$$

$$K(R_i+R_j) = \frac{\sum_{k_1+k_2=4}\frac{4!}{k_1!k_2!}\sigma_1^{k_1}\sigma_2^{k_2}coK(R.,R.,R.,R.)}{(\sigma_i^2+\sigma_j^2+2cov(R_i,R_j))^2}$$

where the cokurtosis (coK) is defined as follows

$$coK(R_i,R_i,R_i,R_j) = \frac{\mathbb{E}[(R_i-\mu_i)^3(R_j-\mu_j)]}{\sigma_i^3\sigma_j};$$
$$coK(R_i,R_i,R_j,R_j) = \frac{\mathbb{E}[(R_i-\mu_i)^2(R_j-\mu_j)^2]}{\sigma_i^2\sigma_j^2}.$$

The cokurtosis is invariant over linear transformation, i.e. $coK(w_iR_i,w_iR_i,w_jR_j,w_jR_j) = coK(R_i,R_i,R_j,R_j)$. Moreover, the particular case of $coK(R_i,R_i,R_i,R_i)$ reduces to the kurtosis of R_i, i.e. $coK(R_i,R_i,R_i,R_i) = K(R_i)$, still invariant over linear transformation. However, it is worth noting that $K(w_iR_i+w_jR_j) \neq K(R_i+R_j)$.

The previous results can be extended to obtain a portfolio kurtosis including N asset, which can be expressed as follows

$$\begin{aligned} K_P = K\left(\sum_{i=1}^N w_iR_i\right) &= \frac{\mathbb{E}\left[\sum_{i=1}^N w_iR_i - \sum_{i=1}^N w_i\mu_i\right]^4}{\left(\mathbb{E}\left[\sum_{i=1}^N w_iR_i - \sum_{i=1}^N w_i\mu_i\right]^2\right)^2} \\ &= \frac{\sum_{\sum_{i=1}^N k_i=4}\frac{4!}{k_1!\cdots k_N!}coK(R.,R.,R.,R.)\prod_{i=1}^N w_i^{k_i}\sigma_i^{k_i}}{\left(\sum_{\sum_{i=1}^N k_i=2}\frac{2!}{k_1!\cdots k_N!}cov(R.,R.)\prod_{i=1}^N w_i^{k_i}\right)^2} \end{aligned} \tag{1}$$

where $k_i = 0,1,2,3,4\ \forall i = 1,\ldots,N$, $cov(R.,R.)$ is the covariance between two assets while $coK(R.,R.,R.,R.)$ is the cokurtosis.

Thus, based on this portfolio kurtosis formula, which can be also written in a closed form for using matrix notation [1], the portfolio kurtosis gradient, needed for

the determination of the marginal risk contributions of the asset classes to portfolio kurtosis, can be computed.

In this connection, it is worth noting that the RP approach is based on the idea that portfolio risk must be equally distributed among asset classes, a goal that can be pursued for any risk measure RM, provided it is homogeneous of degree one in the weights. This is also the case of the portfolio kurtosis. All methodological details are reported in [1] while in the next section we show the effect of selecting a KRP strategy compared with a SRP one.

3 Data and Results

The empirical application is based on seven equity indices from January 2001 to December 2020, provided by Morgan Stanley Capital International. In particular, the selected indexes are MSCI EMU, MSCI UK, MSCI USA, MSCI CANADA, MSCI JAPAN, MSCI PACIFIC EX-JAPAN, MSCI EMERGING MARKETS (expressed in euros). Here weekly returns are considered. Further empirical results considering also monthly data can be found in [1].

Main summary statistics, together with the Jarque-Bera test, are displayed in Table 1. Looking at it, we see that the indexes show negative skewness and positive excess kurtosis. The Jarque-Bera test is always significantly rejected.

Table 1 Summary statistics of the standardized returns assets of the investment universe

Asset	Mean	Volatitily	Skewness	Kurtosis	JB test (p-value)
MSCI EMU	0,0010	0,0291	-0,8840	9,4767	0
MSCI UK	0,0006	0,0273	-0,7391	12,7522	0
MSCI USA	0,0015	0,0261	-0,4513	6,7135	0
MSCI CANADA	0,0014	0,0291	-0,7728	10,1454	0
MSCI JAPAN	0,0008	0,0264	-0,1579	5,4569	0
MSCI PACIFIC EX JP F	0,0017	0,0271	-0,8073	10,4945	0
MSCI EM	0,0019	0,0284	-0,1937	8,4607	0

The test bed for our methodological proposal takes advantage of a rolling sample approach [2], considering different estimation window lengths and applying alternative portfolio rebalancing frequencies to determine the portfolio weights with the KRP and SRP strategies. In particular, here we consider an estimation window that covers 5 years (260 weeks) over which we have calculated portfolios with quarterly (i.e. every 12 weeks) and semi-annual (i.e. every 24 weeks) rebalancing windows. Figure 1 presents the results for the selected estimation window, together with the alternative rebalancing frequencies considered.

It emerges that the KRP strategy turns out to lead to a much more unbalanced and erratic portfolio structure compared with the one of the SRP. The Shannon Entropy measure also shows a different behavior for the two risk parity strategies (Figure 2).

According to Figure 2, the Shannon index is close to 2 when the SRP strategy is adopted, while it shows a very unstable and oscillating behavior when the KRP strategy is adopted. It is worth noting that exposures to asset classes within the KRP-based portfolios are inversely linked to their contribution to the portfolio kurtosis.

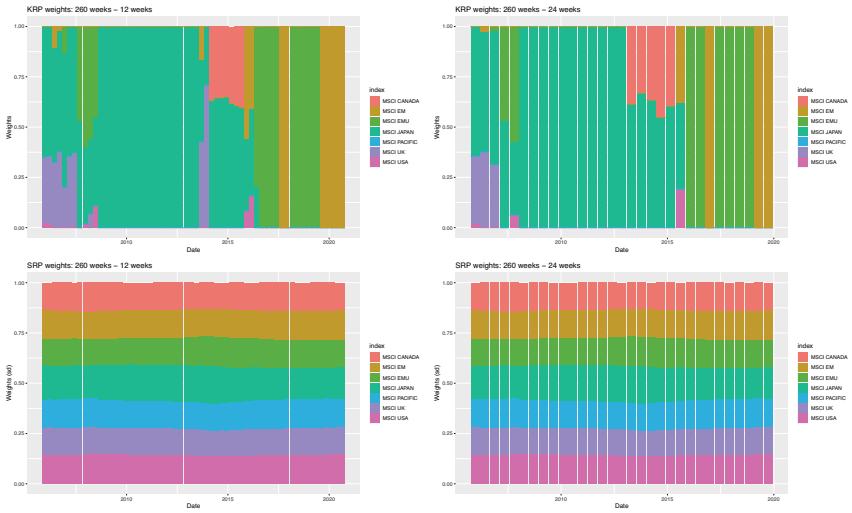


Fig. 1 Bar-charts representing the weights worked out using the KRP (top panels) and the SRP strategy (bottom panels). The estimation windows is 260 weeks long. The left panels refer to a rebalancing frequency of 12 week, while the right ones to 24 weeks.

The relevant characteristics of the out-of-sample returns provided by the two risk parity strategies have been also investigated. The KRP strategy systematically implies a lower kurtosis of out-of-sample returns. Finally, the risk-adjusted performance of the competing risk parity strategies, using the Sharpe ratio, the Sortino ratio and the Omega ratio have been evaluated. From Table 2, it results that KRP strategy provides persistently better reward per unit of risk taken.

Table 2 Sharpe, Omega and Sortino ratios obtained with the KRP and SRP strategies by using different rebalancing frequencies

Estimation window	Rebalancing frequency	Risk Parity strategy	Sharpe ratio	Omega ratio	Sortino ratio
260	12	KRP	0,380	1,159	0,520
		SRP	0,144	1,052	0,199
260	24	KRP	0,245	1,102	0,325
		SRP	0,126	1,046	0,171

Fig. 2 Shannon entropy using KRP and SRP strategies. The estimation windows is 260 weeks long. The top panel refer to a rebalancing frequency of 12 week, while the bottom one to 24 weeks.

4 Conclusions

Our empirical results reveal that a kurtosis-based risk parity strategy, compared to the classic risk parity, produces asset allocation solutions characterized by extremely unbalanced portfolio weights.

in addition we note that the "democratization" of kurtosis also helps its mitigation at least in comparison with the SRP strategy.

The methodological proposal of the paper, even if requires a more complex procedure for its implementation and leads to portfolio weights with a more unstable behavior, allows a better risk-return profile than SRP.

References

1. Braga, M.D., Nava, C.R., Zoia, M.G.: Kurtosis-based Risk Parity: Methodology and Portfolio Effects. ArXiv, (2022)
2. DeMiguel, V., Garlappi, L., Uppal, R.: Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? The review of Financial studies, **22**(5), 1915–1953, (2009)
3. Litterman, R.: Hot spots and hedges. Journal of Portfolio Management, (1996)
4. Maillard, S., Roncalli, T., Teïletche, J.: The properties of equally weighted risk contribution portfolios. The Journal of Portfolio Management. **36**, 4, 60–70, (2010)

Fully reconciled probabilistic *GDP* forecasts from Income and Expenditure sides

Riconciliazione completa delle previsioni probabilistiche del PIL dal lato del reddito e della spesa

Tommaso Di Fonzo and Daniele Girolimetto

Abstract We propose a complete reconciliation procedure of probabilistic *GDP* forecasts, resulting in *GDP* forecasts coherent with both Income and Expenditure sides' forecasted series, and evaluate its performance on the Australian quarterly *GDP* series, as compared to the original proposal by Athanasopoulos *et al.* (2020).
Abstract *In questo lavoro viene proposta una procedura di riconciliazione delle previsioni probabilistiche del PIL e delle sue componenti tanto dal lato del Reddito quanto da quello della Spesa, volta a produrre previsioni coerenti rispetto ad entrambi i lati. Tale procedura, applicata alle serie trimestrali del PIL australiano, viene posta a confronto con la proposta originale di Athanasopoulos et al. (2020).*

Key words: probabilistic forecast reconciliation, linearly constrained multiple time series, GDP, Income, Expenditure

1 Introduction and summary

Probabilistic forecasting has become relevant and widely used in many fields in recent years (Gneiting and Katzfuss, 2014). Panagiotelis *et al.* (2020) propose an effective definition of reconciled (coherent) probabilistic forecasts in the case of genuine[1] hierarchical/grouped time series, that has been further worked out by Wickramasuriya (2021) in a Gaussian framework.

The reconciliation of point and probabilistic forecasts of Australian *GDP* from Income and Expenditure sides, each one distinctly dealt with, was originally considered by Athanasopoulos *et al.* (2020) in order to perform aligned decision making and to improve forecast accuracy. In their empirical study they consider 95 Australian Quarterly National Accounts time series, describing the Gross Domestic Product (*GDP*) at current prices from Income and Expenditure sides, interpreted as two distinct hierarchical structures. In the former case (Income), *GDP* is on the top of 15 lower level aggregates, while in the latter (Expenditure), *GDP* is the top level aggregate of a hierarchy of 79 time series (for details, see Athanasopoulos *et al.*, 2020, and Bisaglia *et al.*, 2020).

T. Di Fonzo (e-mail: tommaso.difonzo@unipd.it); D. Girolimetto (e-mail: daniele.girolimetto@phd.unipd.it); Department of Statistical Sciences, University of Padova.

[1] In a genuine hierarchical/grouped time series, the link between the top-level and $n_b > 1$ bottom variables is uniquely defined (Di Fonzo and Girolimetto, 2021).

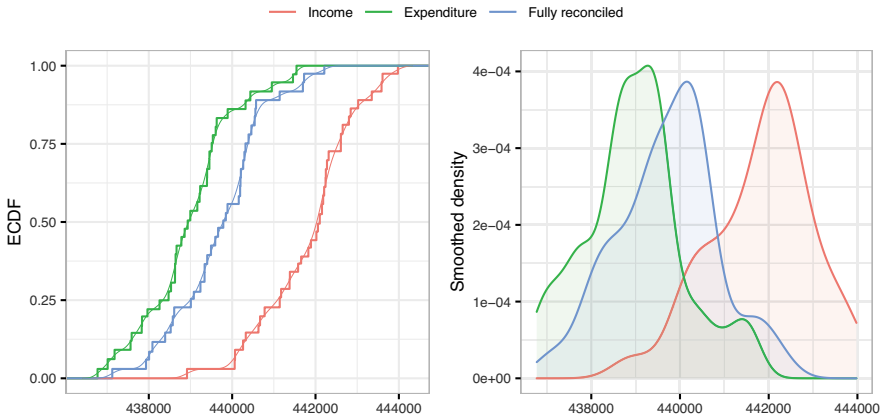


Fig. 1 *GDP* empirical 1-step-ahead forecast distributions for 2018:Q1, MinT-shr reconciliation approach. Empirical Cumulative Distribution Function (left), and Smoothed density (right).

In this paper, the results shown in Bisaglia *et al.* (2020) as for the reconciliation of point macroeconomic forecasts, are extended to a probabilistic forecasting framework. We re-consider the results of Athanasopoulos *et al.* (2020), where the probabilistic forecasts of the Australian quarterly *GDP* aggregates are separately reconciled from Income ($\widetilde{GDP}^{I}$) and Expenditure ($\widetilde{GDP}^{E}$) sides. This means that the empirical forecast distributions $\widetilde{GDP}^{I}$ and $\widetilde{GDP}^{E}$ are each coherent (according to Theorem 3.5 in Panagiotelis *et al.*, 2020) within its own pertaining side with the other empirical forecast distributions, but in general $\widetilde{GDP}^{I} \neq \widetilde{GDP}^{E}$ at any forecast horizon. This circumstance could confuse the user, mostly when the difference between the empirical forecast distributions is not negligible, as shown in Figure 1, where the *GDP* empirical forecast distributions from Income and Expenditure sides for 2018:Q1 are presented along with their fully reconciled counterparts through the MinT-shr approach (Athanasopoulos *et al.*, 2020). Bisaglia *et al.* (2020) propose a complete reconciliation strategy, able to produce a 'one number point forecast' of the *GDP* figure. However, a similar result is still missing for the probabilistic reconciliation setting. We fill this gap, showing how (incoherent) base probabilistic forecasts for general linearly constrained multiple time series (i.e., not necessarily genuine hierarchical/grouped time series) may be coherently reconciled[2].

2 Coherent probabilistic forecast reconciliation

We consider a linearly constrained n-dimensional multiple time series $\{\mathbf{y}_t\}_{t=1}^{T}$, with constraints expressed in homogeneous form as $\mathbf{U}'\mathbf{y}_t = \mathbf{0}_{r\times 1}$, where $\mathbf{U}'$ is a $(r \times n)$ matrix of known coefficients. For a genuine hierarchical/grouped time series, setting $n_a = r$, the standard structural representation (Hyndman *et al.*, 2011) holds:

[2] Note that the naive practice of averaging *GDP* forecasts from different sides yields a single forecast, that is though inconsistent with the component variables from both sides.

$$\mathbf{y}_t = \mathbf{S}\mathbf{b}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{S} = \begin{bmatrix}\mathbf{C}' & \mathbf{I}_{n_b}\end{bmatrix}'$ is the summing matrix mapping the n_b bottom time series $\mathbf{b}_t$ into the complete vector $\mathbf{y}_t = \begin{bmatrix}\mathbf{a}_t' & \mathbf{b}_t'\end{bmatrix}'$. The $(n_a \times 1)$ vector $\mathbf{a}_t$ contains the aggregated (upper) time series, $\mathbf{C} \in \{0,1\}^{n_a \times n_b}$ is the contemporaneous aggregation matrix (Di Fonzo and Girolimetto, 2021), $\mathbf{a}_t = \mathbf{C}\mathbf{b}_t$, and $\mathbf{U}' = \begin{bmatrix}\mathbf{I}_{n_a} & -\mathbf{C}\end{bmatrix}$.

To exploit definitions and results of Panagiotelis *et al.* (2020), valid for genuine hierarchical/grouped time series expressed as in (1), we need to find out a relationship like expression (1) for a general linearly constrained multiple time series as well. In general, for example when the number of constraints makes manual handling of the problem difficult, this can be done in the following steps:

1. compute $\mathbf{R}'$, the $(n_v \times n_f)$ reduced row echelon form (*rref*, Strang, 2019) of matrix $\mathbf{U}'$, where possible $r - n_v$ zero rows are not considered;
2. find out a $(n \times n)$ permutation matrix $\bar{\mathbf{P}}$, such that $\bar{\mathbf{U}}' = \mathbf{R}'\bar{\mathbf{P}}'$ has the form $\bar{\mathbf{U}}' = \begin{bmatrix}\mathbf{I}_{n_v} & -\bar{\mathbf{C}}\end{bmatrix}$, where $\bar{\mathbf{C}}$ is a $(n_v \times n_f)$ matrix of known coefficients (differently from matrix $\mathbf{C}$, the items of matrix $\bar{\mathbf{C}}$ are not necessarily equal to either 0 or 1, and may be negative);
3. define the matrix $\bar{\mathbf{S}} = \begin{bmatrix}\bar{\mathbf{C}}' & \mathbf{I}_{n_f}\end{bmatrix}'$ (analogous of the summing matrix $\mathbf{S}$), and $\bar{\mathbf{y}}_t = \bar{\mathbf{P}}\mathbf{y}_t = \begin{bmatrix}\mathbf{v}_t' & \mathbf{f}_t'\end{bmatrix}'$, where $\mathbf{f}_t$ is a $(n_f \times 1)$ vector of 'free' variables, and $\mathbf{v}_t$ is a $(n_v \times 1)$ vector of 'basic' variables, such that $\mathbf{v}_t = \bar{\mathbf{C}}\mathbf{f}_t$, $t = 1, \ldots, T$;
4. define the 'structural' representation of a general linearly constrained multiple time series as
$$\bar{\mathbf{y}}_t = \bar{\mathbf{S}}\mathbf{f}_t, \quad t = 1, \ldots, T; \tag{2}$$
5. consider the general optimal combination reconciliation formula according to the projection approach (Di Fonzo and Girolimetto, 2021), $\tilde{\mathbf{y}}_h = \mathbf{M}\hat{\mathbf{y}}_h$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{W}_h\mathbf{U}\left(\mathbf{U}'\mathbf{W}_h\mathbf{U}\right)^{-1}\mathbf{U}'$, with $\mathbf{W}_h$ $(n \times n)$ p.d. matrix, and express the same result as:
$$\bar{\mathbf{P}}\tilde{\bar{\mathbf{y}}}_h = \bar{\mathbf{S}}\bar{\mathbf{G}}\bar{\mathbf{P}}\hat{\bar{\mathbf{y}}}_h, \tag{3}$$
where $\bar{\mathbf{G}} = \left(\bar{\mathbf{S}}'\bar{\mathbf{P}}'\mathbf{W}_h^{-1}\bar{\mathbf{P}}\bar{\mathbf{S}}\right)^{-1}\bar{\mathbf{S}}'\bar{\mathbf{P}}'\mathbf{W}_h^{-1}$;
6. given n_f (and $n_v = n - n_f$), $\tilde{\mathbf{y}}_h$ is uniquely determined and does not depend on a particular choice of the free variables. This point is formally clarified in the extended version of the paper.

We consider 3 cases for the approximation of the covariance matrix $\mathbf{W}_h$:

- ols: $\mathbf{W}_h = \sigma^2\mathbf{I}_n$ (Hyndman *et al.*, 2011),
- wls: $\mathbf{W}_h = \widehat{\mathbf{W}}_D = \operatorname{diag}\widehat{\mathbf{W}}_1$ (Hyndman *et al.*, 2016),
- shr: $\mathbf{W}_h = \widehat{\mathbf{W}}_{shr} = \lambda\widehat{\mathbf{W}}_D + (1-\lambda)\widehat{\mathbf{W}}_1$ (Wickramasuriya *et al.*, 2019),

where $\widehat{\mathbf{W}}_1$ is the $(n \times n)$ covariance matrix of the in-sample one-step-ahead base forecasts errors, and $\widehat{\mathbf{W}}_{shr}$ is its shrinked version (Wickramasuriya *et al.*, 2019).

Representation (2) means that $\bar{\mathbf{y}}_t$ lies in an n-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $\bar{\mathbf{S}}$, called "coherent subspace" and denoted by $\mathscr{S}$ (Panagiotelis *et al.*, 2020). Now, let $\mathscr{F}_{\mathbb{R}^{n_f}}$ be the Borel σ-algebra on $\mathbb{R}^{n_f}$, $\left(\mathbb{R}^{n_f}, \mathscr{F}_{\mathbb{R}^{n_f}}, \nu\right)$ a probability space for the free variables, and $\bar{s}: \mathbb{R}^{n_f} \to \mathbb{R}^n$ a continuous mapping matrix. Then a σ-algebra $\mathscr{F}_{\mathscr{S}}$ can be constructed from the collection of sets $\bar{s}(\mathscr{B})$ for all $\mathscr{B} \in \mathscr{F}_{\mathbb{R}^n}$.

Definition 1. (*Coherent probabilistic forecast for a linearly constrained multiple time series*) Given the triple $\left(\mathbb{R}^{n_f}, \mathscr{F}_{\mathbb{R}^{n_f}}, \nu\right)$, we define a coherent probability triple $\left(\overline{\mathscr{S}}, \mathscr{F}_{\overline{\mathscr{S}}}, \check{\nu}\right)$ such that $\check{\nu}(\overline{s}(\mathscr{B})) = \nu(\mathscr{B})$, $\forall \mathscr{B} \in \mathscr{F}_{\mathbb{R}^{n_f}}$.

In order to extend forecast reconciliation to the probabilistic setting, let $(\mathbb{R}^n, \mathscr{F}_{\mathbb{R}^n}, \hat{\nu})$ be a probability triple characterizing base (incoherent) probabilistic forecasts for all n series, and let $\psi : \mathbb{R}^{n_f} \to \mathbb{R}^n$ be a continuous mapping function defined by Panagiotelis *et al.* (2020) as the composition of two transformations, $\overline{s} \circ \overline{g}$, where $\overline{g} : \mathbb{R}^n \to \mathbb{R}^{n_f}$ is a continuous function corresponding to matrix $\overline{\mathbf{G}}$ in equation (3).

Definition 2. (*Probabilistic forecast reconciliation for a linearly constrained multiple time series*). The reconciled probability measure of $\hat{\nu}$ with respect to ψ is a probability measure $\tilde{\nu}$ on $\overline{\mathscr{S}}$ with σ-algebra $\mathscr{F}_{\overline{\mathscr{S}}}$ such that

$$\tilde{\nu}(\mathscr{A}) = \hat{\nu}\left(\psi^{-1}(\mathscr{A})\right), \quad \forall \mathscr{A} \in \mathscr{F}_{\overline{\mathscr{S}}}, \tag{4}$$

where $\psi^{-1}(\mathscr{A}) = \{x \in \mathbb{R}^n : \psi(x) \in \mathscr{A}\}$ is the pre-image of $\mathscr{A}$.

3 Joint bootstrap-based probabilistic forecast reconciliation

Since an analytical expression of the forecast distribution is either unavailable, or relies on unrealistic parametric assumptions, the empirical evaluation of the results will be grounded on reconciled samples obtained according to Theorem 3.5 in Panagiotelis *et al.* (2020)[3]:

Theorem 1. *(Reconciled samples). Suppose that* $\left(\hat{\mathbf{y}}^{[1]}, \ldots, \hat{\mathbf{y}}^{[L]}\right)$ *is a sample drawn from an incoherent probability measure* $\hat{\nu}$*. Then* $\left(\tilde{\mathbf{y}}^{[1]}, \ldots, \tilde{\mathbf{y}}^{[L]}\right)$*, where* $\tilde{\mathbf{y}}^{[\ell]} := \psi\left(\hat{\mathbf{y}}^{[\ell]}\right)$ *for* $\ell = 1, \ldots, L$*, is a sample drawn from the reconciled probability measure* $\tilde{\nu}$ *as defined in (4).*

According to this theorem, reconciling each member of a sample obtained from an incoherent distribution yields a sample from the reconciled distribution. As a consequence, coherent probabilistic forecasts may be developed through a post-forecasting mechanism analogous to the point forecast reconciliation setting. For this purpose, the bootstrap procedure by Gamakumara *et al.* (2018) is applied:

1. appropriate univariate models M_i for each series in the system are fitted based on the training data $\{y_{i,t}\}_{t=1}^{T}$, $i = 1, ..., n$, and the one-step-ahead in-sample forecast errors are stacked in an $(n \times T)$ matrix, $\widehat{\mathbf{E}} = \{\hat{e}_{i,t}\}$;
2. $\hat{\mathbf{y}}_{i,h}^{[l]} = f_i\left(M_i, \hat{e}_{i,h}^{[l]}\right)$ is computed for $h = 1, ..., H$ and $l = 1, ..., L$, where $f(\cdot)$ is a function of the fitted univariate model and associated error, $\hat{\mathbf{y}}_{i,h}^{[l]}$ is a sample path simulated for the $i-$th series, and $\hat{e}_{i,h}^{[l]}$ is the $(i,h)-$th element of an $(n \times H)$ block bootstrap matrix containing H consecutive columns randomly drawn from $\hat{\mathbf{E}}$;
3. the optimal combination reconciliation formula (3) is performed for each $\hat{\mathbf{y}}_{h}^{[l]}$.

[3] Extension to a linearly constrained multiple time series for the Gaussian case (Wickramasuriya, 2021) is currently under study, and will be presented in the extended version of the paper.

The accuracy of the probabilistic forecasts is evaluated using the Cumulative Rank Probability Score (CRPS, Gneiting and Katzfuss, 2014, Panagiotelis *et al.*, 2020). In addition, we employ the Energy Score (ES), that is the CRPS extension to the multivariate case, to evaluate the forecasting accuracy for the whole system (Gamakumara *et al.*, 2018).

4 Reconciled probabilistic forecasts of the Australian *GDP*

For the complete Australian *GDP* accounts from both Income and Expenditure sides, it is $n = 95$ ($n_f = 62$ and $n_v = 33$), and the homogeneous constraints are described by matrix $\mathbf{U}'$ shown in Bisaglia *et al.* (2020). In addition, the available time series span over the period 1984:Q4 - 2018:Q1.

Athanasopoulos *et al.* (2020) produced base forecasts for each time series using univariate ARIMA models (with the `auto.arima` function of the R-package). Our reconciliation proposal is applied within the same forecasting experiment, that considers forecasts from $h = 1$ quarter ahead up to $h = 4$ quarters ahead, using an *expanding* window, where the first training sample is set from 1984:Q4 to 1994:Q3. The base forecasts are reconciled using the `htsrec` function of the R-package `FoReco` (Girolimetto and Di Fonzo, 2022).

Figure 2 shows the *skill scores* of CRPS for the *GDP* reconciled forecasts (top panel), and of ES for Income and Expenditure sides, and the whole system (bottom panel), that is the percentage changes registered by these indices for the considered reconciliation procedures, computed such that positive values signal an improvement in forecasting accuracy over the base forecasts. 'Income' and 'Expenditure' panels, respectively, refer to the results found by Athanasopoulos *et al.* (2020), while the 'Fully reconciled' panel shows the *skill scores* of the fully reconciled probabilistic forecasts. From the bottom panel of Figure 2, it appears that the reconciliation improves on the base forecasts' accuracy, shr and wls offering almost always the best performance. For *GDP*, ols outperforms both wls and shr, whatever side is considered. In general, the simultaneously reconciled probabilistic forecasts give results as good as those of Athanasopoulos *et al.* (2020). In addition, the newly proposed approach produces forecasts that are fully coherent with all economic constraints coming from National Accounts relationships.

References

1. Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R.J., Affan, M. (2020), Hierarchical Forecasting, in Fuleky, P. (ed.), *Macroeconomic Forecasting in the Era of Big Data*, Cham, Springer, 689–719.
2. Bisaglia, L., Di Fonzo, T., Girolimetto, D. (2020), Fully reconciled GDP forecasts from Income and Expenditure sides, in Pollice, A., Salvati, N., Schirripa Spagnolo F. (eds.), *Book of short papers SIS 2020*, 951–956, Pearson.
3. Di Fonzo, T., and Girolimetto, D. (2021), Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives, *International Journal of Forecasting*, in press.
4. Gamakumara, P., Panagiotelis, A., Athanasopoulos, G., Hyndman, R.J. (2018), *Probabilistic Forecasts in Hierarchical Time Series*, Department of Econometrics and Business Statistics, Monash University, Working Paper 11/18.
5. Girolimetto, D., Di Fonzo, T. (2022), *FoReco: Point Forecast Reconciliation*, R package version 0.2.4, `https://CRAN.R-project.org/package=FoReco`.
6. Gneiting, T. and Katzfuss, M. (2014), Probabilistic forecasts, *Annual Review of Statistics and Its Application*, 1, 125-151.

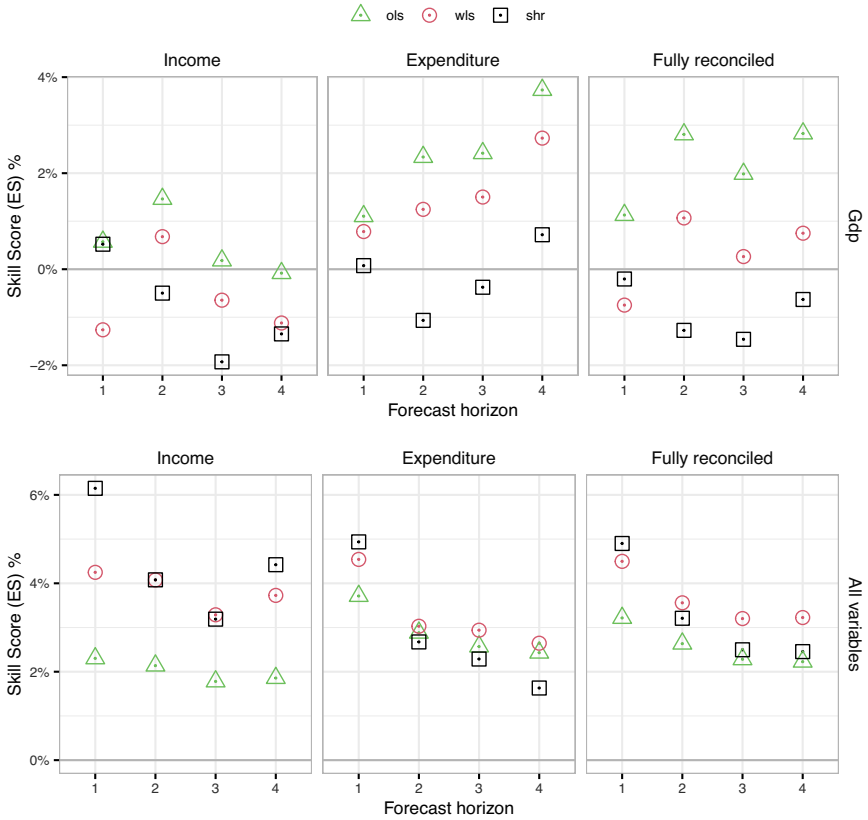


Fig. 2 Skill scores (relative to base forecasts) of CRPS and ES indices for probabilistic forecasts from alternative reconciliation approaches.

7. Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L. (2011), Optimal combination forecasts for hierarchical time series, *Computational Statistics and Data Analysis*, 55, 2579–2589.
8. Hyndman, R. J., Lee, A.J., and Wang E., (2016), Fast computation of reconciled forecasts for hierarchical and grouped time series, *Computational Statistics and Data Analysis*, 97, 16–32.
9. Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R.J. (2020), *Probabilistic Forecast Reconciliation: Properties, Evaluation and Score Optimisation*, Monash University, Department of Econometrics and Business Statistics, Working Paper 26/20.
10. Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R.J. (2021), Forecast reconciliation: A geometric view with new insights on bias correction, *International Journal of Forecasting*, 37, 1, 343–359.
11. Strang, G., (2019), *Linear algebra and learning from data*, Wellesley, Cambridge Press.
12. Wickramasuriya, S.L., (2021), Probabilistic forecast reconciliation under the Gaussian framework, `arXiv.2103.11128`.
13. Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J. (2019), Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization, *Journal of the American Statistical Association*, 114, 526, 804–819.

Acknowledgments

The authors acknowledge financial support from project PRIN2017 "HiDEA: Advanced Econometrics for High-frequency Data", 2017RSMPZZ.

Latent Class models

Latent thresholds model in classification tasks

Modello delle soglie latenti nei compiti di classificazione

Giuseppe Mignemi, Andrea Spoto, Antonio Calcagnì

Abstract Categorical rating scale with two, three or more ordered categories are widely used in many contexts such as screening and diagnostic assessment, quality control, sport refereeing or emergency. The present research aims to model classification tasks in which raters can evaluate the belonging of an item to a certain category according to several ordered levels. It is supposed that each rater has his/her own thresholds (i.e., a reference point) to perform the classification task. The classification outcomes have been described as a function of two independent latent sources of uncertainty, one related to the item, the other one related to the rater. Model parameters have been estimated using an ABC-based algorithm. Finally, the characteristics of the proposed method have been discussed by means of a real case study.

Abstract *Valutazioni categoriali con due, tre o più categorie ordinate sono ampiamente utilizzate in molti contesti come lo screening e la valutazione diagnostica, il controllo della qualita, l'arbitraggio sportivo o situazioni d'emergenza. Il modello proposto mira a descrivere le attività di classificazione in cui rater differenti possono valutare il grado d'appartenenza di un oggetto ad una determinata categoria secondo diversi livelli ordinati. Si suppone che ogni valutatore abbia le proprie soglie per svolgere il compito di classificazione. I risultati della classificazione sono stati descritti in funzione di due fonti di incertezza latenti indipendenti, una relativa all'item, l'altra relativa al valutatore. I parametri del modello sono stati stimati utilizzando un algoritmo ABC-based. Infine, le caratteristiche del metodo proposto sono state discusse attraverso un caso studio reale*

Giuseppe Mignemi
University of Padova e-mail: giuseppe.mignemi@phd.unipd.it

Andrea Spoto
University of Padova e-mail: andrea.spoto@unipd.it

Antonio Calcagnì
University of Padova e-mail: antonio.calcagni@unipd.it

Key words: Classification process, Conditional model, Approximate bayesian computation

1 Introduction

Categorical rating scales with two, three or more ordered categories are widely used in many contexts such as screening and diagnostic assessment [9], sport refereeing [10] or emergency department [4]. Consensus is expected among raters when they classify the same items. It is deemed that raters' agreement is evidence of an accurate and reliable classification, although their opinion may differ, resulting in a controversial classification. For instance, several research works in biostatistics have been moving toward the modelling of this systematic individual variability using GLMM (e.g., [6, 9, 8]). The following contribute aims to model classification tasks in which raters can evaluate the belonging of an item to a certain category according to several ordered levels. It is supposed that each rater has his/her own levels thresholds (i.e., a reference point) to make this kind of classification. Examples of such a situation include the case of a physician who has to classify a patient's disease severity or an emergency operator who has to classify the urgency of intervention.

2 Model specification

Consider the case where a rater $i \in \{1,...,I\}$ has to classify one item $j \in \{1,...,J\}$ according to an ordered set of disjoint categories $\mathscr{C} \doteq \{1,...,C\} \subset \mathbb{N}$. For each item, we observe I different unit vectors $\mathbf{w}_{ij} = [w_{ij1},...,w_{ijC}] \in \{0,1\}^C$, one for each rater, in which the element $w_{ijc} = 1$ indicates that the rater i classifies the item j into the c-*th* category. The observed data are modeled as a function of two independent latent processes, the first of which is related to specific characteristics of the item, whereas the other is related to rater-specific characteristics. As for other psychometric models,the item-specific component is represented by the latent variable $\eta_j \in (0,1)$, whereas the rater-specific component is represented by a vector of independent latent variables, $\boldsymbol{\xi}_i = [\xi_{i1},...,\xi_{iC}] \in \mathbb{R}^C_+$. The vector of observed values $\mathbf{w}_{ij}$ is function of η_j and $\boldsymbol{\xi}_i$ so that the complete sample space is $\mathscr{Y} : (0,1)^J \times \mathbb{R}^C \times \{0,1\}^{J\times I\times C}$. We propose the following model specification for the aforementioned process. For a generic pair (i,j), the model is specified as follows:

$$\eta_j \sim Beta(\eta_j;\alpha_j,\beta_j), \tag{1}$$

$$\xi_{ic} \sim logNormal(\xi_{ic};\mu_c,\sigma_c), \quad \xi_{ic} \in \boldsymbol{\xi}_i \tag{2}$$

$$\pi_{ijc} = \mathscr{P}(w_{ijc} = 1;\boldsymbol{\theta}) = \int_{t(\boldsymbol{\xi}_i,c-1)}^{t(\boldsymbol{\xi}_i,c)} Beta(\eta_j;\alpha_j,\beta_j)\,d\eta, \quad \pi_{ijc} \in \boldsymbol{\pi}_{ij}, \tag{3}$$

$$\mathbf{W}_{ij}|\eta_j,\boldsymbol{\xi}_i \sim Multinomial(\mathbf{W}_{ij};n=1,m=C,\pi=\boldsymbol{\pi}_{ij}), \tag{4}$$

where $\theta=[\alpha_{\eta_j},\beta_{\eta_j},\mu_1,\ldots,\mu_C,\sigma_1,\ldots,\sigma_C]$ with parameter space $\Theta:\{\alpha_{\eta_j},\beta_{\eta_j},\mu_1,\ldots,\mu_C,\sigma_1,\ldots,\sigma_C\in\mathbb{R}_+\}$. The function $t:\mathbb{R}_+^C\times\mathscr{C}\rightarrow(0,1)$ is applied to $\boldsymbol{\xi}_i$ so that

$$t(\boldsymbol{\xi}_i,c)=\frac{\sum_{c=1}^{c}\xi_{ic}}{\sum_{c=1}^{C}\xi_{ic}}\ .$$

The following properties are assumed: $\eta_{ij}\perp\!\!\!\perp\boldsymbol{\xi}_i$, $\xi_{ic}\perp\!\!\!\perp\xi_{ic+1}$ and the $t(\boldsymbol{\xi}_i,c)=0$ if $c=1$ and $t(\boldsymbol{\xi}_i,c)=1$ if $c=C$.
The joint density function is

$$p(\boldsymbol{\eta},\boldsymbol{\xi},\mathbf{w};\boldsymbol{\theta})=\prod_{i=1}^{I}\prod_{j=1}^{J}\left(\prod_{c=1}^{C}\pi_{ijc}^{w_{ijc}}\right)B(\alpha_j,\beta_j)^{-1}\eta_j^{\alpha_j-1}(1-\eta_j)^{\beta_j-1}\times$$
$$\times\prod_{c=1}^{C}\frac{1}{\sqrt{2\pi^C|\boldsymbol{\Sigma}|}}exp\{-\frac{1}{2}\left[\log(\xi_{ic})-\mu_c\right]^T\boldsymbol{\Sigma}^{-1}\left[\log(\xi_{ic})-\mu_c\right]\} \tag{5}$$

2.1 Parameter estimation

Because of the incomplete nature of observed data, the likelihood function implied by the data generating process is computational prohibitive, it requires integrating the joint likelihood function over all the complete sample space [12, 14, 1]. Therefore a *likelihood-free* approach might be applied to circumventing these problems. The *Approximate Bayesian computation* (ABC) methods bypass the direct numerical evaluation of the likelihood function by using a measure of distance d (e.g., Euclidean distance, Mahalanobis distance) between an observed summary statistics based on the actual data $S(y_{obs})$ and the one obtained by the simulated data $S(y_b)$, $b=1,...,B$, where B is the number of the simulations. If the distance $d(S(y_b),S(y_{obs}))$ is less than a fixed threshold (i.e., tolerance acceptance condition) the related parameter value θ_b is accepted. The set of the accepted parameter values $\pi_{ABC}(\theta|y_{obs})$ represents a sample from the approximated posterior distribution $\tilde{\pi}(\theta|y_obs)$. Several techniques have been developed to improve this approximation, including regression adjustments [2], strategies to choose $S(y_{obs})$ [11], or sampling procedures [13, 15, 7].

3 Application

We analyzed a data set from the retail context. We consider a convenience sample of $N_r=203$ consumers, who were administered with a survey in which they were asked to classify each of $N_i=10$ different products of the same brand according to $C=3$ mutually exclusive categories (i.e.,*"Bad"*,*"Neutral"*,*"Good"*). Their choices have been converted into boolean variables (e.g.,*"Bad"*:$=[1,0,0]$, *"Neutral"*:$=[0,1,0]$)

resulting in an array-structured data set. As required by the ABC approach a collection of summary statistics was computed. The $N_i \times C$ integer matrix containing the marginal counts of the categories chosen for each product (i.e., how many times a product was classified into each of the three categories) was used as summary statistics.

Two parameters per product and two parameters per category had to be estimated, resulting in a vector of $N_i \times 2 + C \times 2 = 26$ parameters. We tested two different set of priors to fit our model. Within the former, all the priors were uniform distributed as follows: $\alpha_j \sim Unif(0.1,5), \beta_j \sim Unif(0.1,5), j = 1,..,J$; $\kappa_1 \sim Unif(0.1,20), \kappa_2 \sim Unif(0.1,30), \kappa_3 \sim Unif(0.1,50), \sigma_c \sim Unif(0.1,1), c = 1,..,C$. For the second one, the priors were specified as follows: $\alpha_j \sim LN(0.5,0.09), \beta_j \sim LN(0.5,0.09), j = 1,..,J$; $\kappa_c \sim LN(3,0.5), \sigma_c \sim Unif(0.1,1), c = 1,..,C$.

The *R* package *Easy-ABC* [5] was used to get the simulated summary statistics based on both the specified priors and the observed summary statistic. The ABC-MCMC algorithms proposed by Marjoram [7] with Wegmann's [15] improvements was used to explore the parameter space. The approximate posterior distributions were estimated through the ABC method applying a regression adjustment accounting for *heteroscedasticity* (i.e., neural networks) using of the *R* package *abc* [3]. In the case of uniform priors, we chose a logit transformation of the parameters before regression adjustment as guarantee that the parameter values lie in the range of the prior.

The two sets of priors were compared by means of the Bayes factors and the posterior model probabilities computed using a logistic regression-based variant of ABC with neural network [1, 2]. The model fitted with uniform priors showed a Bayes factor of 12796.3817 and a posterior model probability of .99. Furthermore, a Principal Component Analysis performed using the a priori simulated summary statistics showed the better fit of the model with uniform priors. Thus, the parameters inference was based on the model with uniform priors. The accuracy of the estimation was assessed through a leave-one-out cross validation; the mean prediction error along all the parameters was 0.61. The features of the approximate posterior distributions of the parameters are shown in Table 1. Figure 1 shows an example the prior and the posterior (with and without regression-based correction) densities of the parameter κ_1, the euclidean distance as function of the parameter values and Q-Q plot of the regression's residuals.

4 Conclusion

We have developed a novel approach to the statistical modeling of raters classification process. Raters' classification outcomes have been described as function of two independent latent sources of uncertainty. One related to the item, the other one related to the rater. However, as for any statistical models, further improvements might be implemented. For instance, covariates might be specified in equations (1,2). The assumption of conditional independence between the elements of $\boldsymbol{\eta}$ may be relaxed

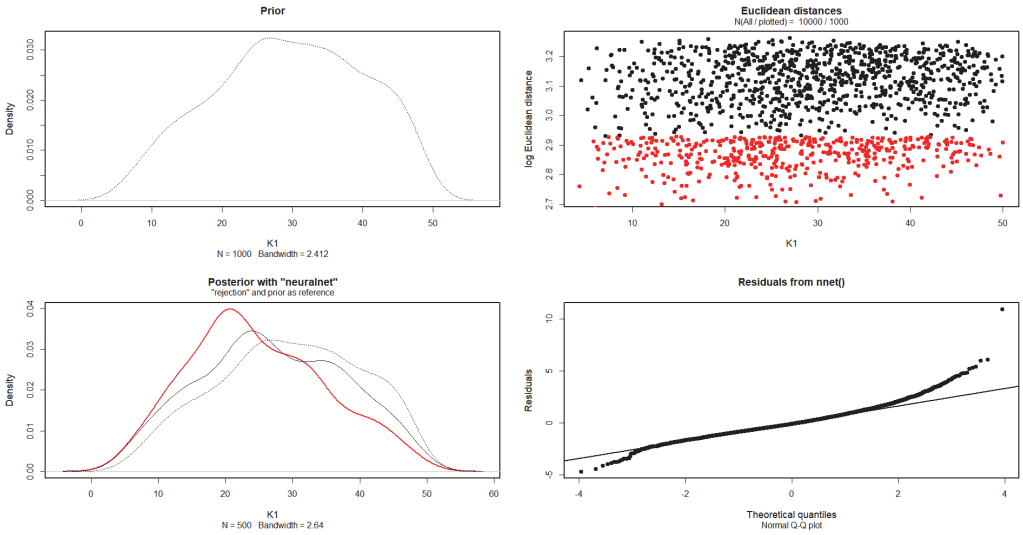


Fig. 1 ABC regression diagnostics for the estimation of the posterior distribution of K1. Posterior density and the accepted points are red.

Table 1 Posterior distributions' features

	Posterior mean	95% HDI	Posterior variance
α_1	2.90	[0.87,4.62]	2.91
α_2	1.96	[0.49,4.51]	2.44
α_3	2.16	[0.48,4.30]	2.27
α_4	2.68	[0.79,4.68]	2.50
α_5	2.16	[0.60,4.57]	2.36
α_6	3.11	[0.60,4.57]	3.22
α_7	2.35	[0.85,4.39]	2.57
α_8	3.43	[1.62,4.88]	3.50
α_9	1.56	[0.46,3.66]	2.03
α_{10}	3.38	[1.57,4.85]	3.28
β_1	2.80	[1.00,4.73]	2.73
β_2	1.21	[0.34,3.23]	1.55
β_3	2.59	[0.81,4.62]	2.88
β_4	3.02	[1.17,4.78]	2.63
β_5	2.64	[0.83,4.60]	2.72
β_6	2.15	[0.71,4.23]	2.27
β_7	2.36	[0.88,4.62]	2.61
β_8	2.91	[1.17,4.66]	2.88
β_9	1.42	[0.49,3.77]	1.47
β_{10}	3.26	[1.51,4.75]	3.21
μ_1	24.01	[8.46,44.79]	28.54
μ_2	16.69	[6.73,25.74]	17.12
μ_3	13.14	[4.85,19.59]	13.40
σ_1	0.61	[0.15,0.97]	0.60
σ_2	0.48	[0.11,0.92]	0.54
σ_3	0.56	[0.13,0.96]	0.56

using a copulas as joint density function. Finally, possible clusters among the items may be addressed in further research works.

References

1. BEAUMONT, M. A. Approximate bayesian computation. *Annual Review of Statistics and Its Application 6*, 1 (2019), 379–403.
2. BLUM, M. G. B., FRANÇOIS, O., AND FRANÇOIS, O. Non-linear regression models for approximate bayesian computation. *Stat Comput 20* (2010), 63–73.
3. CSILLÉRY, K., FRANÇOIS, O., AND BLUM, M. G. B. abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution 3*, 3 (2012), 475–479.
4. DIPPENAAR, E. Reliability and validity of three international triage systems within a private health-care group in the middle east. *International emergency nursing 51* (7 2020).
5. JABOT, F., FAURE, T., AND DUMOULIN, N. Easyabc: performing efficient approximate bayesian computation sampling schemes using r. *Methods in Ecology and Evolution 4*, 7 (2013), 684–687.
6. KIM, C., LIN, X., AND NELSON, K. P. Measuring rater bias in diagnostic tests with ordinal ratings. *Statistics in medicine 40* (7 2021), 4014–4033.
7. MARJORAM, P., MOLITOR, J., PLAGNOL, V., AND TAVARÉ, S. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences 100* (12 2003), 15324–15328.
8. NELSON, K. P., AND EDWARDS, D. On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics 36* (9 2008), 411–426.
9. NELSON, K. P., AND EDWARDS, D. Measures of agreement between many raters for ordinal classifications. *Statistics in medicine 34* (10 2015), 3116.
10. POWIS, B., AND MACBETH, J. L. "we know who is a cheat and who is not. but what can you do?": Athletes' perspectives on classification in visually impaired sport:. *https://doi.org/10.1177/1012690218825209 55* (2 2019), 588–602.
11. PRANGLE, D. Summary statistics. *Handbook of Approximate Bayesian Computation* (12 2018), 125–152.
12. RODRIGUES, G. S., FRANCIS, A. R., SISSON, S. A., AND TANAKA, M. M. Inferences on the acquisition of multi-drug resistance in mycobacterium tuberculosis using molecular epidemiological data. *Handbook of Approximate Bayesian Computation* (12 2020), 481–511.
13. SISSON, S. A., AND FAN, Y. Abc samplers. *Handbook of Approximate Bayesian Computation* (12 2020), 87–123.
14. SISSON, S. A., FAN, Y., AND BEAUMONT, M. A. Overview of abc. *Handbook of Approximate Bayesian Computation* (12 2018), 3–54.
15. WEGMANN, D., LEUENBERGER, C., AND EXCOFFIER, L. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics 182* (8 2009), 1207–1218.

Adaptive filters for time-varying correlation parameters

Filtri adattivi per parametri di correlazione che variano nel tempo

Michele Lambardi di San Miniato, Ruggero Bellio, Luca Grassetti, Paolo Vidoni

Abstract Adaptive filters from control theory are recursive estimators for time-varying parameters. We propose an adaptive filter targeting spatial and temporal correlation parameters, that we conjecture to be more stable in the case of regime shifts. This proposal should be suitable for online environmental monitoring tasks. We illustrate an application to sensor data with a spatio-temporal conditional autoregressive model.

Abstract *I filtri adattivi, afferenti alla teoria del controllo, sono stimatori ricorsivi per parametri variabili nel tempo. Proponiamo un filtro adattivo per parametri di correlazione spaziale e temporale, che congetturiamo essere più stabile di fronte a cambi di regime. Questa proposta pare adatta per applicazioni di monitoraggio ambientale in tempo reale. Un'applicazione a dati provenienti da sensori illustra la metodologia, mediante un modello condizionale autoregressivo spazio-temporale.*

Key words: adaptive filters, conditional autoregression, sensor data, spatio-temporal statistics

Michele Lambardi di San Miniato
University of Udine, e-mail: michele.lambardi@uniud.it

Ruggero Bellio
University of Udine, e-mail: ruggero.bellio@uniud.it

Luca Grassetti
University of Udine, e-mail: luca.grassetti@uniud.it

Paolo Vidoni
University of Udine, e-mail: paolo.vidoni@uniud.it

1 Adaptive filtering approach

Environmental monitoring often requires to collect physical data by means of sensors and to provide predictions in real time. A suitable model can be highly nonlinear, complicated, and even time-varying. One may wait and collect a long training set, then estimate a complicated model. Here, we opt for an adaptive filtering approach [2], where even a simple model can be made flexible by assuming a time-varying parameter vector and then tracking it via recursive estimators.

An adaptive filter provides a *current* estimate $\hat{\theta}_n$ for a real-valued parameter vector θ at step n and updates into $\hat{\theta}_{n+1}$ upon observing a new data chunk y_{n+1}, possibly without parsing older data $y_n, y_{n-1}, \ldots$ again. The update steers towards parameter values that are more supported in the light of y_{n+1}, while compromising with $\hat{\theta}_n$. A hyperparameter $\lambda \in]0,1[$ is involved, called learning rate, which weighs y_{n+1} relative to $\hat{\theta}_n$. Following [2], the update can be formulated as

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \lambda\, C_{n+1}^{-1}\, g_{n+1}\,,$$

where $g_{n+1} = g(\hat{\theta}_n; y_{n+1})$ and $g(\cdot; y)$ is the gradient with respect to θ of some objective function $G(\theta; y)$; C_n is a positive-definite condition matrix that tunes the efficiency aspects of the filter. In a likelihood framework, it is sound to define $G(\cdot; y)$ as the log-likelihood function conditional on the past observations, $g(\cdot; y)$ as the score function, and then set

$$C_{n+1} = (1-\lambda) C_n + \lambda\, g_{n+1}\, g_{n+1}^{\top}\,, \tag{1}$$

which emulates the Fisher information matrix: we refer to this as *Type 1 conditioning*. We also propose the following *Type 2 conditioning*:

$$C_n = \mathrm{diag}(\sqrt{v_n})\,, \quad \text{where} \quad v_{n+1} = (1-\lambda) v_n + \lambda\, g_{n+1} \circ g_{n+1}\,, \tag{2}$$

where $\circ$ is the Hadamard product, thus $v \circ v$ is just the vector v with squared components, while $\mathrm{diag}(v)$ is a diagonal matrix with vector v as its main diagonal. Thus, C_n retains only the diagonal square-rooted entries of the information matrix.

This Type 2 formulation is less efficient, but also less memory- and compute-intensive. It is numerically stable as there is virtually no risk of C_n becoming near-singular. Type 2 filters may also be better than Type 1 at discounting initial conditions and thus coping with regime shifts; this is only a conjecture, based on the resemblance with AdaGrad stochastic gradient descent [5], and supported to some extent by the numerical example provided in the following.

Here we apply Type 1 and 2 filters for a spatio-temporal conditional autoregressive model (STCAR) [4] and then compare the two estimates on a data stream, based on the Intel Lab Data, available in the public domain [1].

2 Spatio-temporal model

Let $(y_{st})_{s,t}$ be a spatio-temporal data matrix, where $s \in \mathscr{S} = \{1,2,\ldots,S\}$ is the space index and $t \in \mathscr{T} = \{1,2,\ldots,T\}$ is the time index. We assume the data are available on a grid, that is, for all pairs $(s,t) \in \mathscr{S} \times \mathscr{T}$, or can be projected onto such a grid.

Let the data be modeled by a random matrix $(Y_{st})_{s,t}$. Define $X_{st} = Y_{st} - \mu$, where $\mu \in \mathrm{R}$ is a constant mean parameter. Let $Y_t = (Y_{st})_s$ and $X_t = (X_{st})_s$ be column vectors, for all $t \in \mathscr{T}$. Let W be a $S \times S$ matrix of known weights with generic s-th row $w_s = (w_{s1},\ldots,w_{sS})^\top$, such that $w_{ss} = 0$ and $\bar{X}_{st} = w_s^\top X_t$ represents a spatial prediction for X_{st}. Following [4], the STCAR model is defined as

$$(\mathrm{Id} - \rho W)(1 - \phi L^{\Delta})(Y_t - \mu) = \varepsilon_t \,.$$

Here, Id is the $S \times S$ identity matrix and

- ρ and ϕ are the spatial and temporal dependence parameters, respectively;
- L is a temporal lag operator, so $L^{\Delta} a_{st} = a_{s(t-\Delta)}$ for all st-indexed quantities a_{st};
- $\Delta \in \{1,2,\ldots\}$ is a temporal lag hyperparameter;
- $\varepsilon_t = (\varepsilon_{st})_s$, and ε_{st} are Gaussian white noise with constant variance σ^2.

As in adaptive filtering approach, we treat the parameter vector $\theta = (\mu, \sigma, \rho, \phi)^\top$ as time-varying. Upon observing $y_{n+1} = y_{st}$, the current estimate $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{\rho}_n, \hat{\phi}_n)^\top$ is updated suitably. Exponential smoothing is used to update $\hat{\mu}_n$ and $\hat{\sigma}_n$ into

$$\hat{\mu}_{n+1} = (1-\lambda)\hat{\mu}_n + \lambda Y_{st} \,, \quad \hat{\sigma}^2_{n+1} = (1-\lambda)\hat{\sigma}^2_n + \lambda \hat{\varepsilon}^2_{st} \,,$$

where $\hat{\varepsilon}_{st} = X_{st} - \hat{\rho}_n \bar{X}_{st} - \hat{\phi}_n (X_{s(t-\Delta)} - \hat{\rho}_n \bar{X}_{s(t-\Delta)})$ and $X_{st} = Y_{st} - \hat{\mu}_n$. In estimating the correlation parameter $\psi = (\rho, \phi)^\top$, the filter $\hat{\psi}_{n+1} = \hat{\psi}_n + \lambda\, C^{-1}_{n+1}\, g_{n+1}$ can be based on the pseudo score function, after replacing μ and σ with their current estimates. Thus, we set for use in (1) and (2)

$$g_{n+1} = \hat{\varepsilon}_{st} \cdot (\bar{X}_{st} - \hat{\phi}_n \bar{X}_{s(t-\Delta)} \,,\, X_{s(t-\Delta)} - \hat{\rho}_n \bar{X}_{s(t-\Delta)})^\top / \hat{\sigma}^2_n \,.$$

3 Data analysis

The Intel Lab Data were collected by means of a mesh network with 54 sensors in the Intel Berkeley Research Lab between February 28 and April 5, 2004 [1]. The sensors were numbered and located in the building as in Figure 1. The sensors collected data asynchronously, roughly every 30 seconds, often with slight delays due to routing. We project the data onto a grid layout: upon observing the generic Y_{st}, we define $Y_{s'(t-k+1)}$ as the k-th most recent observation from sensor s'. This is a pragmatic approximation for the case when it is not perfectly known whether all sensors will be able to provide readings on time as requested.

The original data include nearly 2.3 million rows. Room temperature in Celsius degrees (°C) is recorded. We dropped all observations taken by sensors whose bat-

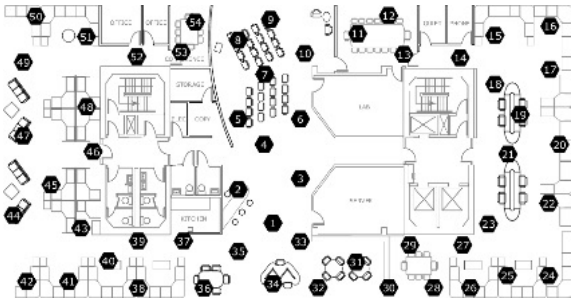


Fig. 1 Map of the Intel Berkeley Research lab. Sensors are numbered from 1 to 54. The image is modified from [1].

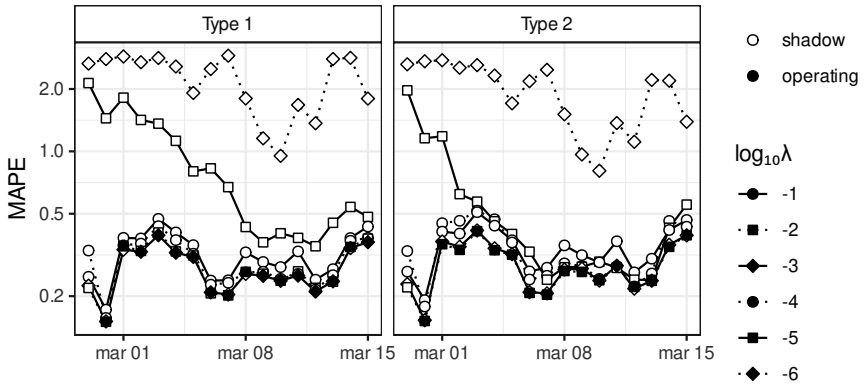


Fig. 2: Daily MAPE depending on λ and filter type. Operating sensors are selected at the end of each day, so there is no operating sensor on the first day.

tery voltage was less than or equal to $2.4V$, since this condition was not nominal. Then, we dropped all the observations after March 15, as too few sensors were still in working order. We thus kept 1.36 million data rows.

The Intel Lab Data are accompanied by connectivity data, which consist of the fractions of successful pings among sensors. We use these data as an automatic proximity measure [3]. We then define W such that $\bar{X}_{st}$ represents a k-nearest-neighbors prediction; for simplicity, we set $k = 1$, since choosing $k > 1$ was seen leading to no significantly different results.

We restrict the domain of the hyperparameter λ and consider the case when it takes values in $\Lambda = \{10^{-i} : i = 1,\dots,6\}$. Multiple configurations of λ are compared based on some performance metric, such as the mean absolute prediction error (MAPE). We predict $Y_{s(t+\Delta)}$ via $\hat{Y}_{s(t+\Delta)} = (1-\hat{\phi}_n)\hat{\mu}_n + \hat{\phi}_n Y_{st}$, where Δ is the lag hyperparameter of the model. We assume $\Delta = 20 \approx 10$ minutes, as this value turns out to provide the best balance between spatial and temporal effects on the predictions.

In Figure 2, the daily MAPE is reported, for all configurations of λ and for both Type 1 and 2 filters. We assume that all the filters run in parallel and are evaluated

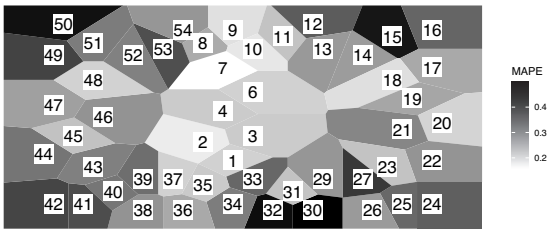


Fig. 3 MAPE, per location, based on the Type 2 filter with $\lambda = 10^{-3}$. Sensor numbers are in line with Figure 1. There were no data for sensors 5 and 28.

at the end of each day, so the best filter insofar is selected as the "operating" one, providing the official predictions of the system for the subsequent day, while the others are hence "shadow" filters acting as a backup. As a by-product of selection, λ is tuned automatically. Both Type 1 and 2 filters seem to harness the intermediate values of λ. The six instances of the Type 2 filter are only slightly less efficient than Type 1, but they are more stable when dealing with more extreme values of λ. The "operating" filter, performs near-optimally on the subsequent day, independently of the filter type. Type 1 operating filters only slightly outperform their Type 2 operating counterparts, whose MAPE is at most 0.03°C higher on March 14 and 0.01°C lower on March 7.

Now, we illustrate the Type 2 filter with $\lambda = 10^{-3}$, which is near-optimal over the whole observation period, as in Figure 2. In Figure 3, the MAPE is reported for each location. The STCAR covariance structure partly explains heteroscedasticity; net of this factor, prediction would look poor only in the table area (bottom center) and the enclosed office rooms (top).

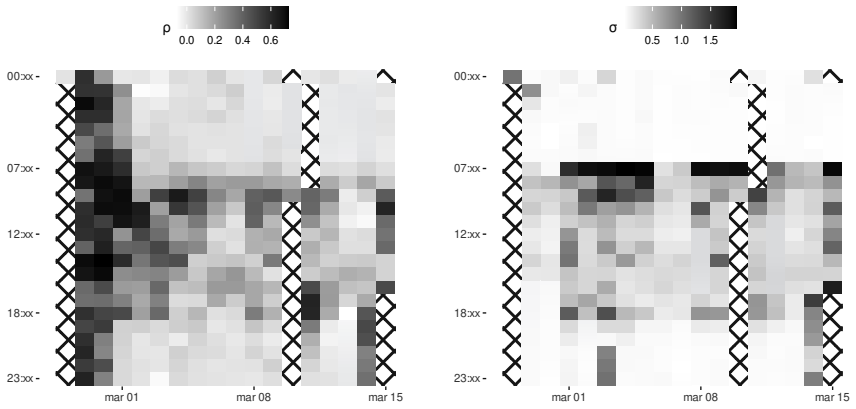


Fig. 4: Hourly averaged parameters, based on the Type 2 filter with $\lambda = 10^{-3}$. Crosshatch pattern in tiles stands for missing or insufficient data.

In Figure 4 the hourly averages of $\hat{\rho}_n$ (left) and $\hat{\sigma}_n$ (right) are reported. Estimates for ϕ are very often close to one, so we do not show them. Also μ is uninteresting, office hours being warmer while nights are colder, in a predictable fashion. In Figure 4, both ρ and σ look substantially time-varying. Nights and weekends are eventless, with little spatial correlation, except for the first few days and Sunday 14 in the evening. As to σ, shocks can be seen early in the morning as the office opens.

4 Final remarks

Adaptive filtering is a natural alternative to time-constant models when the obsolescence of estimates is suspected. In our example, the approach was even more useful as the parameters were seen to vary *within* days. Instead of resorting to complicated models, it may be advisable to keep things simple, as the complexities of both estimation and prediction are affected. Computational easiness also leaves room for self-tuning of systems with limited computing capabilities.

While we call for adaptive filters, their good behavior should not be taken for granted, though. Stability, both numerical and with respect to initial conditions, can be an appealing property to look for in filters. It may be worth considering multiple filter types and instances at least as a backup, or during warm-up, in the light of their complementary advantages.

Acknowledgements This work was supported by the Competence Centre ASSIC - Austrian Smart Systems Integration Research Center, co-funded by the Austrian Federal Ministry for Transport Innovation and Technology (BMVIT), the Austrian Federal Ministry for Education, Science and Research (BMWFW) and the Austrian Federal Provinces of Carinthia and Styria within the Competence Centres for Excellent Technologies Programme (COMET). The research of Michele Lambardi di San Miniato was supported by the European Social Fund (Investimenti in favore della crescita e dell'occupazione, Programma Operativo del Friuli Venezia Giulia 2014/2020) - Programma specifico 89/2019 - Sostegno alla realizzazione di dottorati e assegni di ricerca, operazione PS 89/2019 ASSEGNI DI RICERCA - UNIUD (FP1956292002, canale di finanziamento 1420_SRDAR8919).

References

1. Buonadonna, P., Gay, D., Hellerstein, J.M., Hong, W., Madden, S.: TASK: Sensor Network in a Box. IEEE (2005). For Intel Lab Data see: `http://db.csail.mit.edu/labdata/labdata.html`
2. Haykin, S.: Adaptive Filter Theory. Pearson (2014)
3. Leskovec, J. Sarkar, P., Guestrin, C.: Modeling link qualities in a sensor network. Informatica (2005)
4. Mariella, L., Tarantino, M.: Spatial temporal conditional auto-regressive model: A new autoregressive matrix. Austrian J. Stat. (2010)
5. Toulis, P., Airoldi, E.M.: Asymptotic and finite-sample properties of estimators based on stochastic gradients. Ann. Statist. (2017) doi:10.1214/16-AOS1506

Bayesian structural learning for Latent Class Model with an application to Record Linkage.

Un approccio Bayesiano al problema di structural learning in Modelli a Classi Latenti con un'applicazione al Record Linkage.

Davide Di Cecco

Abstract We consider an extended family of Latent Class models which relaxes the local independence assumption by modeling additional dependencies via graphical models. We propose a Bayesian model averaging procedure to avoid the difficulties of model selection within this family and learn the dependence structure from the data. We show that, by focusing on decomposable dependence graphs, we can design two Gibbs–based MCMC algorithms to sample over the space of eligible models. The procedure is applied on real data on probabilistic Record Linkage.
Abstract *In questo lavoro consideriamo una famiglia estesa di modelli a Classi Latenti che ammette una struttura di dipendenza più ricca dell'ipotesi classica di indipendenza condizionata. Per stimare le dipendenze aggiuntive proponiamo un model averaging Bayesiano e mostriamo come, focalizzandoci sui modelli caratterizzati da grafi delle dipendenze decomponibili, sia possibile costruire due algoritmi MCMC di tipo Gibbs sampling per lo scopo. Presentiamo un'applicazione della procedura a dati sul Record Linkage probabilistico.*

Key words: Latent Class Models. Bayesian Model Averaging. Record Linkage.

1 Introduction

Latent class models (LCM) constitutes a popular family of models utilized for multiple categorical data in many different areas. The classic conditional independence assumption (CIA) implies the independence of all observed variables conditionally on the latent variable. If we do not question that assumption, model selection within this family is necessarily limited to the identification of the number of latent classes (or components) and, in the recent literature, there has been various contributions

Davide Di Cecco
Università di Roma La Sapienza e-mail: davide.dicecco@uniroma1.it

focusing on Bayesian model selection and model averaging techniques, with several important developments (e.g., [4], [10], [13], [3]).

In the present work we consider a different setting, where the number of latent classes is fixed. If in many applications the latent classes are not given a particular meaning and are used to improve the goodness of fit, there are situations where the different components have a specific interpretations, and their number is fixed in the hypothesis of the model. An example is in the Fellegi–Sunter approach to probabilistic Record Linkage ([5], [6], [7]) where we fix two latent classes: one representing pairs of records that constitute matches, and one for non–matches.

In these cases, relaxing the CIA is of particular importance since we cannot tune the model fitting by changing the number of latent classes. There are various proposals in literature to detect violation of the CIA ([9]), and to overcome its limitation. The latter have also been explored in the context of record linkage: for example, random effects ([1]). We focus on the approach including additional dependencies by modeling them as interaction parameters in a graphical model.

Even if the number of latent classes is fixed, selecting the appropriate model in this extended family is still a difficult task in many cases, especially because estimating all possible models can be computationally demanding. As pointed out by many authors even in the context of Record Linkage (see [12], [1]), ignoring existing violation of the CIA can have disastrous consequences on the estimates. However, at the same time, an ill–defined dependence structure can lead to worse linkage results than the simple CIA model ([14]). Here we propose Bayesian model averaging (BMA) as a natural way to overcome this difficulty. In fact, through a BMA procedure we consider several possible models at once, which appears a suitable strategy to address the structural learning problem.

We show that, by focusing on decomposable models, it is possible to design a simple Gibbs–based MCMC algorithm visiting all eligible alternatives. The restriction to decomposable models, in fact, guarantees a tremendous computational advantage as we can exploit analytical results for the posterior probability of each model, and develop a simple Gibbs sampler to generate values over all selected models. At the same time, this restriction does not appear to limit the usability of the procedure.

The applications on probabilistic Record Linkage typically involves considerable quantity of (possible pairs of) records. As a consequence, in some applications the algorithm occasionally remains stuck over a single model. To avoid this, we present a second algorithm based on a Collapsed Gibbs sampler to enhance the mixing properties of the MCMC.

2 The model

The Fellegi-Sunter model consists of a LCM with k manifest binary variables $\mathbf{Y} = (Y_1, \ldots, Y_k)$ representing the agreement/disagreement on each field, and a binary latent variable X representing the true link status. The model assumes the CIA:

$$P(\mathbf{Y}=\mathbf{y}) = p_{\mathbf{y}} = \sum_{x=0}^{1} p_x \prod_{i=1}^{k} p_{y_i|x}, \tag{1}$$

where $p_{y_i|x}$ indicates the conditional probability $P(Y_i = y_i|X = x)$. The dependence graph of this model is a star–shaped graph, where the node representing X is connected to all other nodes, like the one in Figure 1 on the left. Any additional interaction term, (i.e., any additional arc in the graph), constitutes a relaxation of the CIA.

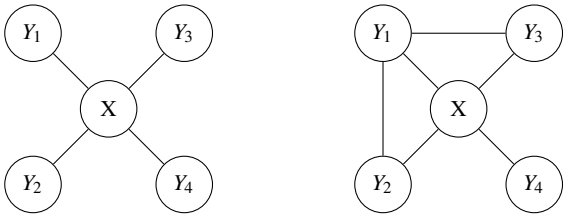


Fig. 1 Two examples of dependence graphs with 4 manifest variables and a latent variable.

Denote the number of observed units presenting profile $\mathbf{y} = (y_1, ..., y_k) \in \{0,1\}^k$ of the manifest variables as $n_{\mathbf{y}}$, and let $\{n_{\mathbf{y}}\}$ represent the whole set of profile counts observed in our data. Let $n_{x,\mathbf{y}}$ be the latent number of units having that profile which belong to latent class x, so that $\sum_x n_{x,\mathbf{y}} = n_{\mathbf{y}}$. Let $N_{x,\mathbf{y}}$ be the random variable associated to $n_{x,\mathbf{y}}$, M be the random variable associated to the model, and Θ_M the set of parameters associated to model M.

2.1 Decomposable models

A model is said decomposable if its dependency graph G is chordal, i.e., if any cycle of G of four or more nodes has an edge connecting two non–adjacent nodes. Let $\{\mathscr{C}_1, \ldots, \mathscr{C}_g\}$ be the maximal cliques of G. Then, it has been proved (see e.g., [2]) that in a decomposable model we have an ordering $\{\mathscr{C}_{\sigma(1)}, \ldots, \mathscr{C}_{\sigma(g)}\}$ of the set, such that, having defined the set of separators $(\mathscr{S}_2, \ldots, \mathscr{S}_g)$ as

$$\mathscr{S}_i = \mathscr{C}_{\sigma(i)} \cap \bigcup_{j=1}^{i-1} \mathscr{C}_{\sigma(j)} \qquad i = 2, \ldots, g,$$

the joint distribution can be written as a product of conditional distributions in the following way:

$$p_G = \prod_{i=1}^{g} p_{\mathscr{C}_i} \left(\prod_{j=2}^{g} p_{\mathscr{S}_j} \right)^{-1} = p_{\mathscr{C}_1} \prod_{i=2}^{g} p_{\mathscr{C}_i|\mathscr{S}_i}, \tag{2}$$

where p over a (sub)graph is to be intended as the (marginal) probability distribution over the variables included in the (sub)graph. [8] utilized a class of priors over the parameters of each decomposable model based on Hyper Dirichlet distribution (see [2]). Such a choice allowed them to give an analytical formula for the posterior probability of each possible decomposable model given the data.

In our models X is latent, so decomposability refers to the augmented dependence graph. However, we can utilize the same analytical formula for the quantity

$$P(\{n_{x,\mathbf{y}}\} \mid M) = \int P(\{n_{x,\mathbf{y}}\} \mid \Theta_M, M)\pi(\Theta_M)\, d(\Theta_M), \tag{3}$$

which can be considered as the marginal augmented likelihood of the model. The integral results in a product of Gamma functions (see [8]).

2.2 Limiting the space of eligible models

Let $\mathscr{M}$ denote the set of eligible models for our BMA. We focus on decomposable models to exploit an analytical formula for the posterior probability of each model in $\mathscr{M}$ at each step of a MCMC algorithm. Obviously, the choice of a prior over the set of models $\mathscr{M}$, as on any other r.v., is subjectively arbitrary. However, we usually just want to exclude some cases, and give a uniform prior on all remaining cases. As a first criterion to exclude some models, we rule out all unidentifiable models. For the case of a binary latent variable, [11] gives a necessary and sufficient condition for the identifiability of any graphical model.

In practice, to utilize the proposed methodology, we have to list out all possible models for a given number of variables. That is, all LCM represented by decomposable graphs containing the star–shaped graph relative to the CIA. Consider the case $k = 4$: in Figure 2.2 we show all possible decomposable identifiable graphs grouped by isomorphism. When $k = 5$ the number of identifiable decomposable models goes up to 445, and when $k = 6$ it goes up to 12830.

Of course, we can restrict $\mathscr{M}$ a priori excluding models that we deem too complex. For example, we can rule out those with too many parameters, or those including interactions of higher order. Higher order interactions in fact, are more difficult to interpret, and, in our experience, lead to instability/slower convergence of the algorithm. As an example, if we exclude all models with 4-cliques, we are left with 285 cases when $k = 5$ and 2925 when $k = 6$.

3 The Gibbs sampler

The t–th iteration of our Gibbs sampler consists of the following steps:

1) sample $n^{(t)}_{x,\mathbf{y}}$ from $\pi(N_{x,\mathbf{y}} \mid M, \Theta_M, N_{\mathbf{y}})$, for all $\mathbf{y}$,

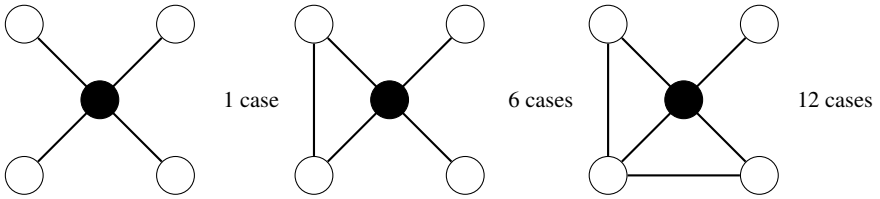


Fig. 2 Identifiable decomposable graph models with 4 manifests (empty nodes) and a latent (black node) grouped by classes of isomorphism

$$N_{x,\mathbf{y}}^{(t)} \sim Mult\left(n_{\mathbf{y}}, \{p_{x|\mathbf{y}}^{(t)}\}_{x=1,\ldots,m}\right) \quad \text{where} \quad p_{x|\mathbf{y}} = \frac{p_{x,\mathbf{y}}}{\sum_x p_{x,\mathbf{y}}},$$

and where the $p_{x,\mathbf{y}}$ are calculated according to the current value of M and Θ_M;

2) sample from $\pi(M, \Theta_M \mid N_{x,\mathbf{y}}) = \pi(\Theta_M \mid M, N_{x,\mathbf{y}})\, \pi(M \mid N_{x,\mathbf{y}})$. That is,

- calculate the posterior probability of each eligible decomposable model in $\mathscr{M}$ conditional on the current value of X, by calculating the quantity

$$\int P(\{n_{x,\mathbf{y}}\} \mid \Theta_M, M)\pi(\Theta_M \mid M)\, d(\Theta_M),$$

  for each model and then normalizing over $\mathscr{M}$. Then sample $M^{(t)}$ from $\pi(M \mid n_{x,\mathbf{y}})$.

- sample all parameters $\Theta_M^{(t)}$ from their posterior conditional distribution $\pi(\Theta_M | N_{x,\mathbf{y}}, M)$, which is a product of Dirichlet distributions.

The algorithm can be used for model selection by simply inspecting the generated values of M, as the relative frequency of each model constitutes an estimate of its posterior probability. Or, it can be used for BMA, by evaluating the posterior distribution of some quantity of interest averaged across the various models. For example, the number of links for each linkage profile $\mathbf{y} \in \{0,1\}^k$.

We also propose a Collapsed Gibbs Sampler (CGS) (see also [13]) which at each iteration sample a model M and a latent configuration $\{n_{x,y}\}$ integrating out the models parameter, to enhance the mixing properties of the MCMC, so to avoid the algorithm being stuck in certain subregions of the posterior distribution.

As we have seen, we can calculate analytically the quantity (3) for any model M and any configuration $\{n_{x,y}\}$ consistent with the observed $\{n_y\}$. Then, we can propose a candidate new configuration $\{n_{x,y}\}^*$, and accept or reject it in a Metropolis-Hastings step.

4 Application

We applied our procedure to a dataset reported in [14] of linked records from patient registries of two hospitals. According to the chosen blocking scheme, the dataset has 590000 record pairs. The number of correct links for each observed linkage profile have been obtained by manual review and are reported in the cited paper. The proposed algorithms took about a minute to run 2 millions of iterations. We do not show here the detailed results for space limitation. However, we considered the posterior averaged distribution of the number of links by profile, and, in all cases, the correct number were included in the 95% credibility intervals. Our results confirmed those obtained in [14] in assessing the inadequacy of the CIA model, and detecting a relevant interaction between (agreement on) telephone number and zip code as linkage variable, as all models with the highest posterior probability included that additional interaction.

References

1. J. Daggy, H. Xu, S. Hui, and S. Grannis. Evaluating latent class models with conditional dependence in record linkage. *Statistics in medicine*, 33(24):4250–4265, 2014.
2. A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.
3. N. Dean and A. E. Raftery. Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11, 2010.
4. D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.
5. I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
6. M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.
7. M. D. Larsen and D. B. Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41, 2001.
8. D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.
9. D. L. Oberski, G. H. van Kollenburg, and J. K. Vermunt. A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3):267–279, 2013.
10. Jia-Chiun Pan and Guan-Hua Huang. Bayesian inferences of latent class models with an unknown number of classes. *Psychometrika*, 79(4):621–646, 2014.
11. E. Stanghellini and B. Vantaggi. Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A):1920–1937, 2013.
12. M. Tromp, N. Méray, A.C.J. Ravelli, J.B. Reitsma, and G.J. Bonsel. Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *Journal of the American Medical Informatics Association*, 15(5):654–660, 2008.
13. A. White, J. Wyse, and T. B. Murphy. Bayesian variable selection for latent class analysis using a collapsed gibbs sampler. *Statistics and Computing*, 26(1-2):511–527, 2016.
14. Huiping Xu, Xiaochun Li, Changyu Shen, Siu L Hui, and Shaun Grannis. Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *The Annals of Applied Statistics*, 13(3):1753–1790, 2019.

Multilevel Latent Class modelling to advise students in self-learning platforms: an application in the context of learning Statistics

Modello multilivello a classi latenti per fornire raccomandazioni agli studenti sulle piattaforme di auto-apprendimento: un'applicazione nell'ambito dell'apprendimento della Statistica

Fabbricatore R., Bakk Z., Di Mari R., de Rooij M., Palumbo F.

Abstract The growing interest in using technology in education has brought new challenges for data analysis, especially when the main aim is providing students with an appropriate feedback on their latent ability level. Any *recommender system* should be able to properly account for the multidimensional nature of such latent trait, while correctly modelling the relationships between variables and individuals at possibly different levels. In this perspective, we focus on a non-standard application of multilevel latent class analysis in the context of learning Statistics. The goal is to detect homogeneous groups of students according to their level of ability, concurrently accounting for the hierarchical structure of the data.

Abstract *Il crescente interesse per l'uso della tecnologia nell'istruzione ha generato nuove sfide per l'analisi dei dati, soprattutto quando il principale obiettivo è fornire agli studenti un feedback appropriato sul proprio livello di abilità latente. Ogni sistema di raccomandazioni dovrebbe essere in grado di considerare la natura multidimensionale di questo tratto latente, identificando al contempo le relazioni tra variabili e individui a possibili differenti livelli. In tal senso, il presente contributo si propone di illustrare l'applicazione di un modello multilivello a classi latenti nell'ambito dell'apprendimento della statistica. L'obiettivo consiste nel rilevare gruppi omogenei di studenti in base al livello di abilità tenendo conto contemporaneamente della struttura gerarchica dei dati.*

Key words: Self-learning platform, Multilevel Latent Class model, Learning Statistics

---

Fabbricatore R., Department of Social Sciences, University of Naples Federico II, e-mail: rosa.fabbricatore@unina.it · Bakk Z., Department of Methodology and Statistics, Leiden University, e-mail: z.bakk@fsw.leidenuniv.nl · Di Mari R., Department of Economics and Business, University of Catania, e-mail: roberto.dimari@unict.it · de Rooij M., Department of Methodology and Statistics, Leiden University, e-mail: rooijm@fsw.leidenuniv.nl · Palumbo F., Department of Political Sciences, University of Naples Federico II, e-mail: fpalumbo@unina.it

1 Introduction

In recent years, there has been a growing interest in using technology to ensure large-scale learning achievements [6]. Self-learning platforms, for example, constitute a very flexible learning environment, which enhances support to students by means of a constant feedback on their ability level [2]. Two main issues emerged concerning (1) how to evaluate learners' proficiency, and (2) the role of psychometric theories and statistical models in analyzing responses - and profiling students accordingly.

Regarding the first issue, a challenging task consists of assessing students' competencies in a multidimensional way, accounting for their intrinsic complexity. In this vein, the Dublin descriptors [4] represent meaningful dimensions used to assess the ability levels a student has achieved with respect to one or more specific topics. More specifically, they define five learning objectives: knowledge and understanding, knowledge application, judgement (critical) ability, and communications and learning skills. Given our focus on learning achievement via technological platforms, only the first three descriptors will be taken into account - i.e., Knowledge (K), Application (A), and Judgement (J).

From a practical point of view, learners' proficiency evaluation is typically based on responses to a set of questions, such as multiple-choice tests, with a different level of difficulty. These questions represent manifest indicators referring to several unobservable abilities (multidimensional latent construct). Notably, moving to the second issue, latent variable models [7] have a key role in the assessment of students' ability, and the detection of homogeneous groups based on their ability level. As such, they are an ideal tool for developing tailored recommendations/remedies from self-learning platforms.

The present work is motivated by an empirical application in the context of learning Statistics. The data include answers to questions (items) covering different Statistic–related topics from students of an introductory course in Statistics. Item-responses are gathered by means of the Moodle survey platform during the course, and will be analysed exploiting the features of multilevel latent class analysis [8].

The remainder of the contribution is structured as follows: Section 2 provides details about the data and the survey procedure, Section 3 describes the central details of the multilevel LC model, Section 4 presents the main results, and Section 5 reports some conclusions, along with possible future developments.

2 Data and procedure

The study involved $N = 202$ Italian students, enrolled in the first year of the psychology course at the University of Naples Federico II, attending the introductory Statistics course. Concerning general descriptive characteristics, attendees were predominantly female (83.6%), and their ages ranged between 18 and 43 (mean = 19.7, sd = 2.77).

Data collection consisted of three waves, each focusing on different statistical topics: descriptive statistics, graphs, tables, and Gaussian distribution (Time 1), probability and random variables (Time 2), hypothesis testing and bivariate statistics (Time 3). For each wave, students were asked to respond to 30 multiple-choice questions, equally divided into K, A, and J, which had four answer options and three different response scores: totally correct answers received two credits, and were scored as 2; partially correct answers received one credit, and were scored as 1; wrong answers received no credit, and were scored as 0. Blank responses were considered missing values.

During data collection, some students dropped out. More specifically, $n = 166$ students remained at Time 2, and $n = 126$ at Time 3.

3 Statistical analysis: Multilevel Latent Class model

The multilevel latent class model is part of the most general multilevel latent variable modeling framework, which allows managing hierarchical data structures with either discrete or continuous latent variables at each level [8]. Herein, we defined a two-level data structure: time–points (low–level/Level 1) clustered within individuals (high–level/Level 2). In particular, at Level 1, three discrete latent variables (K, A, J), conditional on Level 2 class membership, are measured by ordinal indicators with categories $j = 0, 1, 2$; at higher level, another discrete latent variable clusters students based on their likelihood to be in one of the lower–level groups for each Dublin descriptor. Following previous analyses on the same data set, the number of latent classes was set to 3 for both Level 1 and Level 2 discrete latent variables. Moreover, to model potential students' ability change over time (namely for different statistical topics), we added time dummies at level 1.

Formally, let Y_{sdk} denotes the response of student s on item k measuring the dimension d (with $k = 1, \ldots, K$ and $d = 1, \ldots, D$). Whereby X_{sd} indicates the d-th latent class variable at Level 1, i is an index of latent class membership, with I being the total number of latent classes. Similarly, let W_s be the discrete latent variable at Level 2, taking value $m = \{1, \ldots, M\}$.

In the specification we focus on, the manifest distribution of the entire response vector $\mathbf{Y}_{sd} = (Y_{sd1}, Y_{sd2}, \ldots, Y_{sdK})$ with realisation $\mathbf{y}_{sd}$ can be expressed, for each dimension, as:

$$P(\mathbf{Y}_{sd} = \mathbf{y}_{sd}) = \sum_{i=1}^{I} P(X_{sd} = i) P(\mathbf{Y}_{sd} = \mathbf{y}_{sd} | X_{sd} = i), \tag{1}$$

where $\sum_{i=1}^{I} P(X_{sd} = i) = 1$, and $P(\mathbf{Y}_{sd} = \mathbf{y}_{sd} | X_{sd} = i) = \prod_{k=1}^{K} P(Y_{sdk} = j | X_{sd} = i)$ due to the *local independence* assumption. We parametrise the response probabilities by means of the following continuation–logit logistic regressions:

$$P(Y_{sdk} = j|X_{sd} = i) = \frac{\exp\sum_{l=1}^{j}(\beta_{li}^{k})}{\sum_{r=0}^{q}\exp\sum_{l=1}^{r}(\beta_{li}^{k})}, \text{ with } j = 1,2. \tag{2}$$

It is worth noting that response probabilities are assumed to be independent of Level 2 latent variable, and covariates.

On the other hand, lower–level class membership probabilities depend on both high-level class membership (i.e., W_s latent variable), and time. Thus, let $\mathbf{Z}_s$ be the vector of covariates, including the time dummies. Multinomial logistic regressions can be used to parametrise conditional class membership X_{sd} as follows:

$$P(X_{sd} = i|W_s = m, \mathbf{Z}_s) = \frac{\exp(\gamma_{0im} + \boldsymbol{\gamma}_{im}\mathbf{Z}_s)}{\sum_{r=1}^{I}\exp(\gamma_{0rm} + \boldsymbol{\gamma}_{im}\mathbf{Z}_s)}, \tag{3}$$

where $\boldsymbol{\gamma}_{im}$ is the vector of regression coefficients, with $i = 2,\ldots,I$ and $m = 1,\ldots,M$. The full model specification can be visually represented as in Figure 1.

The estimation process was performed through Latent GOLD 6.0 [9] that exploits an Expectation-Maximization (EM) algorithm to obtain maximum likelihood (ML) parameter estimates [8].

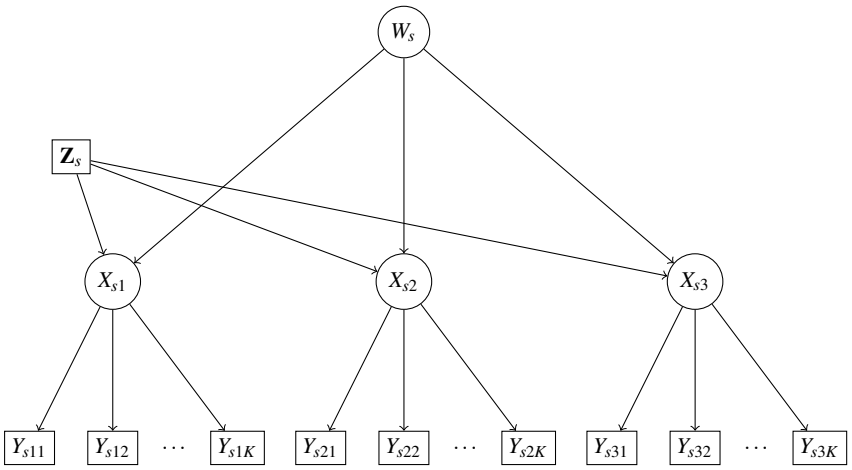


Fig. 1 Multilevel Latent Class model specification for the current application. Note that X_{s1}, X_{s2}, and X_{s3} correspond to the Level 1 latent variables Knowledge, Application, and Judgement, respectively. Each of them was measured by $K = 30$ indicators ($Y_{sd1},\ldots,Y_{sdK}$), 10 for each time point. The Time variable was considered as nominal and was added as a covariate ($\mathbf{Z}_s$) affecting Level 1 class membership probabilities. Finally, Level 2 latent variable $\mathbf{W}_s$ defined groups of students based on their likelihood to be in one of the Level 1 latent classes for each Dublin descriptor (K, A, J).

4 Results

Overall model's summary statistics show no general estimation problems. The entropy-based R^2 reveals a high separation between latent classes for all the discrete latent variables: $R^2_{ENTR} = 0.86$ for Level 2 group variable, $R^2_{ENTR} = 0.77$ for Knowledge, $R^2_{ENTR} = 0.75$ for Application, and $R^2_{ENTR} = 0.74$ for Judgement latent variables at Level 1.

Item conditional response probabilities can be used to characterise the classes of the discrete latent variables at Level 1. Low-, medium-, and high-level ability classes were found for each Dublin descriptor. Regarding Application, the largest class (Class 2) was constituted by high performers, and comprised 37% of the students, followed by Class 3 with a medium level of ability - with 34% of the sample. Conversely, for Judgement, the largest class (Class 1) was that with low performers (49%), followed by the high-performer class - i.e. Class 2, embracing the 36% of the students. Instead, the Knowledge dimension presented more balanced groups, with percentages equal to 34% (Class 1), 35% (Class 3), and 31% (Class 2) for low-, medium-, and high-performance classes, respectively.

Concerning the effect of the Level 2 latent variable, Table 1 shows parameter estimates for the low-level class proportion, conditional on higher-level class membership. We observe (Table 1) that the discrete latent variable at Level 2 detected two well-defined groups of high- (Group 1) and low- (Group 2) performer students in all the Dublin descriptor domains, as well as a third group (Group 3) with moderate ability level in Knowledge and Application but a greater uncertainty related to the Judgement dimension. Specifically, for the latter class, we report almost equal proportions for medium and low levels of ability.

Table 1 Low-level class proportion conditioned to the higher-level class membership. The weight of latent classes at Level 1 and Level 2 was also added.

	Knowledge			**Application**			**Judgement**			
Level 2 class	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*	*Weight*
1	0.01	0.71	0.28	0.05	0.84	0.11	0.16	0.82	0.02	*0.43*
2	0.95	0.01	0.04	0.82	0.01	0.17	0.97	0.01	0.02	*0.30*
3	0.21	0.01	0.78	0.09	0.01	0.90	0.47	0.02	0.51	*0.27*
Weight	*0.34*	*0.31*	*0.35*	*0.29*	*0.37*	*0.34*	*0.49*	*0.36*	*0.15*	

Finally, the analysis revealed a significant effect of time on Knowledge ($p = 0.01$), and Judgement ($p < 0.01$), but a non-significant one on Application ($p = 0.24$). More specifically, students' performance in Knowledge was higher in Time 2 and Time 3 with respect to Time 1; wheres students' critical ability in Statistics increased in Time 2 and decreased in Time 3. All in all, the results highlighted a great difficulty of students in gathering and evaluating information to exert appropriate judgement in hypothesis testing and bivariate statistics. In addition, we report significant inaccuracies and vagueness in the acquisition of theoretical concepts related to descriptive statistics, graphs, tables, and normal distribution.

5 Conclusion

In order to analyse data from self-learning platforms, the present contribution proposes a novel classification strategy exploiting non-parametric multilevel latent variable models. On a practical level, the students' clustering can be straightforwardly used to develop tailored recommendations, and remediations for each group. Moreover, model parameters can also be employed for prediction of class membership of new individuals according to their response vector.

Although we have shown the merit of multilevel latent class analysis in our motivating empirical problem, their potential usefulness is much broader. Further extensions might be conceived, with the goal of bringing the latent underlying structure of data into light. For instance, adding a level in the model to define K, A, and J as continuous latent variables allows to define item response probabilities according to the Item Response Theory framework [5]. Subsequently, a comprehensive evaluation of the effect of item characteristics on students' response probabilities could be performed.

Moreover, it could also be interesting to investigate the effect of demographic and psychological variables (e.g., statistical anxiety or attitude towards statistics) on class membership probabilities. In this respect, several studies (see for example [1, 3]) reported that psychological factors assume a relevant role in students' achievement in Statistics, especially for those enrolled in non-STEM degree courses (where STEM is the abbreviation for Science, Technology, Engineering, and Mathematics). Future works might also address these issues.

References

1. Chiesi, F., Primi, C.: Cognitive and non-cognitive factors related to students' statistics achievement. Stat. Educ. Res. J. **9**, 6–26 (2010)
2. Fabbricatore, R., Parola, A., Pepicelli, G., Palumbo, F.: A latent class approach for advising in learning statistics: implementation in the ALEAS system. In: Limone, P., Di Fuccio, R. (eds.) Proceedings of the First Workshop on Technology Enhanced Learning Environments for Blended Education -The Italian e-Learning Conference 2021, 2817. CEUR Workshop Proc. (2021)
3. Ghani, F.H.A., Maat, S.M.: Anxiety and achievement in statistics: A systematic review on quantitative studies. Creat. Educ. **9**, 2280–2290 (2018)
4. Gudeva, L.K., Dimova, V., Daskalovska, N., Trajkova, F.: Designing descriptors of learning outcomes for higher education qualification. Procedia Soc Behav Sci **46**, 1306–1311 (2012)
5. Hambleton, R.K., Swaminathan, H.: Item response theory: Principles and applications. Kluwer-Nijhoff, Boston (1985)
6. Khatun, R.: Rapidly changing globalized economy and its impact on education in the era of globalization. Res. Rev. Int. J. Multidiscip. **3085**, 1197–1200 (2019)
7. Skrondal, A., Rabe-Hesketh, S.: Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Chapman and Hall/CRC (2004)
8. Vermunt, J.K.: Multilevel latent class models. Sociol. Methodol. **33**, 213–239 (2003)
9. Vermunt, J.K., Magidson, J.: Upgrade manual for Latent GOLD 6.0. Statistical Innovations Inc (2020)

Latent Markov models with associated mixed responses

Modelli Markoviani latenti con risposte associate e di natura mista.

Alfonso Russo and Alessio Farcomeni

Abstract We derive a multivariate latent Markov model to analyse associate responses of mixed nature, including individual covariates in the manifest distribution by suitable link functions. The residual dependency structure in simultaneous outcomes is captured by a Gaussian copula. Outcomes can be any mix of continuous, discrete, ordinal variables. The framework introduced allows for the possibility of modeling multivariate panel data, with separate and distinct margins, without requiring independence assumptions. Inference is conducted by means of an EM algorithm.

Abstract *In questo articolo, si propone un modello Markoviano latente per l'analisi di variabili risposta correlate e di natura mista. E' possibile includere covariate individuali per mezzo di appropriate funzioni link. La struttura di dipendenza nelle osservazioni è gestita con una copula Gaussiana. La procedura inferenziale è condotta tramite un algoritmo EM.*

Key words: Copula; Mixed responses; Panel data

1 Introduction

Latent Markov (LM) models for panel data provide a flexible framework to analyse univariate and multivariate responses. They can be seen as mixed models where random effects are discrete and their evolution over time is Markovian ([9],[3]). They are based on a latent discrete random variable U_{it} that captures the dynamic

Alfonso Russo
Department of Economics and Finance, University of Rome "Tor Vergata",
e-mail: alfonso.russo@uniroma2.it
Alessio Farcomeni
Department of Economics and Finance, University of Rome "Tor Vergata",
e-mail: alessio.farcomeni@uniroma2.it

unobserved heterogeneity and is assumed to follow a first-order Markov chain with $k \in \mathcal{N}$ latent masses. Response variables are modelled with assumptions of *local independence*, i.e. each outcome is independent of its history and other outcomes conditionally on covariates and the latent process. Several extensions have already been proposed in the literature, see [4] for a general review and for details on the inclusion of individual covariates. Some contributions have dealt with the issue of residual dependency structures for the observations, particularly [1] formulate a class of LM models where categorical response variables are affected by contemporary dependence. The case of mixed outcomes is cumbersome, and to the best of our knowledge the only viable approach at this time involves independence assumptions (e.g., [2]).

In this paper, we develop an extension to the classical LM formulation that aims at relaxing the assumption of local independence in the presence of outcomes that can be a mix of continuous, categorical, binary, and ordinal variables. Specifically, the dependency structure is captured by copula functions.

2 Background

We start by formalizing copulas, following [8] as a general reference. A d-dimensional copula, $\mathscr{C} : [0,1]^d \to [0,1]$, is a cumulative distribution function with uniform marginals $u_1, \ldots, u_d$. We write $\mathscr{C}(\boldsymbol{u}) = \mathscr{C}(u_1, \ldots, u_d)$. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ with continuous and increasing marginals $F_{X_1}, \ldots, F_{X_d}$. We can specify the joint distribution of $F_{X_1}(X_1 \mid \boldsymbol{\theta_1}), \ldots, F_{X_d}(X_d \mid \boldsymbol{\theta_d})$ as a copula $\mathscr{C}_{\boldsymbol{X}}$ of the form:

$$\begin{aligned} \mathscr{C}_{\boldsymbol{X}}(u_1, \ldots, u_d) &= \mathbb{P}\left[F_{X_1}(X_1 \mid \boldsymbol{\theta_1}) \leq u_1, \ldots, F_{X_d}(X_d \mid \boldsymbol{\theta_d}) \leq u_d\right] \\ &= F_{\boldsymbol{X}}\left[F_{X_1}^{-1}(u_1 \mid \boldsymbol{\theta_1}), \ldots, F_{X_d}^{-1}(u_d \mid \boldsymbol{\theta_d})\right] \end{aligned} \tag{1}$$

from which, according to Sklar's Theorem, follows immediately that

$$F(x_1, \ldots, x_d) = \mathscr{C}_{\boldsymbol{x}}[F_1(x_1 \mid \boldsymbol{\theta_1}), \ldots, F_d(x_d \mid \boldsymbol{\theta_d})] \tag{2}$$

for all $x_i \in [-\infty, +\infty]$, $i = 1, \ldots, d$. Its density takes the usual form:

$$c_{\boldsymbol{X}}(u_1, \ldots, u_d) = \frac{\partial^d \mathscr{C}_{\boldsymbol{X}}(u_1, \ldots, u_d)}{\partial u_1, \ldots, u_d} \cdot f_{X_1}(x_1 \mid \boldsymbol{\theta_1}), \cdot \ldots \cdot f_{X_d}(x_d \mid \boldsymbol{\theta_d}) \tag{3}$$

where $u_i = F_{X_i}(x_i \mid \boldsymbol{\theta_i}). \forall\, i = 1, \ldots, d$
Among several parametric alternatives (t, Gumbel, Clayton), we choose Gaussian copulas to model the response variables, as usually there are very few of them. Following [6], in this work we make use of *mid*-distribution functions, a modification of the standard CDF that plays an important role when discrete quantiles are at stake. Unlike quantiles, mid-quantiles have good theoretical properties even in the

presence of ties ([7]). For a generic random variable Y with CDF $F_Y(y)$, the function

$$G_Y(y) \equiv \mathbb{P}(Y \leq y) - \frac{1}{2}\mathbb{P}(Y = y)$$

is the mid-CDF of Y. If Y is discrete, $G_Y(y)$ is a step function, a downward-shifted version of $F_Y(y)$. If Y is instead continuous, $G_Y(y)$ reduces simply to $F_Y(y)$ since $\mathbb{P}(Y = y) = 0$.

3 The model

Let Y_{itr} denote the r-th endpoint of interest, measured for the i-th subject at time t; with $r = 1,\ldots,d$, $t = 1,\ldots,T_i$, and $i = 1,\ldots,n$. We also assume there exists a discrete unidimensional latent variable U_{it} with support in $\{1,\ldots,k\}$, where k is known. The latent variable is assumed to evolve over time according to a homogeneous first-order Markov chain, with initial probabilities $\Pr(U_{i1} = u) = \pi_u$, collected in a vector $\boldsymbol{\pi}$ and transition probabilities $\Pr(U_{it} = v|U_{i,t-1} = u) = \pi_{uv}$, collected in a transition matrix $\boldsymbol{\Pi}$. For the marginal distributions, conditionally on U_{it}, we assume a natural exponential family

$$p(Y_{itr}|U_{it} = u, \boldsymbol{\eta}, \psi_r) = \exp\{(Y_{itr}\eta_{utr} - c(\eta_{utr}))/(a(\psi_r) - b(Y_{itr}, \psi_r))\}, \quad (4)$$

where $p(\cdot)$ can either be a PDF or PMF. Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known. The parameter of interest is η_{utr}, while ψ_r is a nuisance parameter. Homogeneity assumptions on ψ_r can easily be relaxed. The regression model will usually be specified after reparameterization through a known link function $g(\cdot)$ as

$$h(\eta_{utr}) = \alpha_{ru} + \boldsymbol{\beta}_r' X_{itr}, \quad (5)$$

where X_{itr} is a vector of time-subject-outcome specific covariates. In (5) only the intercept depends on U_{it}, as customary with latent Markov models. Also this assumption can be easily relaxed. Usually, a conditional independence assumption is put forward so that Y_{itr} is independent of Y_{itq} for all $q \neq r$, conditionally on U_{it}. We relax this assumption through a Gaussian copula model, which requires an unstructured correlation matrix $\boldsymbol{\Sigma}$ as additional parameter. Let $\Phi^{-1}(\cdot)$ denote the quantile function of a univariate standard Gaussian distribution. We assume that $\boldsymbol{Y}_{it}$ has density, conditionally on $U_{it} = u$,

$$\boldsymbol{c}(\boldsymbol{Y}_{it} \mid U_{it} = u) = \frac{\phi_{\boldsymbol{\Sigma}}(\Phi^{-1}(m_{it1u}), \Phi^{-1}(m_{it2u}), \ldots, \Phi^{-1}(m_{itdu}))}{\prod_{r=1}^{d} \phi(\Phi^{-1}(m_{itru}))} \prod_{r=1}^{d} g_r(Y_{itr} \mid U_{it} = u, \boldsymbol{\eta}, \psi_r),$$

with $\phi(\cdot)$ being the density function of a univariate standard normal, and $\phi_{\boldsymbol{\Sigma}}$ the density of a multivariate zero-centered Gaussian with correlation matrix $\boldsymbol{\Sigma}$; $m_{itru} = G_Y(Y_{itr} \mid U_{it} = u, \boldsymbol{\eta}, \psi_r)$ and $g_r(\cdot)$ being the corresponding marginal density/mass

functions governed, together with $G_Y(\cdot)$, by state-specific parameters. Part of the conditional independence assumption still holds as we assume $\boldsymbol{Y}_{it}$ to be independent of $\boldsymbol{Y}_{is}$, for $s < t$, conditionally on U_{it}. Clearly, also Y_{itr} is assumed to be independent of Y_{jsh} for $i \neq j$ as usual.

4 Inference

4.1 Observed likelihood

The observed likelihood can not be conveniently computed directly, as it would involve a telescopic sum over all possible $k^{\max T_i}$ configurations of the latent variables. We adapt, as customary in the latent Markov literature, a simple forward recursion. The computational complexity is linear in $\sum_i T_i$, and the recursion allows to exactly evaluate the observed likelihood. Define $a_{it}(u) = f(\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{it}, U_{it} = u)$. Let, by definition,

$$a_{i1}(u) = \pi_u \mathscr{C}(\boldsymbol{Y}_{it} \mid U_{it} = u).$$

It is possible to show with some algebra that if $T_i > 1$, for $t = 2, \ldots, T_i$,

$$a_{it}(u) = \mathscr{C}(\boldsymbol{Y}_{it} \mid U_{it} = u) \sum_{h=1}^{k} a_{i,t-1}(h)\pi_{hu}.$$

The recursion shall be repeated for $i = 1, \ldots, n$. By definition of $a_{it}(u)$, the observed log-likelihood at the parameter $\boldsymbol{\theta}$ is

$$l(\theta) = \sum_{i=1}^{n} \log\{L_i(\theta)\} = \sum_{i=1}^{n} \log\left(\sum_{u=1}^{k} a_{iT_i}(u)\right).$$

A backward recursion is also implemented to obtain the posterior distributions of each latent states and pairs of consecutive ones. Define $s_{it}(u) = f(\boldsymbol{Y}_{i,t+1}, \ldots, \boldsymbol{Y}_{i,T_i} | U_{it} = u)$. This recursion is initialised with

$$s_{iT_i}(u) = 1 \qquad \forall i$$

and then is based on the following steps

$$s_{it}(u) = \sum_{v=1}^{k} \pi_{uv}\, s_{i,t+1}(v)\, \mathscr{C}(\boldsymbol{Y}_{it} \mid U_{it} = u)$$

to be performed for $t = T_i - 1, \ldots, 1$. For the maximisation of $l(\boldsymbol{\theta})$ an EM algorithm is implemented. We need therefore to introduce the log-likelihood of the complete data, which takes form

$$l^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k} w_{i1u}\log(\pi_u) + \sum_{i=1}^{n}\sum_{t=2}^{T_i}\sum_{u=1}^{k}\sum_{v=1}^{k} z_{ituv}\log(\pi_{uv}) + \sum_{i=1}^{n}\sum_{t=1}^{T_i}\sum_{u=1}^{k} w_{itu}\log\left[\mathscr{C}(\boldsymbol{Y}_{it} \mid U_{it}=u)\right]$$

where w_{itu} is equal to 1 if subject i is in state u at time t and 0 otherwise; while z_{ituv} takes value 1 if subject i at time t moves to state v from u and 0 otherwise.
The EM algorithm iterates the following two steps until convergence.

(a) The *E-step* amounts to computing the conditional expectation of $l^*(\boldsymbol{\theta})$ given the observations and the current value of the parameters. This is actually equivalent to compute the conditional expected value of the variables w_{itu} and z_{ituv} as follows:

$$\mathbb{E}\left[w_{itu}\right] = \frac{a_{it}(u)s_{it}(u)}{\sum_{u=1}^{k} a_{iT_i}(u)}$$

$$\mathbb{E}\left[z_{ituv}\right] = \pi_{uv}\frac{a_{it}(u)s_{i,t+1}(v)}{\sum_{u=1}^{k} a_{it}(u)s_{it}(u)} \times \mathscr{C}(\boldsymbol{Y}_{i,t+1} \mid U_{it}=v)$$

(b) In the *M-step*, the parameters are updated by maximisation of the expected value of $l^*(\boldsymbol{\theta})$ calculated in the previous step. This is performed as follows:

$$\hat{\pi}_u = \frac{\sum_{i=1}^{n}\mathbb{E}\left[w_{i1u}\right]}{\sum_{i=1}^{n}\sum_{h=1}^{k}\mathbb{E}\left[w_{i1h}\right]}$$

$$\hat{\pi}_{uv} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i-1}\mathbb{E}\left[z_{ituv}\right]}{\sum_{i=1}^{n}\sum_{t=1}^{T_i-1}\sum_{v=1}^{k}\mathbb{E}\left[z_{ituv}\right]}$$

The parameters modulating the Gaussian copula are updated separately. First, the matrix $\boldsymbol{\Sigma}$ could in principle be updated by weighted versions of traditional measures like Spearman's rho or Kandall's tau. However, [**?**] prove that, when marginals are not continuous, this strategy might lead to bias. Once the model is fully specified, margin-specific correction factors can be derived to correct these estimators. One step of a Newton-Raphson algorithm is finally implemented to update the remaining free parameters.

5 Conclusions

We have proposed a model for dealing with multivariate outcomes of mixed nature in panel data, when the local independence assumption shall be relaxed. In further work we will apply the described methodology to an original data set about living conditions, poverty, and employment, for selected European countries.

Acknowledgements

The authors are grateful to Prof. Roberto Zelli for advice on a first draft.

References

1. Bartolucci F., Farcomeni ,A., (2009) A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. Journal of the American Statistical Association 104:816–831
2. Bartolucci F., Farcomeni ,A., (2015) A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics 71:80-89
3. Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC Press, Boca Raton, FL
4. Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion). TEST, 23, 433–486.
5. Genest, C., Nešlehová, J., A primer on copulas for count data, ASTIN Bull. 37 (2007) 475–515.
6. Geraci, M., Farcomeni, A., (2022). Mid-quantile regression for discrete responses. Statistical Methods in Medical Research.
7. Ma, Y., Genton, M. G. and Parzen, E. (2011). Asymptotic properties of sample quan- tiles of discrete distributions. Annals of the Institute of Statistical Mathematics 63 227-243.
8. NELSEN, R.B. (1999) An Introduction to Copulas, volume 139 of Lecture Notes in Statistics. Springer, New York.
9. Zucchini, W. and MacDonald, I. L. (2009). Hidden Markov Models for time series: an introduction using R. Springer-Verlag, New York.

Methods for health studies

Beyond the fragility index

Oltre l'indice di fragilità

Piero Quatto and Enrico Ripamonti

Abstract Results of randomized clinical trials (RCTs) can be evaluated through the Fragility Index (FI). Information provided via FI may supplement *p*-values. However, the FI presents with intrinsic weaknesses and shortcomings, which limit its application in biomedicine. Based on the general principles of the analysis of fragility, we propose a new index of strength (SI) of an RCT. This can be obtained in case of either a significant or a non-significant result. SI is straightforward to be calculated, presents with compelling advantages as compared to FI, and solves some of its inherent problems and weaknesses. SI can be obtained as a normalized index, and a threshold for its application can be provided.

Abstract *I risultati degli studi clinici randomizzati (RCT) possono essere valutati attraverso l'indice di fragilità (FI). Le informazioni fornite da FI possono integrare quelle fornite dal p-value. Tuttavia, FI presenta debolezze e carenze intrinseche, che ne limitano l'applicazione nel contesto biomedico. Sulla base dei principi generali dell'analisi della fragilità, proponiamo un nuovo indice di forza (SI) di un RCT. Questo può essere ottenuto in caso di risultato significativo o non significativo. SI è semplice da calcolare, presenta vantaggi convincenti rispetto a FI e risolve alcuni dei suoi problemi e punti deboli intrinseci. SI può essere ottenuto come indice normalizzato e viene suggerita una soglia per la sua applicazione pratica.*

Keywords: Randomized Clinical Trials, Comparing Binomial Proportions, Fragility Index.

[1]Piero Quatto, Department of Economics, Management, and Statistics, University of Milan-Bicocca, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: piero.quatto@unimib.it.

Enrico Ripamonti, Department of Economics and Management, University of Brescia, Brescia, Italy, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: enrico.ripamonti@unibs.it

1. Introduction

The analysis of the fragility of a randomized clinical trial (RCT) is a very controversial issue in the literature. To operationalize the concept of "fragility", Feinstein introduced in the literature the concept of Fragility Index (FI) [1]. The FI has been revised by Walsh et al. in case of a significant result [2]. Since then, its use in the applied medical literature has increased. A relative version of FI has also been proposed, known as Fragility Quotient (FQ) [3]. FI or FQ can also be adopted in case of a non-significant result [4]. Today, there are many applications of FI in several fields of biomedical research [5,6]. The FI has objective weaknesses, and it has been strongly criticized [7]. However, the index is in agreement with one of the recent statements of the American Statistical Association, namely "scientific conclusions and business or policy decisions should be not based only on whether a *p*-value passes a specific threshold" [8]. Thus, it is worth re-considering the analysis of fragility setting it in a broader framework than the current one, so that it could supplement the *p*-value and provide reliable information to researchers.

2. A new index of strength of a randomized clinical trial

2.1 The Fragility Index

In the context of an RCT, the FI can be defined as the minimal number of outcomes that must change their status from "non-event" to "event" to transform a significant result into a non-significant one [2]. The FI is calculated by virtually adding events to the group that contains less events, until the *p*-value p exceeds the threshold α with the use of Fisher's exact test. In case of a non-significant result, the adopted procedure is to subtract events from the group that contains a smaller number of events, until $p \leq \alpha$ is reached. In this case, the FI is generally known as Reverse Fragility Index (RFI). The FQ is obtained as a transformation of FI, namely by dividing FI by the total sample size.

The FI has some objective weaknesses, which can be summarized as follows: (i) it can only be calculated for dichotomous data; (ii) it does not present a threshold above which the result can be considered robust, according to a common terminology in the literature; (iii) it grows as the sample size grows, virtually suggesting that experiments with greater sample size may be less fragile than others; (iv) it is negatively correlated with the *p*-value, therefore large *p*-values are considered fragile and small *p*-values are considered robust.

2.1 Beyond the Fragility Index

Let us consider a general setting of comparison of $\theta_1$and θ_2, indicating the unknown proportions of events in the treatment groups T_1 and T_2, respectively:

$$H_0\colon \theta_1 = \theta_2 \ vs.\ H_1\colon \theta_1 \neq \theta_2. \tag{1}$$

Let us indicate with $\frac{x_1}{n_1}$ and $\frac{x_2}{n_2}$ the proportions of events verified in T_1 and T_2, respectively.
Without losing generality, we assume $\frac{x_1}{n_1} \leq \frac{x_2}{n_2}$ so that $\hat{\theta} = \frac{x_2}{n_2} - \frac{x_1}{n_1}$ is an estimate of $\theta = |\theta_1 - \theta_2|$ and

$$\hat{\sigma}^2 = \frac{x_1}{n_1^2}\left(1 - \frac{x_1}{n_1}\right) + \frac{x_2}{n_2^2}\left(1 - \frac{x_2}{n_2}\right) \tag{2}$$

is a suitable estimate of its variance. We suppose that the difference of proportions can be approximated by a Normal r.v..
We consider the quantity Q satisfying the Equation:

$$2\Phi\left(-\frac{\left|\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q\right|}{\hat{\sigma}}\right) = \alpha \tag{3}$$

where Φ is the distribution function of a Normal r.v. That is to say that, to make the *p*-value equal to the threshold α, it is sufficient to add to x_1 (approximately) Qn_1 events, or, equivalently, to subtract from x_2 (approximately) Qn_2 events, according to the equalities:

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q = \frac{x_1 + Qn_1}{n_1} - \frac{x_2}{n_2} = \frac{x_1}{n_1} - \frac{x_2 - Qn_2}{n_2}.$$

Expression (3) is equivalent to

$$\left|\frac{x_1}{n_1} - \frac{x_2}{n_2} + Q\right| = \hat{\sigma} z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the percentile of order 1 -$\alpha/2$ of the standardized Normal r.v. Z. In particular, for Q we propose the explicit formula

$$Q = \frac{x_2}{n_2} - \frac{x_1}{n_1} - \hat{\sigma} z_{1-\frac{\alpha}{2}} = \left(z_{1-p/2} - z_{1-\alpha/2}\right)\hat{\sigma} \tag{4}$$

where $z_{1-p/2} = \frac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{\hat{\sigma}} = \frac{\hat{\theta}}{\hat{\sigma}}$ and $p = 2\Phi\left(-\frac{\hat{\theta}}{\hat{\sigma}}\right)$ is the two-sided *p*-value for the null hypothesis (1).
If the experiment led to a significant result, Expression (4) takes a positive value, while it takes a negative value in case of non-significant result. More precisely, from

(4) it follows $Q \leq \frac{x_2}{n_2} - \frac{x_1}{n_1}$ in case of significance and $-\hat{\sigma} z_{1-\frac{\alpha}{2}} \leq Q \leq 0$ in case of non-significance.

In light of these results, we can introduce the normalized index SI given by

$$SI = \begin{cases} \dfrac{Q}{\frac{x_2}{n_2} - \frac{x_1}{n_1}} = 1 - \dfrac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\frac{x_2}{n_2} - \frac{x_1}{n_1}} = 1 - \dfrac{z_{1-\frac{\alpha}{2}}}{z_{1-p/2}} & : p \leq \alpha \\ \dfrac{Q}{-\hat{\sigma} z_{1-\alpha/2}} = \dfrac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{-\hat{\sigma} z_{1-\alpha/2}} + 1 = 1 - \dfrac{z_{1-\frac{p}{2}}}{z_{1-\frac{\alpha}{2}}} & : p > \alpha \end{cases} \quad (5)$$

In general, for a fixed p-value, Q decreases as sample sizes grow and, if $p \leq \alpha$,

$$SI \geq 1 - \frac{1}{\sqrt{2}} \Leftrightarrow \frac{\frac{x_2}{n_2} - \frac{x_1}{n_1}}{\hat{\sigma}} \geq \sqrt{2} z_{1-\frac{\alpha}{2}} \Leftrightarrow p \leq 2\Phi\left(\sqrt{2} z_{\frac{\alpha}{2}}\right)$$

according to [9]. Details for $p > \alpha$, further discussion, and examples are in Quatto et al. [10].

3. Conclusion

The FI can be considered as a measure of robustness of an inferential procedure, in the sense that small variations in the data can subvert the *p*-value, modifying it from significant ($p < 0.05$) to non-significant or vice versa. The FI has raised a lot of criticism, although it continues to be used in the applied literature.

The index (5) is a normalized index of strength of RCT, which varies as a function of the *p*-value and for which a threshold can be provided [10], contrary to FI, as highlighted in observation (ii). In its formulation, the index solves many shortcomings of the FI, and can also be extended to the case of unidirectional hypotheses.

References

1. Feinstein AR. The unit fragility index: an additional appraisal of "statistical significance" for a contrast of two proportions. J Clin Epidemiol. 1990;43(2):201–9.
2. Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol. 2014;67(6):622–8.
3. Ahmed W, Fowler RA, McCredie VA. Does sample size matter when interpreting the fragility index? Crit Care Med. 2016;44(11):e1142--e1143.
4. Docherty KF, Campbell RT, Jhund PS, Petrie MC, McMurray JJ V. How robust are clinical trials in heart failure? Eur Heart J. 2017;38(5):338–45.
5. Khan MS, Fonarow GC, Friede T, Lateef N, Khan SU, Anker SD, et al. Application of the reverse fragility index to statistically Nonsignificant randomized clinical trial results. JAMA Netw Open. 2020;3(8):e2012469--e2012469.
6. Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of

optimizing patient care. JAMA Surg. 2019;154(1):74–9.
7. Potter GE. Dismantling the Fragility Index: A demonstration of statistical reasoning. Stat Med [Internet]. 2020; Available from: https://doi.org/10.1002/sim.8689
8. Wasserstein R, Schirm A, Lazar N. Moving to a world beyond “p<0.05.” Am Stat. 2019;73(S1):1–19.
9. Held L. The assessment of intrinsic credibility and a new argument for p< 0.005. R Soc Open Sci. 2019;6(3).
10. Quatto P, Ripamonti E, Marasini D. A new index of strength for a clinical trial. Milan; Mimeo, 2022.

Evaluation of the diagnostic-therapeutic paths for schizophrenic patients through state sequences analysis

Analisi dei percorsi diagnostico-terapeutici dei pazienti schizofrenici tramite tecniche di analisi delle sequenze

Laura Savaré[1,2,3], Giovanni Corrao[3,4], Francesca Ieva[1,2,3]

Abstract To date, there are still few attempts to monitor the quality of mental health care through tools that are easily accessible from databases used for health care. The aim of this work is to analyze the diagnostic-therapeutic pathways, provided to patients diagnosed with schizophrenia, through the analysis of the sequences. This methodology, not yet used in the field of mental health, consists in considering the patient's therapeutic path as a conceptual unit, i.e. a sequence, composed by a succession of different states, which in this case represent the weekly coverage of different treatments. This technique seems to be a valid support for describing complex scenarios, such as the heterogeneity of the paths followed and grouping patients with more similar paths. Therefore, it stands as a precious tool for informing decision making and supporting the definitions of new guidelines in a real-world evidence prespective.

Abstract *Ad oggi, sono ancora pochi i tentativi di monitorare la qualità dell'assistenza sanitaria, in ambito di salute mentale, attraverso strumenti facilmente fruibili dai dati disponibili. Lo scopo del presente lavoro è analizzare i percorsi diagnostico-terapeutici, seguiti dai pazienti schizofrenici, tramite l'analisi delle sequenze. Questa metodologia, non ancora usata in questo contesto, consiste nel considerare il percorso terapeutico del paziente come un'unità concettuale, ovvero una sequenza, composta dalla successione di diversi stati, che in questo caso rappresentano la copertura su base settimanale dei diversi trattamenti. Questa tecnica sembra essere un valido sostengno nel descrivere scenari complessi, come l'eterogeneità dei percorsi seguiti e raggruppare i pazienti con percorsi più simili. Pertanto, rappre-*

[1] MOX, Department of Mathematics, Politecnico di Milano, Milan 20133, Italy
[2] CHDS, Center for Health Data Science, Human Technopole, Milan 20157, Italy
[3] National Centre for Healthcare Research & Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy
[4] Department of Statistics and Quantitative Methods, Laboratory of Healthcare Research & Pharmacoepidemiology, Unit of Biostatistics, Epidemiology and Public Health, University of Milano-Bicocca, Milan, Italy
e-mail: laura.savare@polimi.it, francesca.ieva@polimi.it, giovanni.corrao@unimib.it

senta uno strumento prezioso per supportare il processo decisionale e supportare le definizioni di nuove linee guida in una prospettiva di evidenza del mondo reale.

Key words: analysis of sequences, diagnostic-therapeutic pathways, mental health

1 Introduction

Currently, the vast majority of people with schizophrenia around the world are not receiving mental health care. Schizophrenia is a chronic brain disorder that affects approximately 24 million people or 1 in 300 people (0.32%) worldwide[1]. When schizophrenia is active, symptoms can include delusions, hallucinations, disorganized speech, trouble with thinking and lack of motivation. However, with treatment, most symptoms of schizophrenia will greatly improve and the likelihood of a relapse can be reduced. At least one third of people with schizophrenia experiences complete remission of symptoms [2]. Some people with schizophrenia experience worsening and remission of symptoms periodically throughout their lives, others a gradual worsening of symptoms over time. Most resources for mental health services are inefficiently spent on care within mental hospitals. In Italy, the transition from a hospital to a community-based system started in 1978, with a reform that led to the gradual closing of psychiatric hospitals [3]. The aim of the work is to develop methods to allow the governance to monitor over time, asses the quality and optimize the diagnostic-therapeutic pathways, provided to patients newly taken-into-care for schizophrenic disorders by the National Health Service (NHS) of the Lombardy Region. In particular, there is the need to describe the treatment paths to then evaluate its association with the incidence of negative outcomes.

2 Data

The data used for this study were retrieved from the healthcare utilization databases on the residents of Lombardy, a Region of Italy that accounts for about 16% (10 million) of its population. The Italian population is covered by the National Health Service that provides hospitalization, major diagnostic procedures, and so-called life-saving drugs to all citizens free or almost free of charge. In addition, a specific automated system concerning psychiatric care gathers data from regional Departments of Mental Health (DMHs) accredited by the NHS. This system provides demographic information and diagnostic codes for patients receiving specialist mental healthcare [4].
The target population of this work consisted of all NHS beneficiaries resident in Lombardy, aged 18-40 years, who during the recruitment period (2015-2018) had at least one contact with a mental health service and had a diagnosis of schizophrenia. We then excluded prevalent patients and those with less than 1 year of follow-up.

The patients included into the final cohort were followed for 1 year after the the onset of schizophrenia (i.e, without experiencing death, emigration or hospitalization during the fist year after diagnosis of schizophrenia). During the first year we have collected all psychiatric visits, psycho-social interventions and anti-psychotic dispensed drugs delivered to each patient. The combination of these three interventions make up the optimal treatment path (OTP).

3 Methods

In order to analyze the specific pathways of each patient, intended as a sequence of treatments received, we used the analysis of the *state sequences*.

A sequence is defined as an ordered list of elements, where an element can be a certain status (e.g., employment, health status or coverage of a treatment), a physical object (e.g. base pair of DNA, protein, or enzyme), or an event (e.g. an hospitalization). The positions of the elements are fixed, ordered by elapsed time and they refer to a relative, not absolute, time point (Figure 1). The addressed methods are for sets of sequences each of which are considered as a whole; i.e., as a conceptual unit [5] .

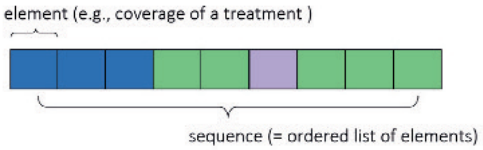


Fig. 1 Sample sequence

In particular, in our case-study, we defined the states of the sequence of each patient on a weekly basis. Based on the OTP, we classified each state (i.e. every week) into four adherence categories, based on how many of the three treatments (i.e, psychiatric visits, psycho-social interventions and anti-psychotic) were dispensed to the patients:

- *Not treated (0 out of 3)*, if in that given week the patient is not covered by any of the three treatments;
- *Low intensity of treatment (1 out of 3)*, if covered by only one treatment;
- *Medium intensity of treatment (2 out of 3)*, if covered by two treatments out of three;
- *High intensity of treatment (3 out of 3)*, if covered by all three treatments.

In particular, we considered for each psychiatric visits a coverage of 2-4 weeks, depending on who provided the visit, for each psycho-social interventions a coverage of 2 weeks and for the anti-psychotic drugs the defined daily dose.

The primary objective of sequence methods is to extract simplified workable information from sequential datasets; that is, to efficiently summarize the data identifying typical patterns and categorizing them into a limited number of groups. This is essentially an exploratory task that consists of computing summary indicators, as well as sorting, comparing and grouping sequences. The resulting groups and real-value indicators may then be submitted to classical inferential methods and serve, for instance, as response variables or explanatory factors for regression-like models.

For the comparison step, we had to determine how sequences should be compared and how the difference between two sequences should be measured. Dissimilarity measures can be classified into measures based on the (minimal) cost of transforming one sequence into the other and those defined as the count of matching attributes. Another interesting distinction is between those that make position-wise comparisons; i.e., that do not allow shifting a sequence or part of it, and those accounting for similar shifted patterns. We decided to build our definition of dissimilarities on counts of common attributes. Let $A(x,y)$ be a count of common attributes between sequences x and y. It is a proximity measure since the higher the counts, the closer the sequences. We derive a dissimilarity measure from it through the following general formula

$$d(x,y) = A(x,x) + A(y,y) - 2A(x,y)$$

where $d(x,y)$ is the distance between objects x and y. The dissimilarity is maximal when $A(x,y) = 0$; i.e., when the two sequences have no common attribute. It is zero when the sequences are identical, in which case we have $A(x,y) = A(x,x) = A(y,y)$. One of the implemented metric is based on the length of $A_S(x,y)$ of Longest Common Subsequence (LCS) [8], which extracts the elements within a group of sequences that occur in the same order, while allowing gaps in between elements, as shown in the Figure 2

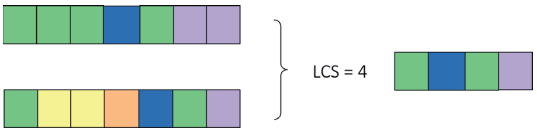


Fig. 2 The Longest Common Subsequence between two sequences

When compared with metrics based on position-wise, the LCS metric reduces distances by accounting for non-aligned matches; i.e., position-shifted similarities. Note that this methodology has a purely sequential and do not take into account any temporal dimensions.

Given the notion of (dis)similaruty, a clustering procedure may be adopted. We used a hierarchical clustering and through the elbow method and silhouette analysis or score techniques we found the optimal number of clusters.

4 Results

Out of the 8'566 patients aged 18-40 years and assisted by the NHS for schizophrenia in the years 2015-2018, 2'739 were incident users. Among these, only 2'329 patients had at least 1 years of observation and were included in the final cohort.
It was observed that during the first year, on average, they were not covered by any treatment for about 5 months, they were covered by 1 or 2 treatments for about 3 months, and for one month they resulted fully covered by all the OTP's treatments. The most stable state is the state in which no treatment is received, i.e. *0 out of 3*, as the probability of remaining in this state, for those who are already there, is 95%. On the other hand, for the other 3 states, the probability of remaining in the same state is 81%, 83% and 87%, respectively. Furthermore, the highest transition rate is from *2 out of 3* to *1 out of 3* and from *3 out of 3* to *2 out of 3* state, that is equal to 0.12. The calculation of the distance through the LCS metric and the study of the optimal number of groups to be considered, as above mentioned in the methods session, allowed us to cluster our patients into 3 groups (as shown in Figure 3):

- 953 patients with *very low intensity* of treatment [1];
- 932 patients with *medium-low intensity* of treatment [2];
- 444 patients with *medium-high intensity* of treatment [3].

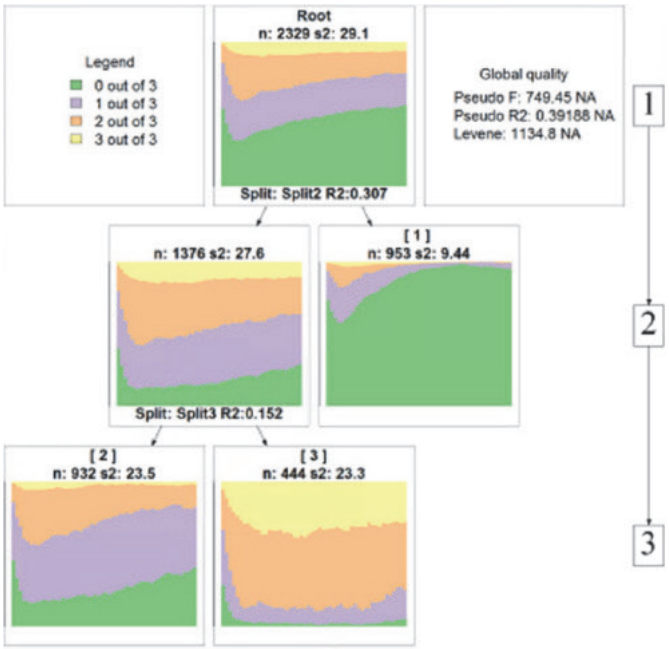


Fig. 3 Hierarchical clustering dendrogram. For each group it is represented the state distribution for each week

5 Conclusions

The analysis of state sequences is useful and effective in this field because it offers a descriptive gain by being able to identify typical patterns. Furthermore, via the inclusion of covariates, it could establish why certain sequence patterns exist and allow the researcher to establish what the effects of a given sequential pattern have on other outcomes. Further developments in these directions will lead to profile the groups according the healthcare rationale through a post-hoc analysis and to capture the potential predictive effect that a specific treatment path has on future negative outcomes. Therefore, sequences analysis could be an intuitive tool supporting providers to assess the quality and effectiveness of the diagnostic-therapeutic pathways provided to these patients and to monitor them over time. Further research could study different definition of the states of the sequences and different metrics used for the distances.

Acknowledgement

References

1. World Health Organization.
   www.who.int/news-room/fact-sheets/detail/schizophrenia
   Accessed: 08/02/2022
2. Harrison, G., Hopper, K., Craig, T., Laska, E., Siegel, C., Wanderling, J. Recovery from psychotic illness: a 15- and 25-year international follow-up study. Br J Psychiatry 2001;178:506-17.
3. Barbui C, Papola D, Saraceno B. The Italian mental health-care reform: public health lessons. Bull World Health Organ 2018; 96: 731–731A
4. Corrao, G., Barbato, A., D'Avanzo, B., Di Fiandra, T., Ferrara, L., Gaddini, A., Monzio Compagnoni, M., Saponaro, A., Scondotto, S., Tozzi, VD., Carle, F., Lora, A.; "QUADIM project", "Monitoring, assessing care pathways (M. A. P.)" working groups of the Italian Ministry of Health. Does the mental health system provide effective coverage to people with schizophrenic disorder? A self-controlled case series study in Italy. Soc Psychiatry Psychiatr Epidemiol. 2021 Jun 16. doi: 10.1007/s00127-021-02114-9. Epub ahead of print. PMID: 34132836.
5. Billari, FC (2001b). "Sequence Analysis in Demographic Research." Canadian Studies in Population, 28(2), 439–458. Special Issue on Longitudinal Methodology.
6. Gabadinho, A., Ritschard, G., Mueller, N.S., Studer M. (2011). Analyzing and visualizing state sequences in R with TraMineR. Journal of statistical software, 40(4), 1-37.
7. Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10: 707–710.
8. Hirschberg, D.S. (1975), 'A linear space algorithm for computing maximal common subsequences', Communications of the ACM 18(6), 341–343.

Optimal timing of bone-marrow transplant in myelodysplastic syndromes through multi-state modeling and microsimulation

Tempo ottimale per il trapianto di midollo osseo nelle sindromi mielodisplastiche attraverso un modello multi stato e di microsimulazione

Caterina Gregorio[1,2], Marta Spreafico[1,3], Francesca Ieva[1,4]

Abstract In myelodysplastic syndromes tailoring the timing of bone-marrow transplant according to patient's genomic profile and characteristics is fundamental since it represents the only curative treatment available in this disease and it can be performed only once. In this work we present a method based on a multi-state model to analytically optimize the timing of the transplant. This method can be used as an alternative to the microsimulation approach based on Monte Carlo. As objective function for the optimization, the Restricted Mean Survival Time was considered together with its first derivative. This decision analysis is the basis for a novel quantitative tool to assist clinicians in their decision process for the treatment of patients affected by this condition.

Abstract *Nelle sindromi mielodispastiche è fondamentare personalizzare il tempo al trapianto di midollo osseo sulla base del profilo genetico del paziente e sulle sue caratteristiche individuali poiché questo è l'unico trattamento curativo attualmente a disposizione e non può essere ripetuto. In questo lavoro presentiamo un approccio basato su un modello multi stato per ottimizare il tempo al trapianto per via analitica. Questo metodo è un'alternativa all'approccio della microsimulazione basato su stima Monte Carlo. Come funzione obiettivo per l'ottimizzazione è stato considerato il Restricted Mean Survival Time e la sua derivata prima. Questa analisi decisionale rappresenta il primo passo per lo sviluppo di un nuovo strumento quantitativo che può assistere i medici nel processo decisionale legato al trattamento dei pazienti affetti da questa patologia.*

Key words: multi-state model, decision analysis, microsimulation, transplant timing, chronic leukemia

[1] MOX – Department of Mathematics, Politecnico di Milano, Milan 20133, Italy
[2] Biostatistics Unit - Department of Medical Science, University of Trieste, Trieste 34100 , Italy
[3] Mathematical Institute, Leiden University, Leiden, The Netherlands
[4]CHDS, Center for Health Data Science, Human Technopole, Milan 20157, Italy
e-mail: caterina.gregorio@polimi.it marta.spreafico@polimi.it francesca.ieva@polimi.it

1 Introduction

The myelodysplastic syndromes (MDS) are a group of tumors affecting haematopoietic stem cells characterized by cytopenia, dysplasia, ineffective haematopoiesis and increased risk of developing acute myeloid leukemia[1]. Nowadays diagnosis and treatments are guided by clinical parameters and biomarkers. However, in the last few years much research has been focused on studying the role of genetic mutations in the progression and prognosis of this disease since it may play a role in therapeutic programs in the future[2]. The only available curative treatment for patients with MDS is allogeneic haematopoietic stem-cell transplantation (HSCT). HSCT can be performed only once and it comes with significance probability of failure because of toxicity and disease relapse. HSCT's success and survival after transplant it is also believed to depend of patients' characteristics as well as their genomic profile. With this in mind, it becomes essential to tailor the timing of HSCT in eligible patients according to individual characteristics and genomics at MDS diagnosis in order to maximize the survival probability in patients. From a statistical point of view, the problem consists in balancing individual pre-HSCT survival probability and post-HSCT in order to maximize patient's overall life-expectancy. The aims of this work are to (1) provide a statistical method to quantitatively assess the individual risk of death before and after the transplant according to different transplant policies with regards to the timing and (2) optimize the HSCT timing by considering the Restricted Mean Survival Time (RMST) and its first derivative.

2 Data

Data comes from a retrospective international cohort study which enrolled patients diagnosed with primary MDS according to 2016 WHO criteria. The data contains demographics, clinical and genomic information at the time of the diagnosis. Moreover, events in the follow-up such as the timing of HSCT (for patients who were transplanted) and the time of death were recorded. 2476 patients were considered in this analysis including 558 who underwent HSCT at some time during the follow-up. Genomic information was summarized using a genomic classification recently proposed for MDS patients[3] which allows to identify patients with a similar mutations profile.

3 Methods

As a first step, we specified a suitable model to describe the clinical problem under study. Using this model, we were then able to tackle the optimization of the HSCT timing following two different methods: one based on microsimulation and one based on an analytic formula hereby proposed.

3.1 Multi-state model

A multi-state model as the one represented in Figure 1 was considered . Let $M(t)$ be the state occupied at time t and let x_i be the vector of baseline covariate information observed for subject i.

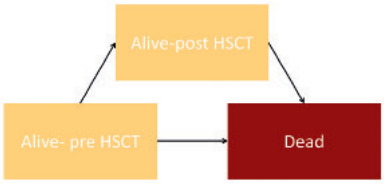


Fig. 1 Multi-state model representation

The model has two "Alive" states corresponding to pre-HSCT and post-HSCT and one absorbing state representing "Dead" status. Specifically, the model corresponds to a standard illness-death model and it is assumed to be semi-markov. The transition from state "pre-HSCT" to death is used to study the natural history of the disease conditionally on baseline covariates relating to disease progression. On the other hand, the transition from "post-HSCT" to death relates to post-transplant survival probability and it depends on the probability of HSCT's success.

3.2 Microsimulation

Similarly to a multi-state model, in a microsimulation model[4, 5] individuals can transit in one state at the time from a sets of mutually exclusive states. We consider now the time to be discrete and finite: $c = \{1, \ldots, c^*\}$ where each time corresponds to a cycle of the model and c^* is the time horizon. For our problem we considered a time cycle of 1 months. The set of possible states at cycle c is denoted by $M(c)$ and $A(c)$ is the set of possible actions available at state $m \in M$ for an individual i. The microsimulation algorithm at each cycle c consists of the following steps:

1. the individual current state at cycle c is recorded;
2. the decision maker chooses an action a;
3. the individual transitions to the next cycle $c+1$ according to a transition probability $P_c(m'|m,a,x_i)$.
4. the individual receives a reward $r_c(m,a)$;

The reward, $r_c(m,a)$, is a real valued function, whereas $P_c(m'|m,a)$ denotes the probability that the individual is in state $m' \in M_{c+1}$, conditionally on state m occupied at time c, the vector of baseline covariates and the action taken at time c. A

policy, $\pi = \{a_1, ..., a_c, ..., a_{c^*}\}$, is defined as the sequence of actions taken, where an action taken is mapping from states to actions, so that $a_c(m) \in A_m$. In our case, the states of the microsimulation correspond to the states of model defined in Section 3.1. The policies that we aimed to compare were π_h: perform HSCT at time $c^h \in \{1, ..., 12\}$ months. Therefore, the policy "perform HSCT at time c^h" induces a probability of transition to state "post-HSCT" at time c^h equal to one minus the probability of dying in c^h for subjects still alive at time c^{h-1} and equal to 0 otherwise. The probabilities of dying at each time c, conditionally on covariates, required for the complete definition of $P_c(m'|m, a, x_i)$ were obtained by the multi-state model. For each policy, performing a simulation based on the previous algorithm for a set of M individuals with covariates x_i allows us to obtain a Monte Carlo estimate of the total reward for the subgroup of patients with characteristics x_i. A reward of 1 was considered if the patient was alive at the end of the cycle and 0 otherwise. As a consequence, the total reward corresponds to an estimate of the RMST over a time horizon of c^* months:

$$\mu^{c^*|x_i} = \sum_{c=0}^{c^*} S(c|x_i) \tag{1}$$

3.3 Analytic Formula

Since our aim is to model the survival probability as a function of the HSCT time, conditionally of baseline covariates, we decompose it using the law of total probability:

$$\begin{aligned} S(t|x_i) &= Pr(T_i > t|x_i) = \\ &= Pr(T_i > t|T > t^h, x_i) Pr(T_i > t^h|x_i) + Pr(T_i > t|T \leq t^h|x_i) Pr(T_i \leq t^h|x_i) \end{aligned}$$

where t^h corresponds to the transplant time. We can now study this formula considering the two possible cases: either time t is before of equal to t^h or is greater than t^h. In the former case we can work on the formula using the definition of conditional probability, whereas in the latter we observe that the probability of surviving at least until t, given that death has occurred before t^h, is 0. We therefore obtain:

$$S(t|x_i) = \begin{cases} Pr(T > t^h|x_i) + Pr(t < T \leq t^h|x_i) = Pr(T > t|x_i) & \text{if } t \leq t^h \\ Pr(T > t|T > t^h, x_i) Pr(T > t^h|x_i) & \text{if } t > t^h \end{cases} \tag{2}$$

Therefore, the RMST becomes:

$$\begin{aligned}\mu^{t^*|x_i} &= \int_0^{t^*} S(u|x_i)du = \\ &= \int_0^{t^h} S(u|x_i)du + \int_{t^h}^{t^*} S(u|x_i)du = \\ &= \int_0^{t^h} Pr(T > u|x_i)du + \int_{t^h}^{t^*} Pr(T > u|T > t^h, x_i)Pr(T > t^h|x_i)du\end{aligned}$$

The RMST can be estimated directly from the model described in Section 3.1 since the probabilities correspond to the multi-state model predicted probabilities of transition.

4 Results

The analysis was stratified according to the genomic group i.e. a separate model was fitted for each genomic group while age and gender were considered as covariates. A Cox type multi-state model was fitted using the `mstate` [7] and `survival` [6] R packages. We applied the methods explained in Section 3.2 and 3.3 in order to obtain the RMST according to different treatment policies. As time horizon for the RMST we considered 36 months. As an example, we reported the results for 57 years old male patients in different genomic groups (Fig.2). Since the microsimulation and the analytic formula leads to the same results, we reported the ones obtained from the analytic formula which has the computational advantage not to require any simulation. Interestingly, it can be observed how most of the variability in the estimated RMST is mostly captured by the patient's genomic profile. To better study the variability related to the timing of HSCT, curves were smoothed using cubic b-splines in order to obtain the first derivative of RMST. It can be observed how latter timing for HSCT seem to be preferred for genomic group A, B, C and D, while the RMST start decreasing after 6 months in genomic group E.

5 Conclusions

Optimizing the time of HSCT in a personalized manner is a non-trivial statistical problem. In this work we presented an analytic formula based on a multi-state model which can be used for the maximization of the RMST. The formula is equivalent to using a microsimulation approach but has a significant advantage in terms of computational effort. Further developments in this direction can lead to a new decision tool for the treatment of patients affected by MDS. Further research could study uncertainty assessment of this method. Moreover, it could be of clinical interest to include biomarkers information in the optimization problem.

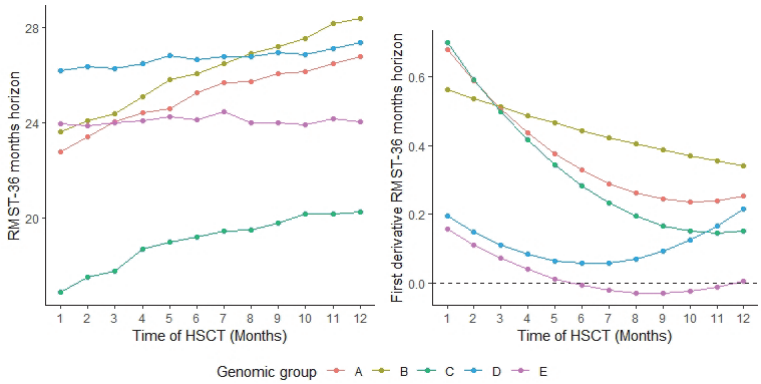


Fig. 2 Estimated RMST for 57 years old male patients. Estimated RMST (left panel) and its first derivative (right panel) according to different transplant policies (x-axis) for different genomic profiles.

Acknowledgements The authors thank 2020 GenoMed4ALL (Genomics and Personalized Medicine for all through Artificial Intelligence in Haematological Diseases - http://www.genomed4all.eu) for sharing the dataset used in this work.

References

1. Campo, E., Harris, N. L., Jaffe, E. S., Pileri, S. A., Stein, H., Thiele, J., Vardiman, J. W.: WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. WHO Press, Geneva, 4th Edition (2008) isbn : 9789283244943
2. Cazzola, M., Della Porta, M. G., Malcovati, L.: The genetic basis of myelodysplasia and its clinical relevance. Blood (2013) doi : 10.1182/BLOOD-2013-09-381665
3. Bersanelli, M., Travaglino, E., Meggendorfer, M., Matteuzzi, T., Sala, C., Mosca, E., ... , Della Porta, M. G.: Classification and Personalized Prognostic Assessment on the Basis of Clinical and Genomic Features in Myelodysplastic Syndromes. Journal of clinical oncology (2021) doi: 10.1200/JCO.20.01659
4. Krijkamp, E. M., Alarid-Escudero, F., Enns, E. A., Jalal, H. J., Hunink, M. M., Pechlivanoglou, P: Microsimulation modeling for health decision sciences using R: a tutorial.Medical decision making (2018). doi: 10.1177/0272989X18754513
5. Rutter, C. M., Zaslavsky, A. M., Feuer, E. J. : Dynamic microsimulation models for health outcomes: a review.Medical decision making (2011). doi: 10.1177/0272989X10369005
6. Therneau T.: A Package for Survival Analysis in R. R package version 3.3-1 (2022). https://CRAN.R-project.org/package=survival.
7. de Wreede L.C., Fiocco M., Putter H.: "mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software (2011), 38(7), 1–30. doi: 10.18637/jss.v038.i07.

A fully Bayesian approach for sample size determination of Poisson clinical trials.

Approccio completamente Bayesiano per la scelta della dimensione campionaria di studi clinici basati su dati di conteggio.

Susanna Gentile and Valeria Sambucini

Abstract In this paper, we exploit a fully Bayesian approach to determine the optimal exact sample size of a single-arm trial based on count data. The idea is to select the sample size by controlling the *Predictive Bayesian Power*, i.e. the predictive probability of obtaining a Bayesian significant result, under the assumption that the treatment is actually effective. An essential element of the procedure is the specification of two different prior distributions and some suggestions about their choice are provided.

Abstract *In questo articolo si sfrutta un approccio completamente bayesiano per la scelta della numerosità campionaria esatta per uno studio a braccio singolo basato su dati di conteggio. L'idea è di selezionare la dimensione del campione controllando la Potenza Bayesiana Predittiva, ovvero la probabilità predittiva di ottenere un risultato statisticamente significativo secondo un ottica Bayesiana, assumendo che il trattamento sia effettivamente efficace. Un elemento essenziale della procedura è l'introduzione di due diverse distribuzioni a priori sulla scelta delle quali vengono forniti alcuni suggerimenti.*

Key words: Exact sample size determination, fully Bayesian approach, Poisson data, Predictive Bayesian Power, two-priors approach.

Susanna Gentile
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: gentile.1705895@studenti.uniroma1.it

Valeria Sambucini
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: valeria.sambucini@uniroma1.it

1 Introduction

The choice of the sample size is a key element of any study design, especially in the field of clinical trials. In this context, Bayesian methods for sample size calculation are particularly attractive because of their flexibility and capability of modelling uncertainty. In this paper, we adopt a "fully Bayesian approach", also called a "proper Bayesian approach", that does not mix frequentist and Bayesian tools, but uses only Bayesian concepts both at the analysis and at the design stage of the trial. As a consequence, the method requires the specification of two different prior distributions, which pursue different purposes: the *analysis* and the *design* priors (see [1], [3] and [7]).

Let us focus on hypothesis testing and consider the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$, where θ denotes the parameter of interest that measures the effect of the experimental treatment. The hypothesis H_1 defines the condition about the parameter that corresponds to a treatment considered sufficiently effective. In a Bayesian setup, we elicit the analysis prior distribution, $\pi^A(\theta)$, that expresses the pre-experimental knowledge, as well as the possible absence of information, about the unknown parameter θ. It is used to compute the posterior distribution, $\pi^A(\theta|y_n)$, where y_n denotes the observed experimental result based on a sample of size n. At the planning stage of the trial, we have not collected the trial data yet and the experimental result is a random quantity, Y_n. Following Spiegelhalter et al. [6], this random result can be considered statistically significant under a Bayesian perspective if it yields a large posterior probability that θ belongs to the alternative hypothesis. In other words, the definition of the *Bayesian significance* provides the condition to obtain a successful experiment that leads to the rejection of the null hypothesis. The sample size determination criterion is based on the computation of the so-called *Predictive Bayesian Power*, $PBP(n)$, that is the predictive probability of obtaining a Bayesian significant result, when the true θ actually belongs to the alternative hypothesis. This latter optimistic assumption is realized by introducing a design prior distribution, $\pi^D(\theta)$, that assigns negligible probability values to the sub-space under H_0 and that is used to compute the prior predictive distribution of Y_n. In general terms, the criterion selects the minimum n that ensures a sufficiently large value for the Predictive Bayesian Power.

In this paper, we apply the fully Bayesian procedure described above for exact sample size calculation in single-arm trials based on count data. As a consequence of using exact methods, we obtain that $PBP(n)$ is not a monotonically increasing function of n, but tends to increase following a saw-toothed behaviour that typically occurs when dealing with discrete distributions of Y_n (see [2]). To take into account this behaviour, in line with Sambucini [5], we suggest to select the optimal sample size as

$$n^{PBP} = \min\{n^* \in \mathbb{N} : PBP(n) > \gamma, \forall n \geq n^*\}, \tag{1}$$

where γ is a fixed threshold. In practice, this more conservative criterion takes into account the behaviour of $PBP(n)$ by imposing that the condition of interest is satisfied for all the sample size values greater than or equal to the optimal one.

The outline of the article is as follows. In Section 2, we illustrate the Bayesian formulation of the problem when a single-arm study with Poisson data is performed and describe the sample size determination criterion based on $PBP(n)$. In Section 3 we show some numerical results and provide some suggestions about the choice of the analysis and the design priors. Finally, Section 4 contains some concluding remarks.

2 Sample size determination based on $PBP(n)$ for Poisson data

Let $(X_1, X_2, ..., X_n)$ be a random sample of count data. We assume that, for $i = 1, ..., n$, X_i represents the number of negative events occurred over a period of time for the i-th patient enrolled in a single-arm study aimed at establishing the efficacy of a new experimental treatment. Each X_i follows a Poisson distribution of rate $\theta > 0$, so that the sampling distribution of the sufficient statistic $S_n = \sum_{i=1}^{n} X_i$ is $f(s_n|\theta) = \text{pois}(s_n; n\theta)$, for $s_n = 0, 1, 2, \cdots$. The experimental treatment can be considered sufficiently effective if the event rate is below a target value θ_0 and, therefore, the interest is focused on the hypotheses $H_0 : \theta \geq \theta_0$ and $H_1 : \theta < \theta_0$.

Under a Bayesian framework, we exploit standard conjugate results and introduce a *gamma analysis prior distribution* to account for pre-experimental information on θ, $\pi^A(\theta) = \text{gamma}\left(\theta; \alpha^A, \beta^A\right)$, with $\alpha^A, \beta^A > 0$. The corresponding posterior distribution is $\pi^A(\theta|s_n) = \text{gamma}\left(\theta; \alpha^A + s_n, \beta^A + n\right)$. At the planning stage of the trial, the *Bayesian significance* of the random result S_n is provided by the condition

$$\mathbb{P}_{\pi^A(\cdot|S_n)}(\theta < \theta_0) = \text{Gamma}(\theta_0; \alpha^A + S_n, \beta^A + n) > \lambda, \tag{2}$$

where λ is a fixed probability threshold, $\mathbb{P}_{\pi^A(\cdot|S_n)}$ denotes the probability measure associated with the posterior distribution of θ and $\text{Gamma}(\cdot; a, b)$ denotes the c.d.f. of a gamma distribution of parameters a and b. The posterior probability in (2) is a decreasing function of S_n and, due to the discreteness of the random result, it is possible to find a non-negative integer k such that

$$\mathbb{P}_{\pi^A(\cdot|k)}(\theta < \theta_0) > \lambda \qquad \text{and} \qquad \mathbb{P}_{\pi^A(\cdot|k+1)}(\theta < \theta_0) \leq \lambda.$$

As a consequence, we can say that the random result S_n is significant under a Bayesian perspective if $S_n \leq k$, where

$$k = \max\left\{u \in \{0, 1, \cdots\} : \text{Gamma}\left(\theta_0; \alpha^A + u, \beta^A + n\right) > \lambda\right\}.$$

In order to derive the Predictive Bayesian Power, we introduce a *gamma design prior distribution*, $\pi^D(\theta) = \text{gamma}\left(\theta; \alpha^D, \beta^D\right)$, with $\alpha^D, \beta^D > 0$, used to

formalize the optimistic design expectations about the efficacy of the experimental treatment. By exploiting this prior density, we obtain the following Negative-Binomial prior predictive distribution of S_n

$$m^D(s_n) = \text{Nb}\left(s_n; \alpha^D, \frac{\beta^D}{\beta^D + n}\right), \quad \text{for } s_n = 0, 1, 2, \cdots.$$

Then, the Predictive Bayesian Power can be obtained as

$$\begin{aligned} PBP(n) &= \mathbb{P}_{m^D(\cdot)}(S_n \leq k) \\ &= \sum_{s_n=0}^{k} \text{Nb}\left(s_n; \alpha^D, \frac{\beta^D}{\beta^D + n}\right), \end{aligned}$$

where $\mathbb{P}_{m^D(\cdot)}$ denotes the probability measure associated with the prior predictive distribution of S_n. In practice, it is given by the sum of the predictive probabilities of all the possible Bayesian significant results, obtained under the assumption that the experimental treatment is actually effective. Given the discrete nature of the Negative-Binomial distribution, we propose to use the conservative criterion in (1) to select the optimal value of n.

3 Numerical results

In this Section, we analyse the impact of the different design parameters on the optimal sample size. Let us assume that $\theta_0 = 1.9$. To elicit the gamma design prior, we find it useful to follow a procedure commonly used for the beta prior densities in the presence of binary data, that consists in expressing the hyperparameters in terms of prior mode and prior sample size (see [4]). In our case, we obtain that by setting

$$\alpha^D = n^D \theta^D + 1 \qquad \text{and} \qquad \beta^D = n^D,$$

$\pi^D(\theta)$ is centred on θ^D and its concentration can be regulated by varying the prior sample size n^D. Since we use this distribution to assume that θ belongs to the alternative hypothesis, it is reasonable to fix θ^D smaller than θ_0 and to set n^D so that the prior probability assigned to the null hypothesis is negligible. In particular, we consider two possible strategies to select n^D: (i) we set $n^D = \infty$, obtaining that $\pi^D(\theta)$ assigns all the probability mass to θ^D and (ii) we select n^D as the smallest value such that $P(\theta < \theta_0) \simeq 0.999$. Note that, in the first case, no uncertainty is introduced at the planning stage and the Predictive Bayesian Power equals the so-called *Conditional Bayesian Power* (see [1] and [5]). For the analysis prior, we similarly ensure that the prior mode is θ^A, but we set the prior sample size n^A equal to 1, in order to represent weak prior information that lets the data be predominant in the posterior analysis.

In Figure 1, we show the behaviour of $PBP(n)$, when $\theta^D = 1.5$, $\theta^A = 1.9$ and $\lambda = 0.95$. Because of the saw-toothed shape, we fix $\gamma = 0.8$ and select the optimal

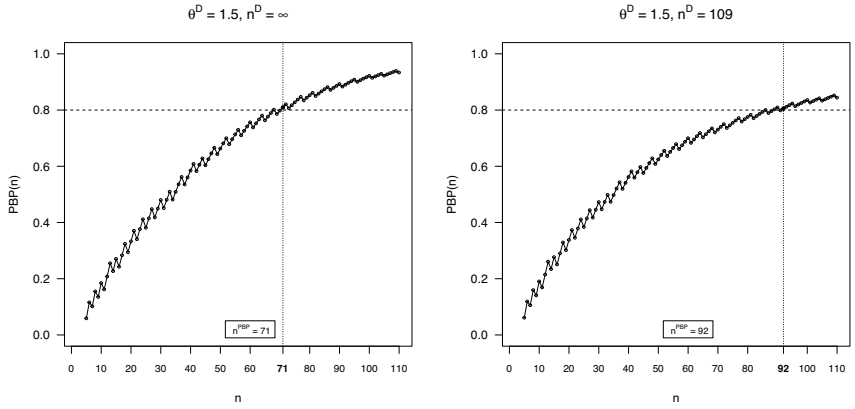


Fig. 1 Behaviour of $PBP(n)$, when $\theta_0 = 1.9$, $\theta^A = 1.9$, $n^A = 1$ and $\lambda = 0.95$. The optimal sample size is selected according to the criterion in (1) for $\gamma = 0.8$.

sample size according to the criterion in (1). We can note that, when $n^D = \infty$ and no uncertainty is accounted for through the design prior, the Bayesian power increases faster and the corresponding n^{PBP} is smaller. The same behaviour can be observed in Table 1, which reports the optimal sample sizes for $\gamma = 0.8$ for different values of θ^A, θ^D, λ and n^D. More specifically, the gap between the two sample sizes increases as θ^D gets closer to θ_0. Regardless of n^D, the selection of θ^D especially impacts the optimal sample sizes because we are considering three completely different scenarios under H_1. In particular, the closer θ^D and θ_0 are, the larger n^{PBP} is. The sample size also increases with λ because we require stronger evidence to consider the result statistically significant. Finally, even though the analysis priors are weakly informative, the optimal sample size increases with the degree of prior scepticism towards the new treatment expressed by $\pi^A(\theta)$. In this case, given the hypothesis system at interest, a larger θ^A expresses stronger scepticism.

4 Conclusion

In the present paper, we address the problem of sample size determination for single-arm studies based on Poisson data. More specifically, we adopt a fully Bayesian approach by exploiting the concept of Predictive Bayesian Power. The proposed strategy has two important advantages in that (i) it allows the introduction of possible pre-experimental knowledge about θ through the specification of an analysis prior distribution and (ii) it models uncertainty on θ at the planning stage with the design prior distribution, avoiding local optimality. Moreover, instead of using normal approximations, we resort to exact methods based on discrete

Table 1 Optimal sample sizes for different values of θ^D, n^D, θ^A and λ, when $\theta_0 = 1.9$, $n^A = 1$ and $\gamma = 0.8$.

		$\lambda = 0.9$			$\lambda = 0.95$			$\lambda = 0.99$		
		θ^A			θ^A			θ^A		
θ^D	n^D	0.9	1.9	2.9	0.9	1.9	2.9	0.9	1.9	2.9
1.5	∞	48	52	57	66	71	75	109	113	118
	109	58	66	71	84	92	97	143	151	156
1.6	∞	86	95	101	121	126	132	196	205	210
	195	107	117	123	152	162	168	260	270	279
1.7	∞	200	212	219	280	287	298	453	464	475
	444	249	262	270	353	366	374	600	612	625

distributions of the data and a conservative criterion is necessary to account for the not monotonic behaviour of $PBP(n)$.

Finally, let us notice that, without loss in generality, we have considered the count of "negative" events, that represent a not-desired outcome for patients. The proposed Bayesian criteria can be similarly derived when the hypotheses are reversed because a "positive" event is considered.

References

1. Brutti P., De Santis F., Gubbiotti S.: Robust Bayesian sample size determination in clinical trials. Stat. Med., **27**, 2290–2306 (2008)
2. Chernick, M.R., Liu, C.Y.: The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. Am. Stat., **56**, 149–155 (2002)
3. Sahu, S.K., Smith, T.M.F.: A Bayesian method of sample size determination with practical applications. J. R. Stat. Soc., **169**, 235–253 (2006)
4. Sambucini V.: A Bayesian predictive two-stage design for phase II clinical trials. Stat. Med., **27**, 1199-1224 (2008)
5. Sambucini, V.: Bayesian vs frequentist power functions to determine the optimal sample size: testing one sample binomial proportion using exact methods. In: Bayesian Inference, Tejedor, J.P. (Ed.) IntechOpen, 77–97 (2017)
6. Spiegelhalter, D.J. and Abrams, K.R. and Myles, J.P.: Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons (2004)
7. Wang, F., Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Stat. Sci., **2**, 193–218 (2002)

Compartmental models in epidemiology: Application on Smoking Habits in Tuscany

Modelli compartimentali in epidemiologia: Applicazione sull'abitudine al fumo in Toscana

Alessio Lachi, Cecilia Viscardi, Maria Chiara Malevolti, Giulia Carreras, and Michela Baccini

Abstract We developed a compartmental model to describe the smoking habits evolution in Tuscany, relying on flexible modelling of age and sex-dependent smoking transition rates. Calibrating on observed data, we estimated the prevalence of current, former and never smokers in the population, and quantified the fraction of deaths attributable to smoking from 1994 to 2033. The model results indicate that smokers prevalence is decreasing over time. We speculate that this reduction could be related to health policies implemented up to now. Results highlight also that smoking habits are different among males and females.

Abstract *Abbiamo sviluppato un modello compartimentale che descrive l'evoluzione delle abitudini al fumo in Toscana, modellando in modo flessibile i tassi di fumo, età e sesso dipendenti. Calibrando su dati osservati, abbiamo stimato la prevalenza di fumatori, ex fumatori e non fumatori nella popolazione e abbiamo quantificato la frazione di decessi attribuibili al fumo dal 1994 al 2033. I risultati del modello indicano che la prevalenza dei fumatori sta diminuendo nel tempo. Tale riduzione potrebbe essere dovuta alle politiche sanitarie sinora attuate. I risultati mostrano inoltre che le dinamiche di abitudine al fumo sono diverse tra uomini e donne.*

Key words: Compartmental Model, Smoke, Tobacco control, Prevalence, Population attributable fraction.

Alessio Lachi
DiSIA (UNIFI), Viale Morgagni 65 - Firenze, e-mail: alessio.lachi@unifi.it

Cecilia Viscardi
DiSIA (UNIFI), Viale Morgagni 65 - Firenze, e-mail: cecilia.viscardi@unifi.it

Maria Chiara Malevolti
ISPRO, Via Cosimo il Vecchio 2 - Firenze, e-mail: m.malevolti@ispro.toscana.it

Giulia Carreras
ISPRO, Via Cosimo il Vecchio 2 - Firenze, e-mail: g.carreras@ispro.toscana.it

Michela Baccini
DiSIA (UNIFI), Viale Morgagni 59 - Firenze, e-mail: michela.baccini@unifi.it

1 Introduction

Smoking is a major risk factor for many common chronic diseases. In particular, over 85% of lung cancers are attributable to smoking. Also, smoking reduces length and quality of life [13].

The World Health Organization Framework Convention on Tobacco Control (WHO FCTC) considers the implementation of tobacco control policies (TCP) as the primary prevention strategy to reduce the prevalence of smokers in the population and the burden of mortality and morbidity attributable to smoking.

Aggregate or compartmental models allow to describe the dynamics of smoking habits in time and to compare the effectiveness of TCPs. They start from an initial population, divided into non overlapping compartments according to observed smoking prevalence, and reproduce the evolution of their sizes through a system of continuous-time dynamic equations [11, 9]. The mechanistic nature of such models results in the easiness of simulation of the system evolution.

This study aims to develop a compartmental model that describes the evolution of smoking habits from 1994 to 2033 in Tuscany, a region of central Italy, estimating the rates of starting, quitting, and relapsing smoking, and forecasts the prevalence of actual, former, and never smokers over time, allowing the evaluation of the impact of previous TCPs.

2 Methods

2.1 Model specification

We considered a compartmental smoking model (CSM) in which people, classified by gender (g) and age (a), are grouped into non overlapping compartments based on their smoking status: Current (C), Never (N), Former (F) smokers, and the related deaths compartments. The compartments C and F are further divided into subgroups denoted by C_i and $F_{i,c}$ for $i \in \{l,m,h\}$ and $c \in \{0,...,15\}$. Subscripts l, m and h stand for smoke intensity: low (cig/day$<$10), medium (10$\leq$cig/day$\leq$19), and high (cig/day$>$19). Subscript c denotes the time from smoking cessation in years.

Transitions of the individuals from a given compartment to another one determine flows that generate changes in the size of the compartment. Such transitions occur with probabilities governed by the annual rate of starting (γ), quitting (ε) and relapsing (η_c) smoking, and by the mortality rates (δ_{C_i}, δ_N, and $\delta_{F_{i,c}}$).

The assumptions underlying the model are: 1) the population is closed to immigration and emigration but opened to newborns, (birth rates from the Italian Institute of Statistics (ISTAT) database of 2011); 2) people start smoking in age $\in \{14,...,24\}$ ($\gamma(a) = 0$ for $a \in \{0,...,13\} \cup \{25,...,100\}$; 3) people are allowed becoming former smokers in age $\in \{25,...,100\}$ ($\varepsilon(a) = 0$ for $a \in \{0,...,24\}$); 4) the probability of starting, quitting, and relapsing smoking are independent from mortality.

Accordingly, the dynamic of the system is described by the following differential equations defined for each cohort [1] (y), and gender (g):

$$\begin{cases} \frac{dN^y(t)}{dt} = & N^y(t)\Big(1-\delta_N(t-y)\Big)\Big(1-\gamma(t-y)\Big) \\ \frac{dC_i^y(t)}{dt} = & C_i^y(t)\Big(1-\delta_{C_i}(t-y)\Big)\Big(1-\varepsilon(t-y)\Big)+N^y(t)\Big(1-\delta_N(t-y)\Big)\gamma(t-y)\pi_{C_i}(t-y)\mathbb{1}_{14\leq(t-y)\leq 24}+ \\ & \sum_{c=1}^{15+}F_{i,c-1}^y(t)\Big(1-\delta_{F_c}(t-y)\Big)\eta_c\mathbb{1}_{(t-y)>24} \\ \frac{dF_i^y(t)}{dt} = & \sum_{c=1}^{15+}F_{i,c-1}^y(t)\Big(1-\delta_{F_c}(t-y)\Big)\Big(1-\eta_c\Big)+C_i^y(t)\Big(1-\delta_{C_i}(t-y)\Big)\varepsilon(t-y) \\ \frac{dDN^y(t)}{dt} = & N^y(t)\delta_N(t-y) \\ \frac{dDC_i^y(t)}{dt} = & C_i^y(t)\delta_{C_i}(t-y) \\ \frac{dDF_i^y(t)}{dt} = & \sum_{i\in\{l,m,h\}}\sum_{c=1}^{15+}F_{i,c-1}^y(t)\delta_{F_c}(t-y). \end{cases} \tag{1}$$

Note that in Eq. (1) a can be viewed as $t-y$ and π_{C_i} represents the percentage of current smokers for each level of intensity $i\in\{l,m,h\}$.

We considered γ, ε, and η_c as unknown parameters to be estimated. In particular, we define γ and ε as natural cubic regression splines of a with 4 and 3 degrees of freedom:

$$\gamma(a)=s(a;\psi_0,\psi_1,\psi_2,\psi_3);\quad \varepsilon(a)=s(a;\phi_0,\phi_1,\phi_2).$$

The relapsing smoke rate, η_c is modelled as a negative exponential function of time from smoking cessation c:

$$\eta(c)=\alpha\beta e^{-12c\beta}$$

where β governs how fast the relapse rate declines over time from smoking cessation, and α governs the lifetime probability of no relapse [7].

The parameters π_{C_i} are fixed to values taken from ISTAT surveys [8].

The mortality rates δ_N, δ_{C_i} and $\delta_{F_{i,c}}$ are assumed gender-specific and constant over time. For each age we computed the mortality rates as follows:

$$\delta_N(a)=\delta_{Pop}(a);\ \ \delta_{C_i}(a)\approx RR_{C_i}\times\delta_{Pop}(a);\ \ \delta_{F_{i,c}}(a)\approx RR_{F_{i,c}}\times\delta_{Pop}(a).$$

We relied on the Relative Risks (RR) estimated in [12] and the mortality $\delta_{Pop}(a)$ rate taken from ISTAT database of 2011.

[1] People born in the same year

2.2 Model calibration and uncertainty quantification

Let us denote by $\theta = (\psi_0, \psi_1, \psi_2, \psi_3, \phi_0, \phi_1, \phi_2, \alpha, \beta)$ the vector of the parameters to be estimated and by $p^{\theta}(t,a) = \left(p_C^{\theta}(t,a), p_N^{\theta}(t,a), p_F^{\theta}(t,a)\right)$ the vector of prevalence computed from the sizes of compartments at time t for the age a [2], given a specific value of θ. A calibration procedure consists of searching the vector θ leading to $p^{\theta}(t,a)$ as closed as possible to $p^{obs}(t,a) = \left(p_C^{obs}(t,a), p_N^{obs}(t,a), p_F^{obs}(t,a)\right)$, the vector of the observed prevalence.

We calibrated our model on the prevalence estimates derived from the multipurpose household surveys data "Aspect of daily life" (AVQ) [8] carried out by ISTAT. Since 1993 the multipurpose surveys AVQ collect fundamental information related to the daily life of individuals and families. In particular, we considered $t \in \{1994, ..., 2019\}$ and we simulated the evolution of the system up to 2033. The model was calibrated separately by gender.

To compare observed and simulated trajectories we considered the following objective function where $H(\cdot,\cdot)$ denotes the Hellinger distance [6]:

$$min\left[\frac{1}{t \times a}\sum_{t,a} H\left(p^{\theta}(t,a), p^{obs}(t,a)\right)\right] = \frac{1}{t \times a \times \sqrt{2}}\sum_{t,a}\sqrt{\sum_{i \in \{C,F,N\}}\left(\sqrt{p_i^{\theta}(t,a)} - \sqrt{p_i^{obs}(t,a)}\right)^2}. \quad (2)$$

We performed a constrained optimization procedure resorting to the R package `nloptr` [14]. To take into account the sampling variability and quantify the uncertainty around point estimates we used a parametric bootstrap procedure [4]. Following an approach similar to [3], we estimated percentile bootstrap confidence intervals assuming that each prevalence followed a Dirichlet distribution, $p^{\theta}(t,a) \sim Dirichlet\left(C^{\hat{\theta}}(t,a), N^{\hat{\theta}}(t,a), F^{\hat{\theta}}(t,a)\right)$ where $C^{\hat{\theta}}$, $F^{\hat{\theta}}$, and $N^{\hat{\theta}}$ represent the size of the compartments corresponding to the best estimate $\hat{\theta}$ derived by minimizing Eq. 2. The sampling procedure was repeated n=1000 times to obtain bootstrap replicates of the unknown parameters used to compute confidence intervals.

We also computed population attributable fraction (PAF) for all cause of death as defined in [5] with never smokers as counterfactual level. These PAFs measure the proportion of deaths that would be avoided if the smoking risk factor in the population were eliminated.

3 Results and discussion

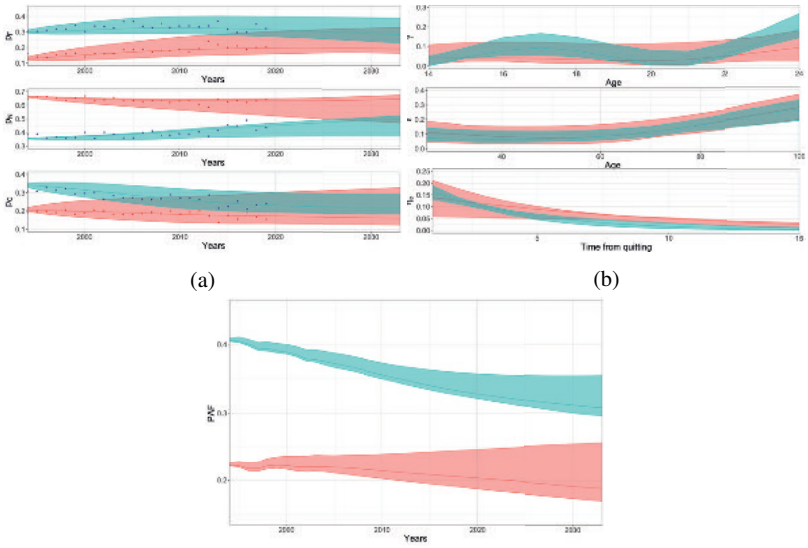
Fig. 1 panel a) shows the estimated prevalence from 1994 to 2033 for each gender. Looking at the observed data (blue and red dots), we observe an adequate model

[2] We considered the age classes 14-17, 18-19, 20-24, 25-34, 35-44, 45-54, 55-59, 60-64, 65-74, 75+

fit. Our forecasts suggest that the smoking prevalence will decrease in the next 15 years. This reduction may be due to the health policies so far implemented.

Panel b) shows the estimates of the rate of starting, quitting and relapsing smoking. From the comparison between rates of starting among males and females, we can see that males are more likely to start smoking than females. Moreover, the female rate exhibits an almost linear behaviour over time while the male rate shows a peak at seventeen. As regards the two rates of quitting, the two groups have the same behaviour, while the relapse rate among females becomes greater than among males after 5 years from quitting.

Panel c) represents PAFs. Our model predicts a reduction of PAF in the next years, due to the reduction of smoking prevalence. Males have a higher PAF than females, due to the higher prevalence of smoking among males.



(a) (b)

(c)

Fig. 1: Figure shows the prevalence estimates (a), the estimates of $\gamma - \varepsilon$ and η_c (b), and PAF (c) along with their 90% confidence bands by gender (males in blue and female in pink).

Several compartmental models have been developed to forecast future smoking rates and assess the impact of TCP [10, 9, 1, 2]. The Simsmoke model developed by Levy [10] has been largely used and applied to several countries, including Italy [9]. The model developed by Carreras et al [1, 2] extended the Simsmoke model by considering time from cessation in former smokers and by taking into account relapsing smoking. Our model adds further elements of novelty to this last model by considering also smoking intensity as a factor affecting mortality. Moreover, we introduced in the compartmental model cubic splines for estimating starting and

quitting smoking rates, thus obtaining more realistic trajectories. Finally, we evaluated parameter variability by estimating confidence intervals based on a parametric bootstrap procedure.

The model has several limitations. In particular, we assume the closeness of the population to immigration and emigration as well as constant mortality, birth rates, and, more in general, constant transition rates. Future work should include multivariate splines to model the dependence of the rates from calendar years. We should also consider second-hand smoking and other risk factors such as environmental pollution.

References

1. Giulia Carreras, Silvano Gallus, Laura Iannucci, and Giuseppe Gorini. Estimating the probabilities of making a smoking quit attempt in italy: stall in smoking cessation levels, 1986-2009. *BMC Public Health*.
2. Giulia Carreras, Giuseppe Gorini, and Eugenio Paci. Can a national lung cancer screening program in combination with smoking cessation policies cause an early decrease in tobacco deaths in italy? *Cancer prevention research*, 2012.
3. Gerardo Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *KeAi*, 2017.
4. Bradley Efron and Ryan Tibshirani. An introdcution to the bootstrap. 1993.
5. GBD. Gbd 2019 tobacco collaborators. spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the global burden of disease study 2019. *The Lancet*, 2021.
6. Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *De Gruyter*, 1909.
7. Rudolf Hoogenveen, Pieter Van Baal, Hendriek Boshuizen, and Talitha Feenstra. Dynamic effects of smoking cessation on disease incidence, mortality and quality of life. *BioMed Central*, 2008.
8. ISTAT. Indagine multiscopo della famiglia: Aspetti della vita quotidiana. 1993-2019. URL: `www.istat.it/it/archivio/91926`.
9. David Levy, Silvano Gallus, Kenneth Blackman, Giulia Carreras, Carlo La Vecchia, and Giuseppe Gorini. Italy simsmoke: the effect of tobacco control policies on smoking prevalence and smoking attributable deaths in italy. *BMC Public Health*, 2012.
10. David Levy, Leonid Nikolayev, Elizabeth Mumford, and Christine Compton. The healthy people 2010 smoking prevalence and tobacco control objectives: results from the simsmoke tobacco control policy simulation model (united states). *Cancer Causes and Control volume*, 2005.
11. David Levy, Hana Ross, Eduard Zaloshnja, Roland Shuperka, and Meriglena Rusta. The role of tobacco control policies in reducing smoking and deaths caused by smoking in an eastern european nation: results from the albania simsmoke simulation model. *Journal of Public Health*, 2008.
12. Michael Thun, Brian Carter, Diane Feskanich, Neal Freedman, Ross Prentice, Alan Lopez, Patricia Hartge, and Susan Gapstur. 50-year trends in smoking-related mortality in the united states. the new england journal of medicine. *The new england journal of medicine*, 2013.
13. WHO. Tobacco smoke and involuntary smoking, iarc monographs on the evaluation of carcinogenic risks to humans. *IARC*, 2004.
14. Jelmer Ypma, Steven Johnson, Hans Borchers, Dirk Eddelbuettel, Brian Ripley, Kurt Hornik, Julien Chiquet, Avraham Adler, Xiongtao Dai, Aymeric Stamm, and Jeroen Ooms. R interface to nlopt. URL: `cran.r-project.org/web/packages/nloptr/nloptr.pdf`.

Covid-19 Assessment and Evaluation 2

We are in the same storm but not in the same boat: Impact of COVID-19 on UK households

Ognuno sta solo sul cuor della terra: Impatto del COVID-19 sulle famiglie del Regno Unito

Demetrio Panarello and Giorgio Tassinari[1]

Abstract The United Kingdom introduced a national lockdown in March 2020, as a means to curb the rising pace of COVID-19 infections in the country. Since then, the various restrictions imposed on citizens have caused enormous social and economic consequences on the UK population. However, full awareness of the mid- and long-term impacts of such measures is still lacking. In this paper, by making use of longitudinal data from the Understanding Society COVID-19 study, consisting of nine survey waves administered to a representative sample of UK citizens from April 2020 to September 2021, we analyse the potential determinants of employment losses and provision of financial assistance, particularly emphasising the differential effects related to individuals' socio-demographic characteristics.

Abstract *Il Regno Unito ha introdotto un lockdown nazionale nel marzo 2020, allo scopo di frenare il ritmo crescente di contagi da COVID-19 nel Paese. Da allora, le varie restrizioni hanno causato enormi conseguenze di tipo sociale ed economico sulla popolazione britannica. Tuttavia, manca ancora una piena consapevolezza degli impatti a medio e lungo termine di tali misure. In questo lavoro, facendo uso di dati longitudinali dell'Understanding Society COVID-19 study, consistente in nove ondate di questionari somministrati a un campione rappresentativo di cittadini britannici da aprile 2020 a settembre 2021, analizziamo le potenziali determinanti della perdita del lavoro e del supporto economico, sottolineando in particolare gli effetti differenziali legati alle caratteristiche socio-demografiche degli individui.*

Keywords: Financial support, Job loss, Pandemic, Understanding Society.

[1] Demetrio Panarello, University of Bologna; email: demetrio.panarello@unibo.it
Giorgio Tassinari, University of Bologna; email: giorgio.tassinari@unibo.it

1 Introduction

It has already been over two years since the introduction of the first national lockdown in the United Kingdom (March 23, 2020), carried out as a response to the rapid spread of COVID-19 cases in the country. Since then, the price of the implemented restrictions has been incredibly high, with strong consequences on the population as a whole and, in particular, on its most vulnerable segments, in terms of mass unemployment, isolation of the people, and widespread financial difficulties. Indeed, among European countries, the United Kingdom was one of the hardest hit by the pandemic. As of April 15, 2022, the national number of deaths due to COVID-19 amounts to about 171,000 (number of people who died within 28 days of being identified as a COVID-19 case by a positive test) and three lockdowns have been implemented in the country so far. The economic consequences at the aggregate level have been extremely strong. In 2020, GDP decreased by 9.7% compared to the previous year, followed by a growth of 6.9% in 2021 [7]; the unemployment rate in 2020 has increased by only 1.5%, mostly as a result of government support, while in 2021, despite the strong recovery in the economic activity level, a further increase of 1.8% is expected, until reaching a level of 7.1% [6]. Moreover, households' consumption fell by 22.4% in the second quarter of 2020, compared with the previous quarter [1]. Recent evidence has shown that the COVID-19 pandemic and the related social and economic interventions, such as physical distancing and closure of production activities, have had different impacts on the various social groups. For instance, in the UK, females and parents are among those who have experienced the largest reduction in subjective well-being [2,9]. Black, Asian and minority ethnic groups (BAME) exhibited a higher COVID-19-related mortality rate than the White population [8] and suffered harsher economic consequences [3]. The available evidence identifies the existence of immediate effects, but our understanding of the mid- and long-term impacts of COVID-19 and related measures remains incomplete. To this end, we analyse a representative sample of the UK population, the UK Household Longitudinal Study (UKHLS), making use of data from the Understanding Society COVID-19 Study [4,5].

Through this study, we aim at contributing to the comprehension of the effects of COVID-19 by focusing on two outcome variables: job loss and receiving financial support. In this regard, it should be mentioned that government interventions to support families and businesses accounted for 5.5% of GDP in 2020 [6]. As regards households, the Coronavirus Job Retention Scheme took shape. It guaranteed transfers to companies to pay 80% of furloughed workers' wages up to a maximum of 2,500 British pounds. In addition, self-employed workers received a taxable subsidy equal to 80% of the average income of the previous three years. Furthermore, the government temporarily increased the unemployment benefits.

In what follows, we will be highlighting the differential effects on the two outcome variables, related to the structural characteristics of individuals and households – such as gender, age, ethnic group, and family size. The next Section illustrates the data and techniques put to use to reach our aims, while the main findings are presented in Section 3. Then, some final remarks are given in Section 4.

2 Data and Methods

The Understanding Society COVID-19 study is a longitudinal survey aimed at capturing the experiences of UK individuals to the COVID-19 pandemic, covering the changing impact of the pandemic on the welfare of the UK population [4,5]. It is part of the UK Household Longitudinal Study (UKHLS) and includes all members of the main Understanding Society sample who participated in waves 8 or 9 (2017-2018), selected through a probability sampling of postal addresses in the UK. The COVID-19 survey was conducted during the first lockdown (April, May, June 2020), in its immediate aftermath (July, September 2020), during the last two lockdowns (November 2020, January 2021, March 2021), and finally in September 2021, for a total of nine survey waves so far.

For our analyses, we estimate panel logit models of two outcome variables: lost employment (Model 1) and financial support received (Model 2). Both models include several regressors related to the socio-demographic characteristics of the individuals and incorporate wave fixed effects. Descriptive statistics, computed for the sample that is not missing for any of the variables, are presented in Table 1.

Table 1: Descriptive statistics of the variables employed in the models.

Variable	*Obs.*	*Min*	*Median*	*Max*	*Mean*	*Std. Dev.*
Lost employment	39574	0	0	1		
Financial support	39574	0	0	1		
White	39574	0	1	1		
Age	39574	16	49	88	47.80	12.18
Male	39574	0	0	1		
Household size	39574	1	3	11	2.809	1.260
Living with a partner	39574	0	1	1		
Long-term health condition	39574	0	0	1		
Feeling lonely (1: Hardly ever or never; 2: Some of the time; 3: Often)	39574	1	1	3		
Payments (1: Up to date with all bills; 2: Behind with some bills; 3: Behind with all bills)	39574	1	1	3		
Currently employed	39574	0	1	1		
Type of employment in the previous wave (1: Employed; 2: Self-employed; 3: Both)	39574	1	1	3		
Current household income amount	39574	0	3000	250001	6665.11	16187.24

3 Main results

Table 2 shows the results from the first model, in which the outcome is a dichotomic variable taking value 1 if the respondent was working in January/February 2020 and has lost his or her job since, and 0 if he or she is still employed.

White people, compared to Black, Asian and minority ethnic groups, have a lower probability of losing their job. The same applies to males, compared to females, to those who are living with a partner, and to those who declared to be both employed and self-employed during the previous survey wave.

Older individuals have a higher probability of losing their job, as well as those with a long-term health condition, those who are experiencing loneliness, and those with higher incomes. Indeed, job loss is positively associated with difficulties in making ends meet, proxied by being behind with bills' payment.

Finally, it becomes more likely to report a job loss since June 2020, which corresponds to the third wave of the Understanding Society COVID-19 study, compared to the responses collected during the second wave (May 2020).

Table 2: Results from Model 1 – Panel logit regression of lost employment.

Variable	***Coefficient***	***Std. Error***
White	-0.284*	0.147
Age	0.035***	0.004
Male	-0.158*	0.094
Household size	-0.015	0.044
Living with a partner	-0.302***	0.114
Long-term health condition	0.250***	0.093
Feeling lonely: Some of the time	0.027	0.099
Feeling lonely: Often	0.468***	0.179
Payments: Behind with some bills	0.583***	0.196
Payments: Behind with all bills	1.780***	0.652
Self-employed in the previous wave	0.027	0.133
Employed and self-employed in the previous wave	-0.776**	0.302
Current household income amount	4.94e-06**	2.27e-06
Wave 3	0.839***	0.183
Wave 4	0.569***	0.193
Wave 5	1.513***	0.178
Wave 6	1.077***	0.191
Wave 7	1.161***	0.194
Wave 8	0.367	0.224
Wave 9	1.702***	0.182
Intercept	-7.036***	0.394
Observations	*39574*	
Log-likelihood	*-3310*	
Wald Chi-Square	*261.16***	

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 3 shows the results from the second model, in which the outcome is a binary variable taking value 1 if the respondent states to be receiving financial support and 0 otherwise.

White people, compared to BAME, have a lower probability of receiving financial assistance. This is in line with the results of the previous model (see Table 2), in which White respondents were shown to have a lower probability of losing their job. A negative association with receiving financial support can also be pointed out for the older people in the sample, male respondents, and those living with a partner. Moreover, as expected, higher incomes are negatively associated with the likelihood of receiving financial support.

Conversely, individuals from larger households are more likely to receive financial support during the pandemic. The same applies to those declaring to be affected by a long-term health condition, to those who are experiencing loneliness, and to those who are behind with payments. Those who declared to be self-employed during the previous wave of the survey also show a higher likelihood of obtaining financial support, which is in line with the provisions of the Coronavirus Job Retention Scheme.

Finally, financial support looks to be unevenly distributed across the analysed survey waves, proving that policy measures aimed at financially supporting the citizens during the pandemic have seen various adjustments over time.

Table 3: Results from Model 2 – Panel logit regression of financial support received.

Variable	***Coefficient***	***Std. Error***
White	-0.551***	0.178
Age	-0.039***	0.005
Male	-0.402***	0.126
Household size	0.174***	0.046
Living with a partner	-0.605***	0.130
Long-term health condition	0.432***	0.118
Feeling lonely: Some of the time	0.216**	0.090
Feeling lonely: Often	0.517***	0.168
Payments: Behind with some bills	1.235***	0.169
Payments: Behind with all bills	2.475***	0.677
Currently employed	-0.017	0.252
Self-employed in the previous wave	6.109***	0.166
Employed and self-employed in the previous wave	3.279***	0.222
Current household income amount	-1.44e-05***	2.79e-06
Wave 3	-2.298***	0.132
Wave 4	-0.039	0.102
Wave 5	-1.895***	0.135
Wave 6	0.420***	0.106
Wave 7	-2.811***	0.163
Wave 8	-0.379***	0.116
Wave 9	0.347***	0.110
Intercept	-3.924***	0.426
Observations	*39574*	
Log-likelihood	*-6458*	
Wald Chi-Square	*1873.47***	

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.

4 Conclusions

In this study, we analysed the consequences of the pandemic on the UK population, analysing data from the Understanding Society COVID-19 Study [4,5]. We estimated two panel logit regression models, focusing on two outcome variables: job loss and provision of financial aid. Our models considered several socio-demographic characteristics and controlled for the effect of time, as policy measures and their impact on the population have greatly changed throughout the pandemic.

Among other things, our outcomes highlight the vulnerability of some social groups. It appears that the ethnic component plays a key role in determining the probability of employment loss: Black, Asian and minority ethnic groups result to be more at risk of losing their jobs. However, such citizens are also more likely to receive financial assistance compared to Whites, proving that the policy measures put in place to help the individuals in need have performed well. The same can be pointed out for women, compared to men. Nevertheless, older adults are more likely to lose their jobs than younger individuals but less likely to get financial support: this is probably due to a higher monetary wealth that they might be able to tap into, making them more capable than younger individuals to absorb income shocks. However, this may not be true for all, and policymakers should ensure that older individuals who got financially affected by the pandemic are adequately assisted.

We also show that, as time goes by, it becomes increasingly likely for citizens to lose their jobs, although financial support does not seem to be constant over time.

As this pandemic keeps spreading and uncertainty increases, understanding how people react is crucial with a view to helping those in need. By doing so, we will be able to build back better in a post-COVID-19 world.

References

1. Brewer, M., Tasseva, I.V.: Did the UK policy response to Covid-19 protect household incomes?. J. Econ. Inequal. **19**(3), 433—458 (2021)
2. Davillas, A., Jones, A.M.: The first wave of the COVID-19 pandemic and its impact on socioeconomic inequality in psychological distress in the UK. Health Econ. **30**(7), 1668—1683 (2021)
3. Hu, Y.: Intersecting ethnic and native–migrant inequalities in the economic impact of the COVID-19 pandemic in the UK. Res. Soc. Strat. Mobil. **68**, 100528 (2020)
4. Institute for Social and Economic Research: Understanding Society COVID-19 Study, 2020-2021 [data collection], 11th Edition, SN 8644. UK Data Service (2021) doi: 10.5255/UKDA-SN-8644-11
5. Institute for Social and Economic Research: Understanding Society COVID-19 User Guide, Version 10.0, October 2021. University of Essex, Colchester (2021)
6. OECD: OECD Economic Surveys, United Kingdom 2020. OECD Publishing, Paris (2020) doi: 10.1787/2f684241-en
7. OECD: OECD Economic Outlook, Volume 2021 Issue 2. OECD Publishing, Paris (2021) doi: 10.1787/66c5ac2c-en
8. Patel, P., Hiam, L., Sowemimo, A., Devakumar, D., McKee, M.: Ethnicity and COVID-19. BMJ **369**, m2282 (2020)
9. Pierce, M., McManus, S., Hope, H., Hotopf, M., Ford, T., Hatch, S.L., ..., Abel, K.M.: Mental health responses to the COVID-19 pandemic: a latent class trajectory analysis using longitudinal UK data. Lancet Psychiat. **8**(7), 610—619 (2021)

A network approach to investigate learning experiences and social support in higher education

Un approccio di rete per esplorare le esperienze di apprendimento e il supporto sociale degli studenti universitari

Ilaria Primerano, Maria Carmela Catone, Giuseppe Giordano, Maria Prosperina Vitale

Abstract The shift to distance learning during the Covid-19 pandemic generated a wide use of online services and platforms. The present contribution - based on a survey of students enrolled in an undergraduate degree course - aims at exploring the impact of online education on students' learning experiences and habits. Using a network analysis approach, we investigated students' learning perceptions, the types of social support, the collaborative behaviours students' learning perceptions and the interconnection of these aspects with the individual performance.

Abstract *Il passaggio alla didattica a distanza durante la pandemia da Covid-19 ha generato un ampio utilizzo di servizi e piattaforme online. Questo contributo – basato sulla realizzazione di un'indagine rivolta agli studenti iscritti ad un corso di laurea triennale - è finalizzato a esplorare l'impatto della didattica online sulle abitudini e sulle modalità di apprendimento degli studenti. Utilizzando un approccio di network analysis è stato possibile indagare le percezioni di apprendimento degli studenti, le tipologie di supporto sociale, i comportamenti collaborativi che emergono negli ambienti virtuali e l'interconnessione di tali aspetti con le performance individuali.*

Key words: Digital skill, Ego-centered network, Social network analysis, Social support, Online learning experience

Ilaria Primerano, Giuseppe Giordano, Maria Prosperina Vitale
Department of Political and Social Studies, University of Salerno, Italy e-mail: iprimerano@unisa.it; ggiordano@unisa.it; mvitale@unisa.it

Maria Carmela Catone
Department of Sociology, University of Barcelona, Spain e-mail: mcatone@ub.edu

1 Introduction

The Covid-19 pandemic has affected the education system forcing universities to adopt emergency forms of teaching by using online digital platforms [1] [2] [3]. Since the first phase of the pandemic, university students experienced different modalities of online teaching, from full distance to blended teaching, showing their resilience in dealing with different situations imposed by the health emergency. The dematerialization of teaching and learning processes have produced several effects on students experience in terms of spaces, times, tools, methods, contents and social relationships.

Several studies investigated the factors affecting university life during pandemic. Some authors explored the expectations underpinning the distance learning processes supported by digital tools and resources [4]. The technological, pedagogical, financial, and social issues were also studied [5] [6]. Furthermore, students' perceptions have been investigated by considering the role of prior digital skills [7], the effects on individual performance [8], the satisfaction for the online teaching and services [9], as well as the kinds of collaborative behaviours and social support among individuals [10].

Moving from this theoretical framework, we aim to address the following research questions: *i) How do prior individual digital skills affect students' abilities during distance learning? ii) Which kind of social support and collaborative behaviours are developed among students during pandemic? iii) How do students perceive and define the online learning experience?*

Based upon these questions, a pilot study is carried out on students enrolled in an undergraduate program to investigate how these aspects are related to individual performances. Different network data arises in terms of adjectival networks and social support networks.

The contribution is structured as follows. Section 2 presents questionnaire design, data collection and methods to deal with network data structures. Preliminary findings and some concluding remarks are reported in section 3.

2 Data collection and methods

This contribution aims at studying the effects of digital learning and digitalization process on university students' experiences and collaborative behaviours.

A pilot study has been designed in order to investigate different aspects, regarding the prior digital skills to support learning, the use of online platforms, the university life experience with its emotional connotation, the kinds of social support arising in virtual environments, the students' learning abilities and their performance. The research instrument designed to collect primary data is an online questionnaire structured in several sections, as described in Table 1.

Specifically, different network structures can be derived by the *perception* and *social support* sections. As regards students' perceptions of distance learning expe-

rience, semantic differential scales give raise to the definition of networks of adjectives [11, 12]. Twelve pairs of opposite adjectives have been chosen, describing the different aspects related to *distance learning activities*, *use of digital platform*, *online interactions*, and *self-study attitude*. Thus, the semantic differential scores are used to define valued networks characterizing the emotional meaning of the investigated concepts.

In the social support section, a 7-points Likert scale has been considered to measure students' agreement about *affiliative*, *informational*, *emotional*, and *instrumental* support items [13]. Moreover, a role relation approach [14] has been adopted to study the tangible and intangible relationships. As result, it is obtained a description of roles played by those figures (colleagues, friends, parents, etc.) who were involved in the students' support networks, helping them during pandemic.

Table 1: Questionnaire sections and descriptions

Section	Description
National and international mobility	Student's choice to move in another university within the national territory or to participate in the Erasmus+ international mobility programme
Use of digital tools	Kinds of device to access online platforms, ease and frequency in the use of interactive tools, and activities carried out together with other students and teachers
Prior digital skills	Skills measured according to the key components developed by the European Digital Competence Framework for Citizens (EU DigComp 2.1)
Perception	Students' perception of distance learning defined by opposite pairs of adjectives included in Semantic Differential scale
Social support	Students' collaborative behaviours emerging in virtual environments and the different kind of support to face with specific issues
Satisfaction	Students' opinions on distance learning experience related to platform's functionality, learning capabilities, organization of activities, effectiveness of distance learning, quality of the relationship between peers and teachers
Services evaluation	Students' assessment on usability of distance learning technological solutions, effectiveness of online teaching methods, and overall satisfaction
Performance	Students' exams grade, number of exams passed, and career progression

3 Preliminary results

The first results about the social support section and the network approach are reported. Specifically, as regards the analysis of perceived social support scales, Fig.1

shows students agreement about the twelve items used to measure the four considered dimensions: affiliative (*I1-I3*), informational (*I4-I6*), emotional (*I7-I9*), and instrumental (*I10-I12*). The seven-points Likert scale ranges from ”fully disagree” to ”fully agree” with a neutral option at the midpoint. In the stacked bar chart in Fig.1 each item corresponds to a bar centred on the neutral value (score 4) located in correspondence of the *O%* value on the horizontal axis. This value splits the disagreement responses (scores 1 to 3, on the left) from the agreement ones (scores 5 to 7, on the right). The responses with neutral score has been equally divided between agreement and disagreement.

Notice that, we combine in the same scale ten positively formulated items and four reversed ones (i.e., *I2*, *I4*, *I7*, and *I9*). Our findings show that respondents feel confident about their perceived social support circle, as they positively rated almost all items. This is particularly evident for the items related to the instrumental support: students are aware that among their acquaintances there is somebody ready to help them if they were dealing with a problematic event. Students also feel integrated in their community of friends with whom spent their time and share their personal affairs. They think to be able to maintain friendship relations and that there is certainly someone who is proud of their achievement.

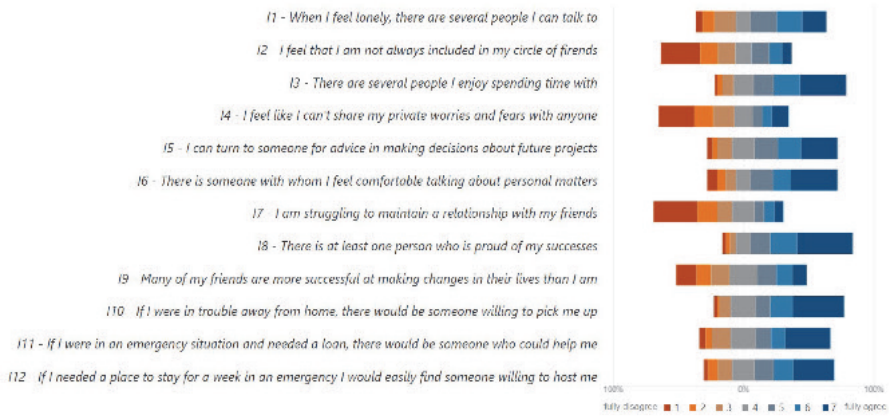


Fig. 1: Stacked bar chart of the perceived social support dimensions: affiliative (*I1-I3*), informational (*I4-I6*), emotional (*I7-I9*), and instrumental (*I10-I12*).

To outline the person profiles to whom students turn for social support, we asked to identify the three profiles they have addressed or would address to deal with specific circumstances. Fig.2 represents the bipartite graphs of the two-mode networks derived from both hypothetical and concrete social support dimensions, where links are defined only between nodes belonging to the two different sets (i.e., students and person profiles). Nodes are shaped and coloured according to the set they belong to and the type of support. Red square is used for person profiles, dots represent

students requiring concrete (black), and hypothetical (grey) support. The network is analysed by looking at the role played by the most relevant person profiles. In general, family of origin members, friends, and students enrolled in the same undergraduate program are the most central profiles in terms of network centrality degree. At the same time, the network representation allows to enhance the joint presence of different person profiles chosen by the same student. For instance, in Fig.2b on the top right, we notice a student pointing out to counseling, student association and professor when asking for emotional support. Specifically, as regards the *instrumental support* (Fig. 2a) some differences can be seen when comparing the relationships established for concrete and hypothetical support. In the first case, students mainly refer to peers and extended family members, while in the second one presence of family of origin (mother and father, brother and sister) is prominent. Furthermore, for *emotional support* (Fig.2b) students address, above all to their family of origin, then in the *informational support* network (Fig. 2c) the importance of peer relationships clearly emerges due to the relevance of both students enrolled in the same university program and friends.

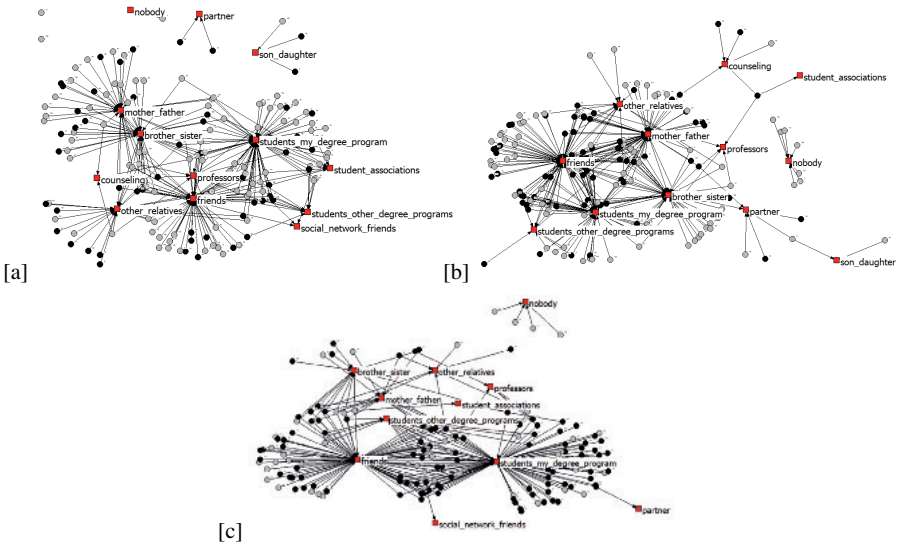


Fig. 2: Bipartite graph visualization of two-mode networks *students*-by-*support person profiles*: a) instrumental support; b) emotional support; c) informational support. Node color: hypothetical support (grey), concrete support (black), types of support person profiles (red)

Furthermore, the survey provides a detailed picture of students experience by highlighting levels of awareness and socialization towards digital tools. The interconnection between technological and social components characterizing the distance learning process is considered. In addition, the effect of relationships, defining

the university social life experience, are evaluated in explaining students' academic performance by controlling for individual covariates.

References

1. Karakaya, K.: Design considerations in emergency remote teaching during the COVID-19 pandemic: A human-centered approach. Educ. Technol. Res. Dev. **69**(1), 295–299 (2021)
2. Hodges, C.B., Moore, S., Lockee, B.B., Trust, T., Bond, M.A.: The difference between emergency remote teaching and online learning (2020)
3. Whalen, J.: Should teachers be trained in emergency remote teaching? Lessons learned from the COVID-19 pandemic. J. Technol. Educ. **28**(2), 189–199 (2020)
4. Cicha, K., Rizun, M., Rutecka, P., Strzelecki, A.: COVID-19 and higher education: first-year students' expectations toward distance learning. Sustainability **13**(4), 1889 (2021)
5. Tejedor, S., Cervi, L., Pérez-Escoda, A., Tusa, F., Parola, A.: Higher education response in the time of coronavirus: perceptions of teachers and students, and open innovation. J. Open Innov.: Technol. Mark. Complex. **7**(1), 43 (2021)
6. Giovannella, C., Marcello, P., Donatella, P.: The effects of the Covid-19 pandemic on Italian learning ecosystems: The school teachers' Perspective at the steady state. Interact. Des. Archit.(s) **45**, 264–286 (2020)
7. Almaiah, M.A., Al-Khasawneh, A., Althunibat, A.: Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic. Educ. Inf. Technol. **25**(6), 5261–5280 (2020)
8. Dickson, M. M., Espa, G., Giuliani, D., Micciolo, R.: Did Covid-19 restrictions in higher education affect students' performances? Evidence from an Italian university. J. Appl. Stat. **33**(2), 123–142 (2021)
9. Schijns, J.M.: Measuring service quality at an online university: using PLS-SEM with archival data. Tert. Educ. Manag. **27**, 161–185 (2021)
10. Furfaro, E., Rivellini, G., Pelle, E., Zaccarin, S.: Constructing personal networks in light of COVID-19 containment measures. Genus **77**, 17 (2021)
11. Primerano, I., Catone, M.C., Giordano, G., and Vitale, M.P.: Assessing undergraduate students' perceptions of distance learning during the COVID-19 pandemic. In: Lombardo, R., Camminatiello, I., and Simonacci, V. (eds.) Book of Short papers. IES 2022 Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, PKE s.r.l., 263–268 (2022)
12. Giordano, G., Primerano, I.: The use of network analysis to handle semantic differential data. Qual. Quant. **52**(3), 1173–1192 (2018)
13. Moretti, M., Simonelli, A., Melloni, M., Ronconi, L.: Interpersonal Support Evaluation List (ISEL): un contributo alla validazione e all'applicazione nel contesto italiano. Psicol. Soc. **7**(3), 419–441 (2012)
14. McCallister, L., Fischer, C.S.: A procedure for surveying personal networks. Sociol. Methods Res. **7**(2), 131–148 (1978)

Physical and cultural activity, internet use and anxiety of Italian university students during the pandemic

Attività fisiche e culturali, uso di internet e ansia degli studenti universitari italiani durante la pandemia

Giovanni Busetta, Maria Gabriella Campolo and Demetrio Panarello

Abstract

The lockdown tends to produce an increase in mental distress such as anxiety through physical and psychological channels, which are correlated with daily life habits such as disturbed eating and sleeping behaviors, as well as variations in physical and cultural activity and internet use. The aim of this paper is to measure the impact of physical, cultural and internet-related activities on anxiety. Specifically, we focused on the differences in this relationship between the pre-pandemic and pandemic periods. To do so, we administered two questionnaires to several university students from three different universities belonging to the South, Center and North of Italy, at the beginning of the pandemic and at one year's distance. The survey also included a psychological test, the STAI-Y, which measures state and trait anxiety.

Abstract

Il lockdown tende a produrre un incremento del disagio mentale come l'ansia, che sono correlati alle abitudini della vita quotidiana. Tali variazioni possono riguardare disturbi nel mangiare e nel dormire, nonché variazioni nell'attività fisica e culturale e nell'uso di Internet. Lo scopo di questo lavoro è misurare l'impatto dell'attività fisiche, culturale e l'utilizzo di Internet, sull'ansia. Nello specifico, ci siamo concentrati sulle differenze in questa relazione tra il periodo pre-pandemia e quello pandemico. Per farlo abbiamo somministrato due questionari agli studenti universitari di tre diversi atenei del Sud, Centro e Nord Italia, all'inizio della pandemia e a distanza di un anno. L'indagine utilizza un test psicologico, lo STAI-Y, che misura l'ansia di stato e di tratto.

Key words: STAI test; Coronavirus outbreak; Pandemic; Mental health; Physical activity; Internet use; State and Trait Anxiety.

1 Introduction

Physical activity (from now on PA) is defined as "any bodily movement produced by skeletal muscles that results in energy expenditures… Exercise is a subset of physical activity that is planned, structured, and repetitive and has as a final or an intermediate objective the improvement or maintenance of physical fitness" (Caspersen et al., 1985). 150 min of moderate or 75 min of vigorous PA activity per week is considered to be protective against factors such as depression (Schuch et al., 2018) and anxiety (Schuch et al., 2019), according to the International PA guidelines (WHO, 2018).

Several studies detected a decrease in PA in multiple countries (Bauer et al., 2020; Schuch et al., 2020; Mattioli et al., 2020) during the current pandemic.

Also, the relationship between internet use and anxiety is considered controversial by the main literature. On the one side, internet-use disorders are generally caused by depression, anxiety, and loneliness (Longstreet et al., 2019). On the other side, gambling, online gaming, and social media use may produce states of anxiety and depression (Brand et al., 2019).

Considering specifically the current COVID-19 pandemic, the connected period of quarantine and the consequent reduction of social contacts acted as stressors on anxiety, negative emotions (Al-Kamdari and Al Sejard, 2020; Gao et al., 2020) and economic difficulties. The association between anxiety and internet use has been specifically studied and identified for the current pandemic (Planchuelo-Gomez et al., 2020; van der Velden et al., 2020).

According to several scholars, being male and/or young increases the risk of internet overuse. Some authors (Elhai et al., 2019; Yang et al., 2020) showed a significant correlation between problematic use of smartphones and mental health symptoms such as depression and anxiety. Finally, the increase in internet use could also be connected to the uncertain future, because of the need to escape from reality (Biolcati et al., 2017), and this increase is usually more pronounced for individuals presenting higher levels of anxiety (see Masaeli and Farhadi, 2021 for a review).

The aim of this paper is to measure the impact of physical, cultural and internet-related activities on anxiety. Specifically, we focused on the differences in this relationship between the pre-pandemic and pandemic periods. To do so, we administered two questionnaires to several university students from three different universities belonging to the South, Center and North of Italy, at the beginning of the pandemic and at one year's distance. The survey also included a particular psychological test, the STAI-Y, which measures state and trait anxiety.

Most previous studies with a similar research question are at risk of being affected by sample self-selection, due to the delicate nature of the topic. In our study, we overcome this problem by observing anxiety levels of the same individuals at different points in time (pre-pandemic, at the beginning of the pandemic and at one year's distance).

2 Data and Methods

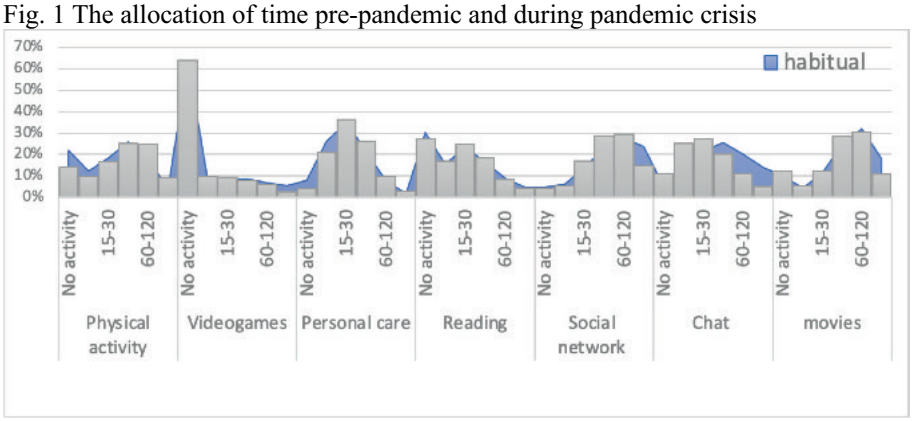
We collected data from an ad hoc questionnaire, administered to the students of three Italian universities located in Messina, Udine, and the Marche Polytechnic University one year after the lockdown, when Italy has been divided into three "colored" zones (red, orange, yellow), characterized by different restrictive measures on the basis of the severity of the spread of COVID-19 at the regional level (Panarello and Tassinari, 2021). The survey collected information on a total sample of 3,580 students.

In addition to gathering demographic information, the survey addresses the following topics: economic, labor, context-based, on-line teaching, time use, and psychological well-being features. About the last topic, to evaluate the level of anxiety we administered the State-Trait Anxiety Inventory (STAI-Y) composed of 40 items rated on a 4-point Likert scale. The first 20 items refer to the respondents' feelings "at this moment" (State), while the last 20 items evaluate the "usual" frequency of feelings (Trait). The score of each subtest ranges between 20 and 80. While the mean value of the State score is 54.85 (Std. Dev. 12.015); the Trait score one is 46.34 (Std. Dev. 10.940).

We collected information about the allocation of time considering two different moments, pre-pandemic and during the pandemic in the following activities: physical activity, videogames, personal care, reading, social network, chat and watching movies.

In the following graph (Fig. 1) we show the time (minutes per day) dedicated to different activities. For each activity we considered the following scale: No activitiy; less than 15 minutes; between 15 and 30 minutes; between 30 and 60 minutes; between 60 and 120 minutes and finally more than 120 minutes. In blue we show the distribution of the time devoted to the activities pre-pandemic and in grey the time spent during the pandemic.

Fig. 1 The allocation of time pre-pandemic and during pandemic crisis



Students who declare they did not exercise before the pandemic were 22% versus 14% during the pandemic, and the percentage of those who perform more than one hour of physical activity increases from 22% (15% 60-120 minutes; 4% over 120 minutes) to 34% (25% and 9%, respectively). More than 60% of students said they did not play video games, while the percentage of those who played more than an hour in the pre-pandemic period was 13% (6% during the pandemic). The percentage of those who did not dedicate time to read decreased, and the percentage of students who spent at least half an hour on this activity during the pandemic increased. Finally, concerning the time dedicated to watch movies, we can see no relevant differences in the trend of the two distributions.

3 Methods and Results

We separately estimate state anxiety and trait anxiety to evaluate the different impact of PA, internet use and cultural activities on habitual (Trait score) and current (State score) level of anxiety. While trait anxiety is defined as "an individual's predisposition to respond and state anxiety as a transitory emotion characterized by physiological arousal and consciously perceived feelings of apprehension, dread, and tension" we use the two typologies of anxiety to measure pandemic and pre-pandemic mental distress.

In the following table we report the results of our models. First, in terms of gender inequalities, being a woman produces an increase in anxiety, both habitual and current. High levels of trait anxiety are correlated with higher increases in anxiety during pandemic. Studying at Università Politecnica delle Marche is correlated with higher levels of anxiety compared to University of Messina and Udine. Furthermore, being more than 26 years old is correlated with higher both pandemic and pre-pandemic levels of anxiety.

PA is always correlated with lower levels of anxiety and this relationship is more pronounced for anxiety during pandemic, with respect to what happened previously. Videogaming increases trait anxiety when it overcome 30 minutes a day, but it has not any statistically significant effect on state anxiety. Social media increases both trait and state anxiety, but this relationship hold only for high levels of its use. The act of reading is negative correlated with both the categories of anxiety. Watching movies is correlated with higher levels of both trait and state anxiety, but this relationship appears only if more than two hours a day are devoted to this activity. Time devoted to personal care is negatively correlated with both pandemic and pre-pandemic anxiety. Not using chat at all is correlated with lower levels of trait anxiety, while using chat more than two hours a day produces an increase in state anxiety.

	State score Coef	Trait Score Coef
Woman	4.20***	2.96***
Age (ref. 18-21)		
22-25	0.84	0.49
26 and more	1.51**	1.17*
University (Ref. Messina)		
Politecnica delle Marche	1.30**	0.21
Udine	0.22	-0.04
Trait anxiety (score)	0.41***	
Physical activity (ref. No activity)		
less than 15 mins	-1.61*	0.46
15 to 30 mins	-2.45***	-0.59
30 to 60 mins	-2.37***	-1.26*
1 to 2 hours	-3.24***	-1.97**
more than 2 hours	-3.06**	-2.30**
Videogames (ref. No activity)		
less than 15 mins	0.33	0.61
15 to 30 mins	0.27	-0.03
30 to 60 mins	1.02	1.86**
1 to 2 hours	-0.51	1.55*
more than 2 hours	-1.27	2.45*
Personal care (ref. No activity)		
less than 15 mins	-2.44***	-3.81***
15 to 30 mins	-4.49***	-5.04***

Contribution Title		
30 to 60 mins	-4.56***	-6.07***
1 to 2 hours	-6.04***	-6.39***
more than 2 hours	-6.75***	-6.16***
Reading (ref. No activity)		
less than 15 mins	-0.67	0.16
15 to 30 mins	-1.00*	-0.23
30 to 60 mins	-1.91***	-0.5
1 to 2 hours	-1.53*	-0.18
more than 2 hours	-1.78*	-0.41
Social network (ref. No activity)		
less than 15 mins	0.73	-1
15 to 30 mins	0.45	0.33
30 to 60 mins	2.17*	1.36
1 to 2 hours	3.74***	2.75**
more than 2 hours	5.46***	3.17**
Chat (ref. No activity)		
less than 15 mins	0.36	-1.35*
15 to 30 mins	0.06	-0.48
30 to 60 mins	0.66	-1.15
1 to 2 hours	0.39	-1.48
more than 2 hours	1.35	-0.26
Watching movies (ref. No activity)		
less than 15 mins	1.05	0.87
15 to 30 mins	0.45	0.33
30 to 60 mins	0.08	0.49
1 to 2 hours	0.09	0.83
more than 2 hours	0.52	2.77***
Constant	34.66***	48.00***

References

1. Al-Kandari, Y.Y., Al-Sejari, M.M., 2020. Social isolation, social support and their relationship with smartphone addiction. Inf. Commun. Soc. 1–19. doi:10.1080/1369118X.2020.1749698.
2. Bauer LL, Seiffer B, Deinhart C, Atrott B, Sudeck G, Hautzinger M, et al. (2020), Associations of exercise and social support with mental health during quarantine and social-distancing measures during the COVID-19 pandemic: a cross-sectional survey in Germany. medRxiv. https:// doi. org/ 10. 1101/ 2020. 07. 01. 20144 105.
3. Biolcati R, Mancini G, Trombini E. Brief report: the influence of dissociative experiences and alcohol/ drugs dependence on Internet addiction. Mediterr J Clin Psychol. 2017;5(1). https://www.researchgate. net/publication/322577619_Brief_report_The_influence_of_dissociative_experiences_and_alcoholdrugs_ dependence_on_Internet_addictions
4. Brand, M., Wegmann, E., Stark, R., Müller, A., Wölfling, K., Robbins, T.W., Potenza, M.N., 2019. The Interaction of Person-Affect-Cognition-Execution (I-PACE) model for ad- dictive behaviors: update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. Neurosci. Biobehav. Rev. 104, 1–10. doi:10.1016/j.neubiorev.2019.06.032.
5. Caspersen CJ, Powell KE, Christenson GM. (1985), Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep.;100(2):126-31. PMID: 3920711; PMCID: PMC1424733.
6. Elhai JD, Levine JC, Hall BJ. The relationship between anxiety symptom severity and problematic smartphone use: a review of the literature and con- ceptual frameworks. J Anxiety Disord. 2019;62: 45–52. doi:10.1016/j.janxdis.2018.11.005.
7. Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., ... Dai, J., 2020. Mental health problems and social media exposure during COVID-19 outbreak. PLoS One 15 (4), e0231924. doi:10.1371/journal.pone.0231924.

8. Longstreet, P., Brooks, S., Gonzalez, E.S., 2019. Internet addiction: when the positive emotions are not so positive. Technol. Soc. 57, 76–85. doi:10.1016/j.techsoc.2018.12.004.
9. Masaeli N, Farhadi H. Prevalence of Internet-based addictive behaviors during COVID-19 pandemic: a systematic review. J Addict Dis. 2021 Oct-Dec;39(4):468-488. doi: 10.1080/10550887.2021.1895962. Epub 2021 Mar 22. PMID: 33749537.
10. Mattioli AV, Sciomer S, Cocchi C, Maffei S, Gallina S. (2020) Quarantine during COVID-19 outbreak: changes in diet and physical activity increase the risk of cardiovascular disease. NutrMetabCardiovasc Dis;30(9):1409–17. https:// doi. org/ 10. 1016/j. numecd. 2020.05. 020.
11. Planchuelo-Gómez, Á., Odriozola-González, P., Irurtia, M.J., de Luis-García, R., 2020. Longitudinal evaluation of the psychological impact of the COVID-19 crisis in Spain. J. Affect. Disord. doi:10.1016/j.jad.2020.09.018, S0165032720327130.
12. Schuch FB, Vancampfort D, Firth J, Rosenbaum S, Ward PB, Silva ES, et al. (2018), Physical activity and incident depression: a meta-analysis of prospective cohort studies. Am J Psychiatry.;175(7):631–48. https:// doi. org/ 10. 1176/ appi. ajp. 2018. 17111 194.
13. Schuch FB, Stubbs B, Meyer J, Heissel A, Zech P, Vancampfort D, et al. (2019) Physical activity protects from incident anxiety: a meta-analysis of prospective cohort studies. Depress Anxiety.;36(9):846–58. https:// doi. org/ 10. 1002/ da. 22915.
14. Schuch FB, Bulzing RA, Meyer J, L.pez-S.nchez GF, Grabovac I, Willeit P, et al. (2020) Moderate to vigorous physical activity and sedentary behavior change in self-isolating adults during the COVID-19 pandemic in Brazil: A cross-sectional survey exploring correlates. medRxiv. https:// doi. org/ 10. 1101/ 2020. 07.15. 20154 559.
15. van der Velden, P.G., Fonds Slachtofferhulp, C.C., Das, M., van Loon, P., Trauma, van, L.C.S., Bosmans, M, 2020. Anxiety and depression symptoms, and lack of emo- tional support among the general population before and during the COVID-19 pan- demic. A prospective national study on prevalence and risk factors. J. Affect. Disord. doi:10.1016/j.jad.2020.08.026, S0165032720326227.
16. WHO. Physical activity. 2018 [cited 15.07.2020]. Available from: https:// www. who. int/ news-room/ fact- sheets/ detail/ physi cal- activity.
17. Yang J, Fu X, Liao X, Li Y. Association of problem- atic smartphone use with poor sleep quality, depres- sion, and anxiety: a systematic review and meta- analysis. Psychiatry Res. 2020;284:112686. doi:10. 1016/j.psychres.2019.112686.

The digital divide in Italy before and during the pandemic phase

Il divario digitale in Italia prima e durante la fase pandemica

Laura Zannella

This study uses microdata from the Community Survey on ICT use by households and individuals to examine the digital divide in Italy on the basis of two dimensions: access to the internet (frequency of use of the network), levels of digital skills. The interest is focused on the analysis of the changes occurred in the digital divide before and during the pandemic phase (2019-2021) with reference both to the overall Italian population aged 16-74 years and to its particular segments. By logistic regression models will be analyzed the changes between the years and their interaction with a set of socio-demographic and territorial variables on both the frequency of Internet use and levels of digital competence

Questo studio utilizza i microdati dell'indagine comunitaria sull'uso delle ICT da parte delle famiglie e dagli individui per esaminare il digital divide in Italia sulla base di due dimensioni: accesso ad internet (frequenza di utilizzo della rete), i livelli delle competenze digitali. L'interesse è concentrato sull'analisi dei cambiamenti intercorsi nel digital divide prima e durante la fase pandemica (2019-2021) con riferimento sia al complesso della popolazione italiana in età 16-74 anni che a suoi particolari segmenti. Mediante modelli di regressione logistica verranno analizzate le variazioni tra gli anni e la loro interazione con un set di variabili socio-demografiche e territoriali sia sulla frequenza di utilizzo d'Internet che sui livelli di competenza digitale.

Key words: digital skill, digital divide, ICT

[1] Laura Zannella, Istat; laura.zannella@istat.it:

Laura Zannella

Introduzione

La pandemia di Covid-19 ha accelerato l'utilizzo delle tecnologie digitali nei diversi ambiti della vita quotidiana evidenziando il potenziale delle ICT (information and communication technology) per consentire l'esercizio dei diritti fondamentali come quelli legati all'istruzione e alla salute, ma ha anche messo in luce come i divari digitali condizionano l'inclusione sociale. Il 19 marzo 2021 la Commissione europea ha proposto una visione per la trasformazione digitale dell'Europa entro il 2030, che si sviluppa intorno a quattro cardini: competenze digitali, infrastrutture, trasformazione digitale delle imprese, digitalizzazione dei servizi pubblici. Nel settembre 2021 la Commissione ha introdotto un quadro di governance per raggiungere gli obiettivi digitali sotto forma di un percorso per il decennio digitale. Successivamente, il 26 gennaio 2022, la Commissione ha proposto al Parlamento europeo e al Consiglio di sottoscrivere una dichiarazione sui diritti e i principi che guideranno la trasformazione digitale nell'UE. Il progetto di dichiarazione riguarda diritti e principi fondamentali per la trasformazione digitale, quali porre al suo centro le persone e i loro diritti, sostenere l'inclusione, promuovere la partecipazione allo spazio pubblico digitale, aumentare la sicurezza, l'autonomia e la responsabilità delle persone e promuovere la sostenibilità del futuro digitale.

L'Istat attraverso l'indagine comunitaria "Uso delle tecnologie dell'informazione e della comunicazione da parte delle famiglie e degli individui" produce annualmente indicatori condivisi a livello europeo per valutare, monitorare e reindirizzare le politiche attuate dal governo in materia di ICT.

Il principale obiettivo di questo lavoro è quello di analizzare l'andamento del divario digitale nel nostro Paese. Mediante l'analisi dei micro-dati dell'indagine Istat si intende rispondere alle seguenti domande di ricerca: tra il 2019 e il 2021 è aumentato il ricorso all'utilizzo di internet? I divari digitali tra i diversi segmenti della popolazione sono aumentati o diminuiti? Ad una maggiore utilizzo della rete ha corrisposto un miglioramento delle competenze digitali?

1.1 Dati e metodi

Questo studio, si basa sui micro dati dell'indagine comunitaria sull'uso delle tecnologie della comunicazione e dell'informazione, condotta annualmente dall'Istat su un campione probabilistico di circa 24.000 famiglie e 54.000 individui. Con l'indagine ICT vengono rilevate informazioni sulle caratteristiche socio-demografiche degli individui (sesso, età, livello d'istruzione, occupazione, risorse economiche, territorio di residenza, etc.) e informazioni dettagliate sull'uso di internet: (tipo di connessione, frequenza di utilizzo, attività svolte, competenze digitali, etc.).

Il concetto di divario digitale viene reso operativo facendo riferimento a due dimensioni: l'accesso e le competenze digitali. Per quanto riguarda l'accesso (divario digitale di primo livello) viene analizzata l'a frequenza con cui ci si

connette ad internet. Sulla base della frequenza sono state individuate tre tipologie di utente: assiduo, è una persona che ha utilizzato internet tutti i giorni nei tre mesi precedenti l'intervista; utente settimanale, che si è collegato ad internet qualche volta a settimana ma non tutti i giorni, l'utente e sporadico che ha utilizzato internet qualche volta al mese nei 3 mesi precedenti l'intervista.

Ad ognuna delle tre definizioni corrisponde una variabile dummy che assume il valore 1 se la persona presenta quella caratteristica e il valore 0 altrimenti. Per analizzare i cambiamenti intervenuti tra il 2019 e il 2021 sono stati utilizzati tre modelli logistici, uno per ciascuna delle seguenti variabili dipendenti: P_1 = probabilità di essere un utente assiduo, P_2 = probabilità di essere un utente settimanale, P_3= probabilità di essere un utente sporadico. Sono state considerate le seguenti variabili esplicative categoriche: Y=Year (2019,2021), X_1= Gender (men, women), X_2= Age group (16-24, 25-34, 35-44, 45-54, 55-64, 65-74), X_3= Education level (high, medium, low), X_4= Occupational level (high, medium, low, unemployed), X_5= Geographic area of residence (north-west, north-est, centre, south, islands), X_6=Type of municipaljty (metropolitan area, over 50.000 inhab, 10,000-50.000 inhab, until 10.000 inhab), X_7=Family economic resource (very googd, good, bad, very bad).

Poiché l'obiettivo dell'analisi è quello di valutare i cambiamenti intercorsi tra i due anni, viene utilizzato un modello logistico comprendente tutti gli effetti semplici e gli effetti interattivi di Year con le variabili esplicative. Indicando con P la variabile dipendente, con m il numero di variabili esplicative, con X_r la r-ma variabile esplicativa e con h_r il numero delle sue modalità, per ciascun individuo i del campione il modello logistico può essere scritto come segue:

$$\text{logit}\left(P_i\right) = \log\left(\frac{P_i}{1-P_i}\right) = \alpha + \sum_{j=1}^{2}\beta_j Y_{ji} + \sum_{r=1}^{m}\sum_{s=1}^{h_r}\gamma_{rsi}X_{ri} + \sum_{j=1}^{2}\sum_{r=1}^{m}\sum_{s=1}^{h_r}\lambda_{jr}Y_{ji}X_{ri} + \varepsilon_i$$

$$i = 1,2,\ldots,n$$

I parametri del modello vengono stimati con il metodo della massima verosimiglianza, utilizzando la procedura proc logistic del Sas. La procedura fornisce le stime dei parametri e i corrispondenti errori standard, calcola i test relativi alla significatività complessiva del modello e delle singole variabili esplicative, consente di effettuare i confronti tra le diverse modalità di ciascuna variabile esplicativa. I confronti tra coppie di modalità sono basati sull'odds ratio, ossia sul rapporto tra gli odds delle due modalità. L'ipotesi nulla di uguaglianza dei due odds viene testata mediante il test di Wald basato sul rapporto tra le funzioni di verosimiglianza.

Successivamente, l'analisi si concentra sui divari digitali di secondo livello legati alle competenze digitali. Per la misurazione delle competenze digitali è stato adottato il quadro concettuale definito dalla Commissione europea in collaborazione con gli istituti nazionali di statistica. Le persone che hanno utilizzato internet negli ultimi 3 mesi vengono classificate sulla base delle attività svolte in quattro domini di competenza digitale: information, comunication, content-creation, problem-solving. Per ciascun dominio sono previsti tre livelli di competenza (below basic, basic, above basic), attribuiti sulla base del numero di attività svolte. A seconda del

dominio di competenza il numero di attività varia da un minimo di quattro a un massimo di nove. Le persone che non hanno utilizzato internet nei tre mesi precedenti l'intervista sono classificate come "without digital skills". I valori dei quattro indicatori di dominio vengono utilizzati per elaborare un indicatore complessivo dei digital skill con quattro modalità:

- above basic. competenze superiori a quelle di base in tutti e quattro i domini.
- basic, competenze di base (o superiori) nei quattro domini
- belowe basic. competenze inferiori a quelle di base in almeno un domini
- none digital skill, nessuna competenza in tutti e quattro i domini

L'analisi delle variazioni dei livelli dei digital skill nella popolazione italiana in età compresa tra i 16 e i 74 anni tra il 2019 e il 2021 è stata condotta applicando un modello logistico lineare analogo al precedente, nel quale la probabilità che un individuo disponga di competenze digitali di base o superiori è messa in relazione con l'anno di riferimento dell'indgine e il set di variabili socio-demografiche e territoriali

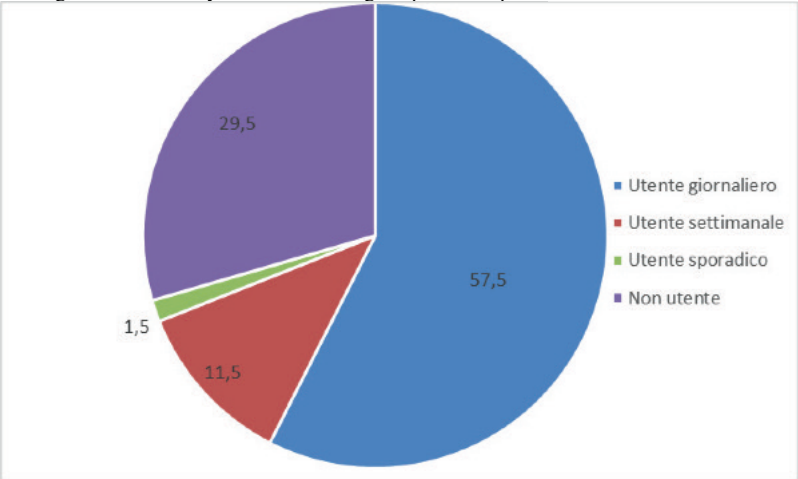
1.2 Risultati preliminari [1]

I risultati descrittivi forniscono una prima evidenza empirica delle relazioni esistenti tra i tassi di utilizzo di Internet e alcune delle caratteristiche "strutturali" della popolazione come il sesso, l'età, il livello di istruzione, status professionale, posizione nella professione, frequenza d'uso di Internet. Come mostrato nella figura1 nel 2020 la distribuzione della popolazione italiana di 6 anni e più genera tre gruppi principali: gli utenti giornalieri 57,5%; i non utenti, circa il 29,5% e le persone che si rapportano a Internet in modo poco intenso. Nel 2020 oltre l'80% dei ragazzi in età 15-24 anni è un utente giornaliero mentre tra i 60-64-enni la quota di internauti assidui scende al 50,1% e arriva al 30% tra le persone di 65-74 anni.

Tra il 2019 e il 2020 si registra un incremento degli utenti di internet. Ad aumentare è la quota di unteti giornalieri (+4 punti percentuali), mentre tra gli utenti settimanali o sporadici non si registrano variazioni significative. Sono soprattutto i ragazzi di 6-14 anni a far registrare gli incrementi maggiori +11,5, incrementi sensibili si registrano anche tra la popolazione anziana di 60-64 anni +6,3. Per quanto riguarda le competenze digitali nel 2019 il 35,8% degli utenti di internet giornalieri di 16-74 anni ha competenze digitali basse, il 27,9% ha competenze di base e solo il 34,6% ha competenze elevate.

Figura1 *Persone di 6 anni e più per frequenza d'uso d'Interne*t .Anno 2020

[1] Le elaborazioni sono limitate al 2020 in quanto i micro dati dell'indagini Istat del 2021 saranno disponibili a breve.

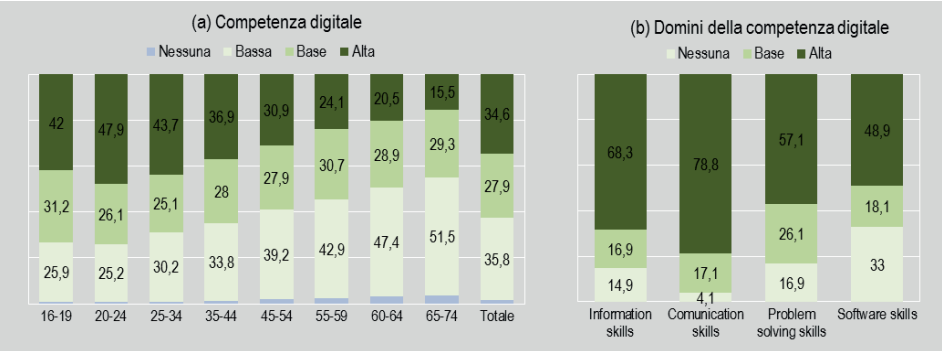


L'età resta un fattore importante ma non decisivo: i giovani di 20-34 anni hanno livelli avanzati di competenze (Figura 2a). Se si analizzano separatamente le quattro dimensioni in base alle quali è calcolato l'indicatore composito emerge che gli internauti hanno competenze digitali più avanzate per e-skill legati al dominio della comunicazione (78,8%) e dell'informazione (68,3%) rispetto a quelli collegati alla capacità di risolvere problemi (57,1%) e di utilizzare software per trattare/veicolare contenuti digitali (48,9%) (Figura 3b).

Table 2*: Persone di 6 anni e più per frequenza d'utilizzo di internet (per 100 persone con le stesse caratteristiche). Anni 2019-2020*

Classi di età	***Utente giornaliero***		***Utente settimanale***		***Utente sporadico***		***Non utente***	
	2019	***2020***	***2019***	***2020***	***2019***	***2020***	***2019***	***2020***
6-14	45,0	56,5	23,7	21,6	3,5	1,5	27,8	20,4
15-19	83,0	85,6	7,5	6,2	1,0	0,9	8,5	7,3
20-24	84,3	85,0	6,6	7,9	1,1	0,5	8,0	6,5
25-34	78,1	81,7	9,0	7,6	0,6	0,7	12,2	9,9
35-44	72,1	76,9	12,3	9,6	1,4	1,4	14,2	12,1
45-54	62,8	67,9	15,4	13,4	2,0	1,6	19,8	17,1
55-59	54,2	59,4	16,2	14,7	2,1	2,1	27,6	23,8
60-64	43,8	50,1	16,3	15,9	2,0	2,0	38,0	32,0
65-74	26,6	30,0	12,6	12,6	2,7	2,2	58,1	55,2
75+	6,9	7,3	4,1	4,8	1,0	1,4	88,1	86,4
Totale	53,5	57,5	12,6	11,5	1,8	1,5	32,1	29,5

Figure 2: *Utenti di 16-74 anni per livello di competenza digitale . Anno 2019*



References

1. Compaine, B. 2001. The digital divide: facing a Crisis or creating a Mith?. MIT Press: Cambridge (Mass)
2. Di Maggio, P.J., Hargattai, E (2002). From the 'digital divide' to 'digital inequality' studying internet use as penetration increases. Working Papers 47, Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Arts and Cultural Policy Studies.
3. Hargattai, E. (2002). Second Level Digital Divide: Differences in People's Online Skills, First Monday (7) 4
4. European Commision (2016). The Digital Competence Framework 2.0 https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework.
5. European Commision (2021) Decennio digitale https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_it

Covid-19 and financial professional advice

Consulenza finanziaria professionale e Covid-19

Marianna Brunetti and Rocco Ciciretti

Abstract This paper empirically investigates the use of financial professional advice by households and the effects on it of the outbreak of Covid-19 pandemic. Based on a new panel dataset provided by Consob, we estimate that the households suffering an income reduction induced by the lockdown due to the Covid-19 outbreak reacted resorting to professional financial advice more than their counterparts having a stable income during the same period. Dissecting the results, we show that this effect is driven by the more financially sophisticated households. These results survive to a variety of robustness checks and support the hypothesis that financial literacy might act as a complement of, rather than a substitute to, financial professional advice.
Abstract *Il lavoro analizza empiricamente la decisione di avvalersi di un consulente finanziario durante la pandemia da Covid-19, sulla base di un dataset appositamente predisposto in collaborazione con Consob. Le stime ottenute mostrano come le famiglie che hanno subìto una riduzione di reddito a causa delle misure restrittive volte ad arginare la diffusione del Covid-19 abbiano reagito rivolgendosi a consulenti finanziari professionali con maggiore probabilità rispetto a famiglie comparabili ma che durante lo stesso periodo non hanno subito riduzioni di reddito. Analisi di eterogeneità mostrano come questo risultato sia guidato dalle famiglie con maggior grado di fiducia nella figura del consulente finanziario e da quelle con maggiori conoscenze finanziarie, supportando l'ipotesi che sofisticazione finanziaria e consulenza professionale siano complementi piuttosto che sostituti.*

Key words: financial professional advice, financial literacy, covid-19, household finance, trust

[1] Marianna Brunetti, University of Rome Tor Vergata, CEFIN & CEIS; email: marianna.brunetti@uniroma2.it

Rocco Ciciretti, University of Rome Tor Vergata, CEIS & RCEA-Rimini Fellow.; email: marianna.brunetti@uniroma2.it

1 Introduction

The outbreak of Covid-19 pandemic in the early months of 2020 has subverted the lives and projects of households under many aspects, albeit the economic consequences of the imposed closures and lockdown have been heterogeneous, depending e.g. on the employment status and the type of occupation (see e.g. Guiso and Terlizzese, 2020). In this paper we empirically investigate the effect it had on the use of financial professional advice. The first key strength of our empirical strategy is that the Covid-19 break we exploit was largely unexpected, thus providing an ideal quasi-natural experiment, which we exploit using a difference-in-differences methodology to compare propensities of a treated group (the households working in the private sector that underwent a sensible reduction of their income due to the Covid-19 first outbreak) with a control group of households (those working in the public sector and those in the private sector having not experienced an income reduction). Hence, we are plausibly able to identify a causal relationship between the reduction in the income, and the hike in financial advice search among the households economically affected by the Covid-19 restrictions and closure implemented to face the first outbreak of the pandemic. Another strength of this study is that the empirical analysis relies on a new and representative household panel survey dataset, provided by the Consob, the public authority responsible for regulating the Italian financial market, which allows us to control for a significant set of household characteristics, not only including the classical demographic and socio-economic controls, but also measures of trust and a pretty thorough measure of financial literacy.

2 Methodology

The dataset is obtained based on a survey administered every year since 2017 by GFK to around 3,000 randomly selected households, half of whom are rotating panel units, producing a final dataset that is population-representative. We use the panel component of the last three waves, referring to 2019, 2020 and 2021. The questionnaire comprises about a hundred questions, ranging from the demographic and socio-economic characteristics of the household, to investment preferences and choices, saving habits, financial knowledge and the consultation of a professional financial advisor, which we use as outcome variable. In the 2020 wave we also have information about whether the household experienced income reduction in the light of the restrictions imposed during the first months of 2020, when the pandemic firstly hit Italy. We use this information, along with the occupational condition of the financial respondent, to identify two groups. One, representing our control group, includes retirees and public employees, who overall have not experienced income reduction and, in many cases, have even increased their savings. The other one, representing our treatment group, includes self-employed, who by construction have to deal with an uncertain income, as well as private employees who declared a

reduction in their income in the verge of the lockdown, having them faced an unprecedented and definitely unexpected shock to their labour income, often having to draw down their savings, if any.

We then estimate the following regression:

$$FA_{it} = \alpha_0 + \alpha_1 \boldsymbol{Treated}_{it} + \alpha_2 \boldsymbol{Post}_t + \alpha_3 (\boldsymbol{Treated}_{it} \times \boldsymbol{Post}_t) + \mathbf{X}_{it}\boldsymbol{\theta} + \varepsilon_{it}$$

where FA_{it} takes value 1 if household *i* at year *t* declared having had contacts with a financial professional advisor in the previous 12 months, and 0 otherwise; $\boldsymbol{Treated}_{it}$ is defined as treatment above $\boldsymbol{Post}_t$ is a binary variable taking the value of 1 if the survey refers to the period after the Covid-19 outbreak, i.e. starting from 2020, and 0 otherwise. Finally, $\mathbf{X}_{it}$ is a vector of control variables, and ε_{it} is the error term

In line with the literature, the vector of controls includes a wide set of socio-economic and demographic characteristics that may affect the households' economic decisions (see, e.g., Guiso and Sodini, 2012), namely area of residence, income and net wealth, household size, and homeownership as well as age, gender, marital status, education employment status of the financial respondent. Moreover, we also control for the level of financial knowledge, assessed based on a set of 5 questions covering notions like the risk-return relationship, real interest rates, compound interest rate, diversification and mortgages.

We are interested in the coefficient α_3, which we estimate via OLS, as it measures the differential effect of the change induced by the pandemic across the households that underwent an income shock (treated) and those that did not (control). The common trend assumption, crucial in any DiD specification, is verified both graphically and statistically.

3 Main Results

Results reported in Table 3, column (1) prove that household experiencing an income reduction due to the Covdi-19 lockdown are 16,6 percentage points more likely to appeal to financial advice after the pandemic than a comparable household in the control group. Moreover, dissecting by level of financial literacy, in Columns (2) and (3), and by level of trust in the financial advisor, Columns (4) to (7), proves that the results are driven by households with a higher financial literacy and those that actually trust these professional figures. Two conclusions can be drawn: first, financial literacy does not act as a substitute to professional financial advice, it rather acts as a complement to it. This result is in line with what reported e.g. in Kim et al (2021). Second, trust in professional advisors is a key element, thus calling for interventions aimed at further increasing the credibility of these figures.

Table 1: Use of financial advisory before and after the pandemic

Estimates	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full sample	Financial literacy		Trust			
		High	Low	No	Neutral	Yes	Not known
α_1	-0.103	-0.156*	-0.024	0.041	-0.257	-0.209*	-0.032
	(0.069)	(0.081)	(0.134)	(0.204)	(0.185)	(0.121)	(0.127)
α_2	-0.016	0.314	0.604	-0.379	-0.569	0.722	3.144*
	(0.461)	(0.591)	(0.854)	(0.945)	(0.908)	(0.881)	(1.706)
α_3	0.166*	0.246**	-0.111	0.180	0.190	0.358**	-0.754
	(0.092)	(0.106)	(0.189)	(0.234)	(0.215)	(0.170)	(0.471)
Obs	626	189	437	146	187	190	103
R^2	0.270	0.362	0.266	0.438	0.367	0.489	0.508

Robust standard errors in parentheses. All regressions are estimated via least squares and include controls for time and space fixed effects, municipality size, homeownership, income, wealth, household size, age and age squared, gender, civil status, education, employment status, and financial literacy of the financial respondent. The symbols ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

References

1. Guiso, L., & Sodini, P. (2012). Household Finance. An Emerging Field. in G. Constantinides, M. Harris & R. Stulz (Eds.), Handbook of the Economics of Finance.
2. Guiso, L. and D. Terlizzese (2020). Quanto pesa lo shock COVID-19 sulle famiglie. Mimeo
3. Guiso, L., & Viviano, E. (2015). How Much Can Financial Literacy Help?. Review of Finance, 19(4), 1347–1382.
4. Kim, H.H, Maurer, R. & Mitchell O., S. (2021) How financial literacy shapes the demand for financial advice at older ages, The Journal of the Economics of Ageing, 20.
5. Lusardi, A., & Mitchell, O.S. (2014). The Economic Importance of Financial Literacy: Theory and Evidence. Journal of Economic Literature, 52(1), 5-44.
6. Lusardi, A., & Tufano, P. (2015). Debt Literacy, Financial Experiences, and Overindebtedness. Journal of Pension Economics and Finance, 14(4), 332-368.
7. van Rooij, M.C.J., Lusardi, A., & Alessie, R.J.M. (2011). Financial Literacy and Stock Market Participation. Journal of Financial Economics, 101(2), 449-472.
8. van Rooij, M.C.J., Lusardi, A., & Alessie, R.J.M. (2012). Financial Literacy, Retirement Planning and Household Wealth. Economic Journal, 122(560), 449-478.

Bayesian modelling and inference 4

Bayesian functional mixed effects model for sports data

Modelli funzionali Bayesiani a effetti misti per dati sportivi

Patric Dolmeta, Raffaele Argiento and Silvia Montagna

Abstract The use of statistical methods in sport analytics is common practice nowadays. In this work, we propose a hierarchical Bayesian model for describing and predicting the evolution of performance over time for shot put athletes. We address seasonality and heterogeneity in results by means of a linear mixed effects model with heteroskedastic errors. The model provides an accurate description of the performance trajectories and allows for prediction of athletes' performance in future seasons. We apply our method to an extensive real world data set on performance data of professional shot put athletes recorded at elite competitions.

Abstract *L'impiego di metodi statistici per lo studio dello Sport è ormai largamente diffuso. In questo lavoro, proponiamo un modello Bayesiano gerarchico per la descrizione e la previsione di risultati nel tempo per lanciatori del peso. Grazie a un modello lineare ad effetti misti con errori eteroschedastici, affrontiamo stagionalità e eterogeneità nei risultati. Applichiamo il metodo ad un dataset reale di grandi dimensioni contenente i risultati del lancio del peso in un gran numero di competizioni internazionali e osserviamo una soddisfacente descrizione dei dati e la possibilità di quantificare l'incertezza nelle previsioni di performance future.*

Key words: Performance analysis, Bayesian functional data analysis, GARCH models, Sport analytics

Patric Dolmeta
Bocconi University, Via Roentgen 8, Milano e-mail: patric.dolmeta@unibocconi.it

Raffaele Argiento
Università degli Studi di Bergamo, Via dei Caniana 2, Bergamo e-mail: raffaele.argiento@unibg.it

Silvia Montagna
Università degli Studi di Torino, C.so Unione Sovietica 218/bis, Torino e-mail: silvia.montagna@unito.it

1 Introduction

Shot put is a track and field event involving throwing ("putting") the shot as far as possible. Shot put events range over the whole year, with indoor competitions held during Winter months and major tournaments during the Summer. So, results display some sort of seasonality, as it is common to all sportive competitions. We underline that here the term "seasonality" is not used to indicate a cyclical behaviour of observations over time, as in the literature of time series but, rather, a time dependent gathering of observations. On one hand competitions are traditionally concentrated in some months of the year, and on the other hand weather and environmental conditions may affect the performances or even the practicability of the sport itself. In shot put, it is reasonable to say that seasons coincides with calendar years and that taking seasonality effects into account is necessary to provide an accurate representation of the data.
In this work, we are interested in describing the evolution of performances of professional shot put athletes throughout their careers. We describe results of each athlete as error prone measurements of seasonal means trough a Bayesian mixed effects model, that describes the seasonal mean for each athlete as a deviation from a grand mean.

2 The World Athletics shot put data set

The data was obtained from an open results database (www.tilastopaja.eu) following institutional ethical approval (Prop_72_2017_18). The dataset comprises 41,000 measurements of World Athletics (the world governing body for track and field athletic sports) recognized elite shot put competitions for 653 athletes from 1996 to 2016. For each athlete, the data set reports the date of the event, the shot distance in meters, an indication of doping violation and some demographic information.
The outcome of interest is the shot distance. Data are collected over time: hereafter we will denote as t_{ij} the time at which the j-th observation for athlete i is recorded. t_{ij} corresponds to the time elapsed from January 1st of each athlete's career starting year to the date of the competition. Having described seasons as calendar years, athletes will compete in a different number of seasons according to their career length. Figure 1 shows the number of athletes per season as well as boxplots of the distribution of their mean performances across the various seasons. A general increasing trend in performance can be observed as a function of career length (right panel in Figure 1).

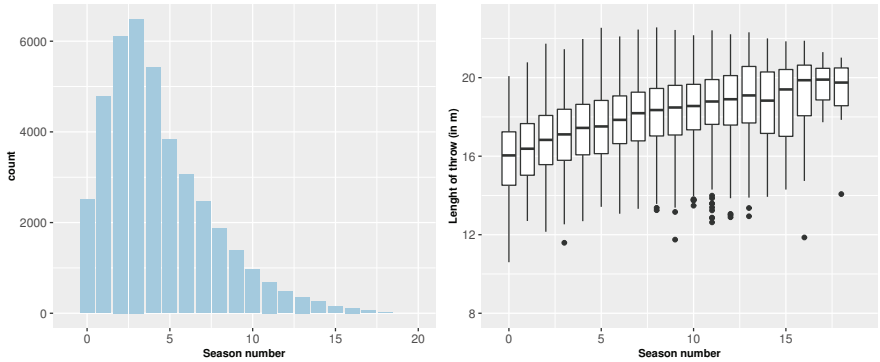


Fig. 1 Left: total number of athletes per season. Right: each boxplot shows the distribution of the athletes' mean performances within each season.

3 The model

Let n be the total number of athletes in the study. Shot put performances for athlete i collected over time represent temporal grouped data that we expect to be correlated when associated to the same individual and season. Hence, we propose a mixed-effects model for the shot put performances for athlete i at time point t_{ij} (y_{ij}), inducing a hierarchical structure in the error variance. Accordingly, y_{ij} are deviations from a seasonal mean function $\mu_i(\cdot)$ that takes constant value μ_{is} for any time point in a given season s:

$$y_{ij} = \mu_{is} + \varepsilon_{ij} \tag{1}$$

with $\varepsilon_{ij} \overset{iid}{\sim} N(0, \psi^2)$ independent errors recorded at time t_{ij}, for $j = 1, \ldots, n_i$ and n_i is the total number of measurements available on athlete i. We further regard the seasonal- and unit-specific random intercept μ_{is} as the latent effect that quantifies the extent to which performances in season s responds above or below the overall mean m:

$$\mu_{is} = m + \zeta_{i,s}, \quad \text{with} \quad \zeta_{i,s} \overset{iid}{\sim} N(0, h_{i,s}) \tag{2}$$

Again, residuals at this higher level of hierarchy are assumed to be normally distributed, uncorrelated with lower-level residuals, but not uncorrelated among themselves. Indeed, the assumption of homogeneity of variance is inadequate here. Early graphical displays, straightforward exploratory analysis and initial modeling choices suggest a significant variability of the average response and its variance across seasons. In particular, suppose the variance of some athlete's performances during a specific time interval is known, then it will provide insight on future variability. Even further, a history of volatility in results provides information about an athlete's potential more than a background of constant, close to the average, performances. Hence, we consider a random intercept model with Normal Generalized Autoregressive Conditional Heteroskedastic (GARCH) errors [1]. Specifically,

$$\mu_{is} \mid m, h_{is} = m + \zeta_{is} \overset{iid}{\sim} N(m, h_{is}) \tag{3}$$

$$h_{is} = \alpha_0 + \alpha_1 \zeta_{is-1}^2 + \varpi h_{is-1} \tag{4}$$

where $\alpha_0 > 0, \alpha_1 \geq 0$ and $\varpi \geq 0$ to ensure a positive conditional variance, and $\zeta_{is} = \mu_{is} - m$ with $h_{i0} = \zeta_{i0} := 0$ for convenience. The additional assumption of wide-sense stationarity with

$$E(\zeta_t) = 0 \tag{5}$$

$$Var(\zeta_t) = \alpha_0 (1 - \alpha_1 - \varpi)^{-1} \tag{6}$$

$$Cov(\zeta_t, \zeta_s) = 0 \text{ for } t \neq s \tag{7}$$

is guaranteed by requiring $\alpha_1 + \varpi < 1$, as proven by [1].
Three parameters of the seasonal component require prior specification: the overall mean m and the conditional variance parameters, ϖ and $\vec{\alpha} = (\alpha_0, \alpha_1)^\top$. For the autoregressive and heteroskedastic parameters of the GARCH model, we propose non-informative priors satisfying the positivity constraint. For the overall mean parameter, we rely on a more informative Normal prior centered around the mean suggested by posterior analysis of preliminary versions of the model. In particular:

$$m \sim N(\mu_{m_0}, \Sigma_{m_0}) \tag{8}$$

$$\vec{\alpha} \sim N_2(\mu_\alpha, \Sigma_\alpha)\, I\{\vec{\alpha} > 0\} \tag{9}$$

$$\varpi \sim N(\mu_\varpi, \Sigma_\varpi)\, I\{\varpi \geq 0\} \tag{10}$$

where $\vec{\alpha} = (\alpha_0, \alpha_1)$ is a bidimensional vector. We complete the model specification assuming that the parameters are statistically independent and noticing that the hypothesis needed for wide-sense stationarity do not translate into actual prior conditions on the parameters. Hence, one of the objects of our analysis becomes to test whether the constraint $\alpha_1 + \varpi < 1$ holds true.
Because of the recursive definition of the conditional variance, no conjugate model exists for the GARCH parameters. Hence, we rely on an adaptive version of the Metropolis Hastings algorithm for posterior updates. In particular, for parameters m, ϖ and $\vec{\alpha}$ we build an adaptive scale Metropolis such that the covariance matrix of the proposal density adapts at each iteration to achieve an *optimal* acceptance rate [2]. We ran our sampler for $50,000$ iterations, with a burn-in period of 50% and a thinning of 5. We analyzed posterior samples for a variety of function estimates at different time points and for a variety of athletes, and the other model parameters. No issues emerged regarding convergence and mixing of the chains.

4 Results

Our goal is in estimating trajectories for athletes' performances. To this end, we generate a fine grid of equispaced time points, $\{t_k\}_{k=1}^{T}$, over our time span and evaluate the trajectory in performance on this grid.
Since we modelled the seasonal mixed effects function as a piecewise continuous function taking individual- and season-specific values, when estimating such function on any point in the time grid, we need to determine which season the time point belongs to. As discussed in Section 2, time is rescaled so that equal values across individuals indicate the same day of the year, possibly in different years. Therefore, season changes, that occur at new year's days, can be easily computed by straightforward proportions. In the following Equation, the indicator variable $\chi_{(t\in s)}$ determines to which season each time point t belongs to and index g ranges over the total number of iterations G:

$$\widehat{y_i(t)} = \frac{1}{G}\sum_{g=1}^{G}\sum_{s=1}^{S_i}\mu_{is}^{(g)}\chi_{(t\in s)} \quad \text{for } t = t_1,\ldots,t_T. \tag{11}$$

Equation (11) represents the point estimate of athlete i's performance at time t. Similarly, 95% credible intervals can be computed to quantify uncertainty around our point estimate. Estimated trajectories, credible bands and one season ahead prediction are displayed in Figure 2 for a random selection of shut put athletes.

5 Discussion

We proposed a hierarchical Bayesian model for the analysis of athletes' performances in a longitudinal context. We addressed the issue of seasonal gathering of sports data with a mixed effects model with GARCH errors, providing evolving random intercepts over different time intervals in the data set. While the motivation of our work comes from the analysis of shot put performance data, the methodology presented in this work is applicable to the analysis of performance data collected in all measurable sports.
The comprehensive nature of the data set suggests exploiting it further, possibly by including the contribution of covariates on the response. In particular, we believe there is potential for a better understanding of the effects of doping, not only on single performances, but on the overall evolution of a career. Further, we hope also doping detection might be targeted. Additional future developments include more sophisticated modeling choices both for the intraseasonal variability and the seasonal intercepts themselves. A nonparametric Bayesian approach to the hierarchy with the intent of clustering observation both across athletes and seasons is already forthcoming. This way we hope to recognize common patterns in similarly evolving careers for enhanced prediction purposes.

References

1. Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics **31**(3), 307 – 327 (1986)
2. Haario, H., Saksman, E., Tamminen, J.: An adaptive metropolis algorithm. Bernoulli **7**, 223–242 (2001)

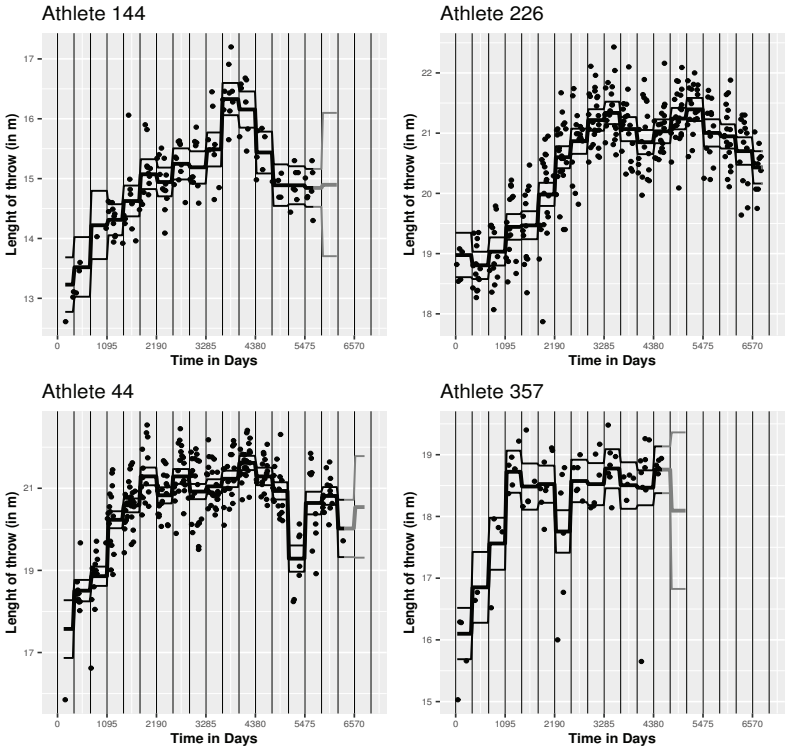


Fig. 2 Performance trajectory estimates for a random selection of athletes. The x-axis denotes the time measured in days from January 1st of the first season of career, whereas on the y-axis there is the length of throw in meters. Vertical lines represent calendar years (seasons in our notation). The final part of each trajectory (grey) for which no observations are available, represents one-season-ahead performance prediction.

Bayesian Optimization with Machine Learning for Big Data Applications in the Cloud

Ottimizzazione Bayesiana Integrata col Machine Learning per Applicazioni Big Data su Cloud

Bruno Guindani, Danilo Ardagna and Alessandra Guglielmi

Abstract Bayesian Optimization is a promising method for efficiently finding optimal cloud computing configurations for big data applications. Machine Learning methods can provide useful knowledge about the application at hand thanks to their predicting capabilities. In this paper, we propose a hybrid algorithm that is based on Bayesian Optimization and integrates elements from Machine Learning techniques to tackle time-constrained optimization problems in a cloud computing setting.
Abstract *L'ottimizzazione bayesiana è un metodo promettente per trovare configurazioni ottimali di applicazioni big data eseguite su cloud. I metodi di machine learning possono fornire informazioni utili sull'applicazione in oggetto grazie alle loro capacità predittive. In questo articolo, proponiamo un algoritmo ibrido basato sull'ottimizzazione bayesiana che integra tecniche di machine learning per risolvere problemi di ottimizzazione con vincoli di tempo in sistemi di cloud computing.*

Key words: acquisition function, cloud computing, Gaussian Process

1 Introduction

Big data analytics are employed in several industries to allow organizations and companies to make better decisions. The most suitable execution environment of big data analytic applications is a cluster of virtual machines (VMs) which allows the adjustment of the allocated resources (CPU, memory, disk, network) to match the application current needs. Choosing the right cloud configuration to minimize execution times and reduce costs is essential to service quality and business competitiveness. However, due to the diverse behavior and resource requirements of analytic jobs, choosing the best configuration for a broad spectrum of applications is a challenging process [1].

Bayesian Optimization (BO) has recently gained notoriety as a powerful tool to solve global optimization problems in which expensive, black-box functions are involved; see the recent paper [6] or the popular tutorial paper [3]. BO is a sequential design strategy that requires few steps to get sufficiently close to the true optimum,

Bruno Guindani[1], Danilo Ardagna[1] and Alessandra Guglielmi[2]
[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy
[2] Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: {bruno.guindani, danilo.ardagna, alessandra.guglielmi}@polimi.it

while requiring no derivative information on the optimized function. Most commonly, it is initialized by choosing and evaluating a small handful of starting points, then fitting a Gaussian process (GP) using these points. The posterior distribution of the fitted GP provides an estimate of both the function value at each point and the uncertainty around the estimate. BO then iteratively chooses new points at which to evaluate the function in a such a way to balance exploration (high uncertainty) and exploitation (best estimated function value), as explained in [4], for instance.

The goal of this work is to integrate Bayesian optimization algorithms with Machine Learning (ML) techniques in the context of cloud computing optimization. The former have proven to be successful [1, 7] in exploring and finding optimal or near-optimal cloud configurations after a small amount of exploratory runs. On the other hand, the latter can provide useful information to be incorporated into the BO mechanism in several ways to improve its performance, for instance in the form of cheap estimates of target quantities to guide the exploration process. This work builds on previous results found in [1], in which the *CherryPick* system is successfully applied to benchmark applications on cloud computing frameworks such as Apache Spark. This system exploits pure BO to find optimal cloud configurations. Our work is motivated by the belief that ML can lend their predicting capabilities to BO to further improve its effectiveness. This topic has been explored in [5], which examines the performance of several ML models in carrying out predictions of execution times of Spark cloud jobs with different types of workloads. The hybrid BO algorithm we propose here is promising since it shows the usefulness of ML in the context of cloud computing configuration.

The setup of this paper is as follows. Section 2 describes the mathematical formulation of the problem, Section 3 presents our proposal of a BO algorithm, while Section 4 collects some preliminary experimental results.

2 Background and mathematical formulation

The goal of BO is to approximate, e.g., the minimum of a given function f, called objective function, by using as few iterations as possible. Namely, we want to find $\widehat{x}$ where $\widehat{x} = \arg\min_{x \in \mathscr{A}} f(x)$. Strong assumptions on f or on the minimization domain $\mathscr{A}$ are not required, and BO algorithms are derivative-free, i.e., they do not require any knowledge about the derivatives of f. For these reasons, BO is often used to optimize expensive black-box objective functions (see [2]), that is, functions for which little to no information is available, and whose evaluation has significant time, resource, and/or monetary costs.

We consider the mathematical formulation for our *constrained global optimization* problem similarly to [1]. Let $x \in \mathscr{A}$ denote the d-dimensional vector representing a configuration for the cloud job, including information such as the number of cores used for the job, with $\mathscr{A} \subset \mathbb{R}^d$ being the domain of all feasible configurations. The *objective function* to be minimized is the total cost $f(x) = P(x)T(x)$, where $T(x)$ is the unknown execution time and $P(x)$ is the price per unit (it is a known, deterministic function). We also assume the constraint that $T(x) \leq T_{max}$, where T_{max} is a given threshold. Hence the problem is to find the minimum of f,

$$\min_{x \in \mathscr{A}} f(x) = P(x)T(x) \quad \text{s.t. } f(x) \leq P(x)\,T_{max}. \tag{1}$$

In this paper, we assume the deterministic price function $P(x)$ as being proportional to the number of virtual machines or cores used by the application job, which is always included in the cloud configuration vector x. Other choices of the price function are possible.

The key idea of BO comes from the Bayesian approach to statistics, in which values taken by f are treated as random variables, and a *prior distribution* represents the a-priori information on the modeled phenomenon – in the case of BO, information on the location of the minimum. The prior distribution is then iteratively updated with information coming from the observed data, obtaining the *posterior distribution*. For the rest of the paper, we assume that observed data, i.e., the evaluations of f, are noise-free. This is justified by the analysis in [5] on the data considered for validation. In a more general context, data can be assumed to have independent, normally distributed additive noise with variance η^2.

In this context, the Gaussian process (GP) is the preferred choice for the prior for f. This means that for any $x \in \mathscr{A}$,

$$f(x) \sim \pi_x(\cdot) = \mathscr{N}(\mu_0(x), \sigma_0^2(x,x)).$$

Functions $\mu_0(\cdot)$ and $\sigma_0^2(\cdot,\cdot)$ are called mean and kernel functions, respectively, and are the GP model hyperparameters. In this work, we assume $\mu_0(\cdot) \equiv \mu_0$ and we use the Matérn kernel with smoothness parameter $\nu = 5/2$ (see [3]):

$$\sigma_0^2(x,x') := \frac{1}{2^{3/2}\Gamma(5/2)} \left(\sqrt{5}\|x-x'\|\right)^{5/2} K_{5/2}\left(\sqrt{5}\|x-x'\|\right).$$

Having observed values $H_n = \{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$ of the objective function, one computes the posterior distribution of each $f(x)$, which is also Gaussian and it is characterized by the posterior mean $\mu_n(\cdot)$ and variance $\sigma_n^2(\cdot)$, i.e.

$$f(x)|H_n \sim \pi_x(\cdot|H_n) = \mathscr{N}(\mu_n(x), \sigma_n^2(x))$$

which can be computed by well-known properties of GPs (see, for instance, [3]).

BO is an iterative algorithm that obtains a new observation at each iteration by solving a proxy problem – the maximization of the *acquisition function* $g(x)$, which depends on the fitted GP model and measures the utility of evaluating the objective function at a given configuration x. This function is optimized at each round of the iterative algorithm, instead of directly optimizing the objective function itself, since it is available in closed form and inexpensive to evaluate. See Figure 1 for a summary of how BO works. Specifically, in the top panel, the objective function to be minimized (in gray) is approximated by the Gaussian Process, in the form of its posterior mean function (in dashed blue) and 95% credible interval (in light blue) after evaluating 3 points (red dots). The bottom panel shows the acquisition function given the current posterior distribution. The red cross indicates the maximum of the acquisition function, i.e. the next point which will be evaluated.

In this paper, we compare different acquisition functions. The *Expected Improvement* (EI) over the best value f_n^* found by the optimization process so far is:

$$EI_n(x) := \mathbb{E}_{\pi_x(\cdot|H_n)}[\max(f_n^* - f(x), 0)] \quad \text{with} \quad f_n^* = \min_{i \le n} f(x_i).$$

The expectation is taken under the current posterior distribution $\pi(\cdot\,|H_n)$ of $f(x)$, given history H_n. We consider a generalization of EI to the constrained optimization setting – the *Expected Improvement with Constraints* (EIC) acquisition function (see [8]), which accounts for the probability of a point of respecting the constraints:

$$EIC_n(x) := EI_n(x) \cdot \mathbb{P}_{\pi_x(\cdot|H_n)}\big(f(x) \le P(x)\, T_{max}\big). \tag{2}$$

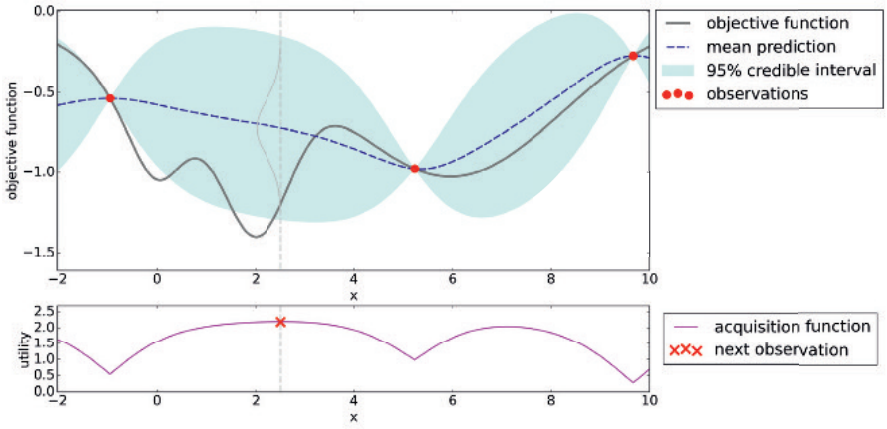


Fig. 1: Bayesian optimization after 3 iterations. Top panel: the objective function and its Bayesian approximation. Bottom panel: acquisition function.

3 Our hybrid algorithm

Our algorithm is based on pure BO, but it integrates elements coming from ML techniques. We use a memory queue for discrete features to prevent exploration of already visited values. The algorithm continues until the evaluated running time at the current iteration is sufficiently close to the time threshold: $T(x_n) \in [0.9\,T_{max},\ T_{max}]$. Our goal is to obtain a configuration that is compliant with the time threshold, but also uses as few resources as possible. Generally speaking, using more resources results in a lower execution time – meaning that a time which is just under the threshold consumes the least amount of resources for that configuration to be feasible. After termination, it is likely that we have found the true optimal configuration because of the convergence properties of the BO algorithm. Afterwards, we execute subsequent runs using such optimal or near-optimal configuration.

As far as the acquisition function is concerned, in this paper we use two variations of EIC (see (2)), which both integrate EIC with information coming from ML.

In particular, at each iteration, a ML model $\widehat{T}(\cdot)$ is trained on the current history H_n to accurately predict execution times $T(x)$ for any configuration x. Then, the EIC acquisition function in (2) is either multiplied by an exponential factor of $\widehat{T}(x)$ (variant B), or is set to zero for values in which $\widehat{T}(x)$ is larger than the threshold T_{max} (variant C). The former encourages the search process towards configurations which respect the time constraint, while the latter prevents the search outright in areas which are predicted to violate such constraint.

4 Experiments

We present preliminary results using the techniques we have discussed in Section 3. We use variants B and C of the algorithm, as well as pure BO, on the Query26 application from the TPC-DS industry benchmark run on Apache Spark with input data size equal to 250 GB and time threshold equal to 150 s. In this case, we optimize the total cost in (1) on the number of cores x. The same three fixed initial points were used for all variants. Figure 2 shows the comparison of pure BO (top row) with variants B (middle row) and C (bottom row) at each algorithm iteration. The left panels display the number of cores x selected by the algorithm, while the middle panels show the cumulative costs of the selected configurations. The percentage errors between the actual execution time and the one predicted by the ML model are displayed on the right column. We apply Ridge regression since [5] shows that it is the most accurate in this context, with a mean absolute error smaller than 5%.

Using our new algorithm, we are able to reduce the number of iterations which produce unfeasible configurations from 15 to 1 or 2. Similarly, cumulative costs associated to unfeasible runs are reduced from 11.67 to 0.80 and 1.51. At the very first iteration, all variants explore high values of the number of cores since there is no enough information on that part of the domain. This also explains the large prediction errors at the same iteration. After that, our ML model is able to predict the execution time of subsequent iterations with very good accuracy. Finally, the termination criterion (see the vertical dotted line in Figure 2) correctly assesses the optimality of the configuration with $x = 22$ cores, and stops the exploration phase.

Acknowledgments

The European Commission has partially funded this work under the Horizon 2020 Grant Agreement number 956137 LIGATE: LIgand Generator and portable drug discovery platform AT Exascale, as part of the EuroHPC Joint Undertaking.

References

1. O. Alipourfard, H. Liu, J. Chen, S. Venkataraman, M. Yu, and M. Zhang. Cherrypick: Adaptively unearthing the best cloud configurations for big data analytics. In *NSDI USENIX*, 2017.
2. E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

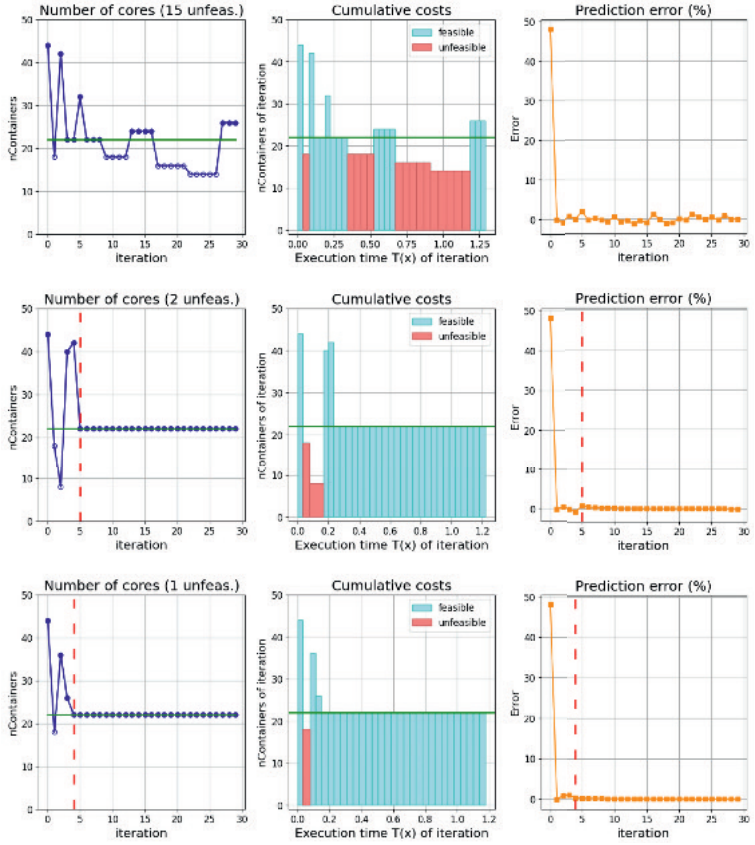


Fig. 2: Comparison of pure BO (top row) with variants B (middle row) and C (bottom row) at each algorithm iteration. Left panel: chosen numbers of cores x; middle panel: cumulative costs of the chosen configuration; right panel: percentage error. The true optimal number of cores is denoted by the green horizontal line.

3. P. I. Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.
4. B. Letham, B. Karrer, G. Ottoni, and E. Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14:495–519, 2019.
5. A. Maros, F. Murai, A. P. C. da Silva, J. M. Almeida, M. Lattuada, E. Gianniti, M. Hosseini, and D. Ardagna. Machine learning for performance prediction of spark cloud applications. In *IEEE 12th International Conference on Cloud Computing (CLOUD)*, pages 99–106, 2019.
6. T. Pourmohamad and H. K. Lee. Bayesian optimization via barrier functions. *Journal of Computational and Graphical Statistics*, (just-accepted):1–23, 2021.
7. B. Reagen, J. M. Hernández-Lobato, R. Adolf, M. Gelbart, P. Whatmough, G.-Y. Wei, and D. Brooks. A case for efficient accelerator design space exploration via Bayesian optimization. In *2017 IEEE/ACM ISLPED*, pages 1–6. IEEE, 2017.
8. M. Schonlau, W. J. Welch, and D. R. Jones. Global versus local search in constrained optimization of computer models. *IMS Lecture Notes-Monograph Series*, pages 11–25, 1998.

Confidence distributions and fusion inference for intractable likelihoods

Distribuzioni di confidenza e inferenza combinata per verosimiglianze intrattabili

Elena Bortolato and Laura Ventura

Abstract In the context of inference in complex models, involving intractable likelihood functions, we show how Confidence Distributions for a parameter of interest can be constructed with an efficient simulation strategy. The procedure is quite similar to those used in Approximate Bayesian Computation, but without requiring an arbitrary tuning tolerance. The proposed methodology allows us to derive proper likelihoods and enables to perform fusion inference, when the sources of information are non homogeneous or partial. An illustrationin the context of meta-analysis is discussed.

Abstract Si considerano problemi di inferenza in modelli complessi, in cui la funzione di verosimiglianza non è trattabile, e si mostra come poter costruire una distribuzione di confidenza per un parametro di interesse, con una strategia di simulazione simile a quella adottata negli approcci *Approximate Bayesian Computation*, senza richiedere una soglia arbitraria per il livello di approssimazione. La metodologia proposta rende possibile il recupero di una funzione di verosimiglianza ed è adattabile ai contesti inferenziali in cui le fonti di informazione non sono omegenee. Il metodo è illustrato in un esempio di metanalisi.

Key words: Approximate Bayesian Computation, Confidence distributions, Fusion inference, Implied likelihood

Elena Bortolato
Department of Statistical Sciences, University of Padua, e-mail: elena.bortolato.1@phd.unipd.it

Laura Ventura
Department of Statistical Sciences, University of Padua, e-mail: ventura@stat.unipd.it

1 Introduction

In many fields of science, it is needed to fuse together information from potentially different types of sources or complex models. In these settings, it has been becoming common practice to consider simulation-based inference (Sisson et al. 2018). Indeed, in the Bayesian framework, Approximate Bayesian Computation (ABC) and Bayesian Synthetic Likelihood (BSL) methods escape the problem of impossible likelihood evaluation. Then, several methods are well established for combining uncertainty from multiple information or models: by updating the posterior distribution when data arrive sequentially, by using robust Bayesian methods, or model averaging (Berger 1990, Wasserman 2000). On the contrary, in the frequentist approach, performing inference with complicated models and models made of modules may be more problematic. For example, Lindley (1958) shows that paradoxes easily arise when combining p-values. Instead of simply considering p-values for the purpose, one could appeal to confidence distributions (CD); see Hannig and Xie (2012). We observe that these can be derived even in complex settings, where an underlying pivotal distribution is not available in closed form.

The outline of this work is the following. In Section 2 and 3 we summarize the essentials of CD and likelihood free inference. In Section 4 we present the proposed construction of the CD and in Section 5 we describe how to suit the methodology within a fusion inference framework. Some examples are discussed in Section 6.

2 Towards confidence distributions and fiducial inference

Confidence distributions are sample-based distribution functions on the parameter space that convey and resume the frequentist inferential results, similarly to the counterpart Bayesian posterior. In order to formalize the mathematical setting, let y_n be the $n-$size sample, realization of a random variable $Y \in \mathscr{Y}$, and let $\theta \in \Theta \subseteq \mathrm{R}$ be the scalar parameter of interest.

Definition 1. *Confidence distribution*
A function $H_n(\cdot) = H_n(y_n, \cdot)$ on $\mathscr{Y} \times \Theta \to [0,1]$ is called a confidence distribution for a parameter θ, if

C1 For each given $y \in \mathscr{Y}$, $H_n(\cdot)$ is a cumulative distribution function on Θ;
C2 At the true parameter value $\theta = \theta_0$, $H_n(\theta_0) = H_n(y_n, \theta_0)$, as a function of the sample y_n, follows the uniform distribution $U(0,1)$.

Definition 2. *Confidence density and CD random variable*
A random variable ξ such that

$$\xi | Y = y_n \sim H_n(\cdot)$$

is called a CD random variable and its probability distribution, called confidence density, is given by $\frac{\partial H_n(\cdot)}{\partial \xi}$.

The CD random variable represents the uncertainty in the estimation of the parameter of interest, or can be seen as a random estimator of it. For any couple of values θ_L and θ_U, the included mass-area under the CD measures the evidence associated to the interval $[\theta_L, \theta_U]$. In practice, exact or asymptotic CDs can be derived from the cumulative distribution of a pivotal quantity, or can be generated using bootstrap methods (see Ruli and Ventura 2021, and references therein). CDs are also related to fiducial inference and structural models (Fraser 2004), which aim at constructing a measure on the parameter space without a prior.

3 Simulation based inference for intractable likelihoods

For some complex models, for which the likelihood cannot be evaluated in a practical amount of time, both in the frequentist and the Bayesian framework, some methods have been proposed to overcome the numerical difficulties, these are mainly based on simulations.

In the frequentist approach, one possibility for the inspection of the likelihood is based on the use of an auxiliary model $M^*(\psi)$ (Gourieroux et al. 1993), where ψ is a parameter. The model is fitted to the data in order to obtain a point estimator, $\hat{\psi}$, then a binding function $\psi(\theta)$ among the original and the auxiliary model is considered and the map is estimated by simulation. In some cases, other solutions for the problem are given by composite likelihoods, see for instance Varin et al. (2011) and reference therein.

Approximate Bayesian Computation (ABC), instead, aims at finding the posterior distribution for the parameter of interest, simulating from the joint distribution of the parameter and the data, $p(y,\theta) = \pi(\theta)p(y|\theta)$, retaining those simulates values θ^*, for which the simulated samples y^* match the observed sample y_n. The statistic $t(\cdot)$ used to compare the actual sample and the simulated ones, is typically low dimensional and possibly sufficient. Also, the matching criterion is relaxed, according to the definition of a distance $d(\cdot,\cdot)$ and a small tolerance ε: thus the values θ^* are accepted if $d(t(y^*),t(y_n)) < \varepsilon$. The choice of a metric together with the summary statistic determine the precision of the results. ABC has also been used in the frequentist approach to obtain approximations of the likelihood function, eventhough the amount of inexactness is notoriously difficult to quantify, and the procedure is non-consistent (Frazier et al. 2015).

4 CD for intractable models

In the preface of their book, "Confidence, probability and likelihood" (2016), Schweder and Hjort state "The price to be paid for an epistemic distribution not based on a prior is that in most models only approximate confidence distributions are available, and they might be more computationally demanding than the Bayesian posterior". While some recent works improved the approximation of CDs,

a lot has to be done on the side of using them for complex models and in likelihood free framework. We note that it is possible to build an exact confidence distribution together with a confidence density from an accept-reject scheme, a summary statistic and the auxilium of a proposal distribution as done in ABC. In Algorithms 1 and 2 the simulation strategy is described. If the distribution of the chosen statistic $t(\cdot)$ is stochastically ordered in the parameter of interest (i.e. $Pr(t(y) < s|\theta_1) \leq Pr(t(y) < s|\theta_2)\ \forall s,\ \theta_1 > \theta_2$, and strictly for at least one s), C1 in Definition 1 is satisfied. Instead, C2 holds since the defined CD has meaning of a p-value function. Note that it is not required that $t(\cdot)$ is a pivotal quantity (i.e. its distribution doesn't depend on the parameter of interest), neither that $t(\cdot)$ is sufficient or its distribution known. The assumption can be checked *a posteriori*, by verifying that the CD is monotone non decreasing.

Algorithm 1 Confidence distribution

Choose a proposal $q(\theta)$, M the number of iterations, a summary statistic $t(\cdot)$, compute $t_n = t(y_n)$
1- sample $\theta^* \sim q(\theta)$ and $y^* \sim p(y|\theta^*)$
2- compute $t^* = t(y^*)$
3- accept θ^* if $t^* \leq t_n$ else reject
4- repeat (1-2-3) M times
5- sample θ^{**} from the accepted θ^* with replacement and probability $1/q(\theta^*)$
Output: $\theta^{**} \propto H_n$ a confidence distribution

Algorithm 2 Confidence density

Fix a desired size R
1- define a grid of values $G = (1/R, 2/R, \ldots, 1)$
2- compute the quantiles of the CD : $\xi \sim H_n^{-1}(G)$
Output: $\xi \sim c(\theta)$ is a CD random variable

5 From confidence to likelihood

It is possible to retrieve a likelihood function by considering the implied likelihood (Efron, 1993). This involves the construction of a fictitious second dataset, doubling the original $y^{II} = (y_n, y_n)$, whose likelihood function is $\mathscr{L}(\theta; y^{II}) \propto \mathscr{L}(\theta; y_n)^2$. The implied likelihood is retrieved by

$$\mathscr{L}(\theta) = \frac{c(\theta; y^{II})}{c(\theta; y_n)}, \tag{1}$$

where $c(\theta; y)$ is the confidence density. In practice, it is possible to perform this step resampling the CD random variables obtained from the analysis of the doubled dataset with importance weights given by the distribution of the CD random variables obtained on the original data.

6 Examples

6.1 Inference in the lognormal model

Consider the lognormal model. Assume to observe as a summary statistic $t_n = 1.439$, the mean of $n = 5$ realizations:

$$t_n = n^{-1} \sum_{i=1}^{n} e^{z_i}, \quad z_i \sim Z \sim N(\mu, 1), \quad i = 1, \ldots, n.$$

For $n > 2$ an analytical form for the distribution of t_n is not available; thus, furnishing a confidence interval for μ based on t_n, despite the formal simplicity of the model, is a very sophisticated problem. We apply our method proposing 10^5 values from the Uniform in (-4,4), when the true parameter is $\mu = 0$. The difference between the acceptance probability (about 49%) is remarkable compared to the ABC's (0.1%). Then, we run a simulation study in which we compute the 95% confidence intervals obtained with CDs and with the approximate likelihood via ABC with a small tolerance ($\varepsilon = 0.01$). The resulting empirical coverage based on 10^4 simulations and 10^4 proposal values for each of them is 94.9% (se= 0.2%), against ABC's 77.1% (se= 0.3%). We show in Figure 1 the CD and the coverage of the intervals along the iterations.

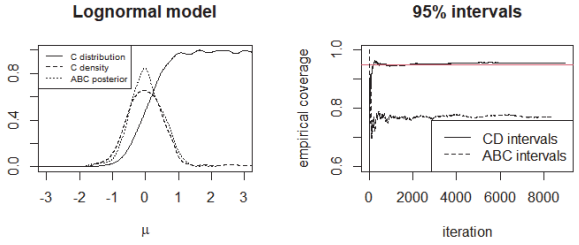


Fig. 1 *Left*: Confidence distribution, confidence density and ABC posterior given $t_n = 1.439$. *Right*: empirical coverage of intervals from a simulation study comparing CD to ABC credible intervals.

6.2 Implied likelihood via CD and Meta-analysis

Suppose three independent laboratories have measured the same quantity of interest, from a sample of size $n = 18$, and we are interested in combining the results to make inference for a parameter μ representing the mean of the phaenomenon. The first lab has analysed the data with a normal model and reported the mean ($S_1 = 31.04$). The second lab preferred to give the median ($S_2 = 30.69$), having observed some presumed outliers, while the third lab considered the logarithm of the measures, trying to correct for some observed asymmetric behaviour, and furnishes the mean of the logarithms ($S_3 = 3.43$). We build the CDs, obtain the implied likelihoods (IL) and combine them, taking the product. We show the results in Figure 2. The comparison to the likelihood, available for the first lab's analysis, indicates good adaptation of

the procedure. The scheme illustrated is quite general, can be used in presence of in-

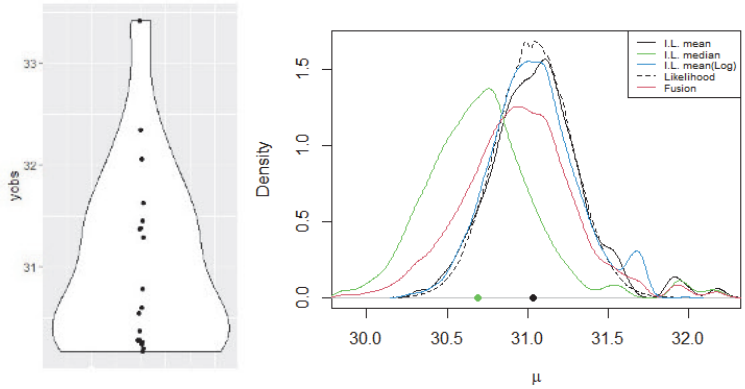


Fig. 2 *Left*: sample analysed by the laboratories. *Right*: combined likelihood (red line), obtained by fusion of three IL, related to different CDs. Black and green dots represent the statistics given by the first and the second lab (S_1, S_2), the third is not in the support of the x-axis $(S_3 = 3.43)$.

tractable or multiple competitor models, trying to account for misspecification. The combined IL, in contrast to p-value recombination, avoids paradoxes and naturally conveys a distribution estimator, without further assumptions.

References

1. Berger, J. O.: Robust Bayesian analysis: sensitivity to the prior. J. Stat. Plan. Inference, **25(3)**, 303-328 (1990)
2. Efron, B.: Bayes and likelihood calculations from confidence intervals. Biometrika, **80(1)**, 3-26 (1993)
3. Fraser, D.A.S.: Structural inference. Encyclopedia of Statistical Sciences, **13** (2004)
4. Frazier, D. T., Martin, G. M., Robert, P. C.: On Consistency of Approximate Bayesian Computation. arXiv:1508.05178 (2015)
5. Gourieroux, C., Monfort, A., Renault, E.: Indirect inference. J. Appl. Econ., **8(S1)**, S85-S118 (1993)
6. Hannig, J., Xie, M.: A note on Dempster-Shafer recombination of confidence distributions. Electron. J. Stat., **6**, 1943-1966 (2012)
7. Lindley, D.V.: Fiducial Distributions and Bayes' Theorem, J. R. Stat. Soc. Series B Stat. Methodol., 102-107 (1958)
8. Ruli, E., Ventura, L.: Can Bayesian, confidence distribution and frequentist inference agree?. Stat. Methods. Appt., **30(1)**, 359-373 (2021)
9. Schweder, T., Hjort, N. L.: Confidence, likelihood, probability. Cambridge University Press, Cambridge (2016)
10. Sisson, S. A., Fan, Y., Beaumont, M. (ed.): Handbook of approximate Bayesian computation. Chapman and Hall/CRC, New York (2018)
11. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Stat. Sin., 5-42 (2011)
12. Wasserman, L.: Bayesian model selection and model averaging. J. Math. Psychol., **44(1)**, 92-107 (2000).

Wasserstein distance and applications to Bayesian nonparametrics

Distanza di Wasserstein e applicazioni in statistica bayesiana nonparametrica

Marta Catalano, Hugo Lavenant, Antonio Lijoi, Igor Prünster

Abstract Bayesian nonparametric models are able to learn complex distributional patterns in the data by leveraging on infinite-dimensional parameters, typically consisting in vectors of random measures. To perform a principled BNP model comparison one thus needs a measure of discrepancy between vectors of random measures. In recent works the authors have proposed two different metrics based on the Wasserstein distance. We here provide new perspectives to our findings, highlighting a universal relation between the two metrics.

Abstract *I modelli bayesiani nonparametrici (BNP) sono in grado in imparare assetti distribuzionali complessi nei dati facendo leva su parametri infinito-dimensionali, che tipicamente consistono in vettori di misure aleatorie. Per portare avanti un paragone tra modelli BNP basato su fondamenta solide, è dunque necessario definire una misura di discrepanza tra vettori di misure aleatorie. In alcuni lavori recenti gli autori hanno proposto due diverse metriche basate sulla distanza di Wasserstein. In questo contributo forniamo una nuove prospettiva che mette in evidenza una relazione universale tra le due metriche.*

Key words: Bayesian model comparison, Lévy measure, Wasserstein distance.

Marta Catalano
University of Warwick, UK, e-mail: marta.catalano@warwick.ac.uk

Hugo Lavenant
Bocconi University, Italy, e-mail: hugo.lavenant@bocconi.it

Antonio Lijoi
Bocconi University, Italy, e-mail: antonio.lijoi@bocconi.it

Igor Prünster
Bocconi University, Italy, e-mail: igor.pruenster@bocconi.it

1 Introduction

The Wasserstein distance provides a comprehensive way to quantify the comparison between two random objects. Its first definition traces back to [8], though it then appeared independently in many other scientific fields. As a result, one can find this *simple measure of discrepancy* (using Gini's words) under different names, such as Gini distance, coupling distance, Monge-Kantorovich distance, Earth Moving distance and Mallows distance; see [5, 12, 11] for reviews.
Random structures are the key to statistical modeling and inference, especially in a Bayesian framework where the parameter of the model is random as well. We argue that a principled way to perform Bayesian model comparison is then to measure the discrepancy between the random parameters of the models, by relying on the Wasserstein distance.
In this contribution we focus on Bayesian nonparametric (BNP) models, which are based on infinite-dimensional parameters typically consisting on vectors of random measures. For example, most BNP models for *partially exchangeable sequences* [4] are built as follows. First, one considers a vector of dependent random measures $\tilde{\mu} = (\tilde{\mu}_1, \dots, \tilde{\mu}_d)$, such as the *completely random vectors* (CRVs) introduced in Section 3. Then, one models d groups of observations $\boldsymbol{X}^i = (X_1^i, \dots, X_{n_i}^i)$ as conditionally independent given $\tilde{\mu}$ with

$$\boldsymbol{X}^i | \tilde{\mu} \overset{\text{ind}}{\sim} T(\tilde{\mu}_i)$$

for $i = 1, \dots d$, where T is some transformation of the random measure.
To perform a BNP model comparison one thus needs the definition of a distance between CRVs, whose law is specified through a multivariate Lévy intensity. Following [3, 2, 1] we here describe two different ways of defining such a distance, one at the level of the random measures and one at the level of the underlying Lévy measures. Generalizing the findings in [1], we highlight that the two metrics are intimately connected and provide a universal relation between them that allows the practionner to exploit the best properties of each one.

2 Wasserstein distance

In its classical formulation, the Wasserstein distance provides the means for comparing two probability measures on a Polish space $\mathbb{X}$ with metric d. Recent developments allow to extend the comparison from probability measures (that is, measures with mass 1) to generic measures with possibly different masses. A particularly interesting scenario arises when the masses are infinite, as in the case of infinitely active Lévy measures. In this section we describe both the Wasserstein distance and its extended version.
We first recall the notion of coupling. Given two probability measures P^1, P^2 on a Polish space $(\mathbb{X}, d)$ we define a coupling to be the law of any random vector (X, Y)

on the product space $\mathbb{X}\times\mathbb{X}$ such that $X\sim P^1$ and $Y\sim P^2$. We denote by $\Gamma(P^1,P^2)$ the set of couplings between P^1 and P^2. A probability P^1 has finite p-th moment if for any $x\in\mathbb{X}$,

$$\int_{\mathbb{X}} d(s,x)^p \mathrm{d}P(s) < +\infty.$$

Definition 1. Let P^1 and P^2 be probability measures on $(\mathbb{X},d)$ with finite p-th moment. The *Wasserstein distance* of order p between P^1 and P^2 is

$$\mathscr{W}_p(P^1,P^2)^p = \inf_{(X,Y)\in\Gamma(P^1,P^2)} \mathbb{E}(d(X,Y)^p).$$

We observe that the finiteness of the p-th moment guarantees the finiteness of the Wasserstein distance thanks to the triangular inequality of the metric d. A very common choice for $(\mathbb{X},d)$ is $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$. In this case, P has finite p-th moment if and only if $\int \|x\|^p \mathrm{d}P(x)$ is finite.
Starting from this scenario we extend the definition to Lévy measures on $\Omega_d = [0,+\infty)^d \setminus \{\mathbf{0}\}$. This distance was studied in [7, 9, 1] for $p=2$. To this end, we insist on an equivalent definition of coupling in terms of pushforward measures, which will be easier to generalize to Lévy measures. For a point $(\boldsymbol{s},\boldsymbol{s}')\in\mathbb{X}\times\mathbb{X}$, we denote by $\pi_1(\boldsymbol{s},\boldsymbol{s}')=\boldsymbol{s}\in\mathbb{X}$ and $\pi_2(\boldsymbol{s},\boldsymbol{s}')=\boldsymbol{s}'\in\mathbb{X}$ its projections. For any measure spaces $\mathbb{X}^1$ and $\mathbb{X}^2$, if μ is a measure on $\mathbb{X}^1$ and $f:\mathbb{X}^1\to\mathbb{X}^2$, the pushforward measure of μ by f is the measure on $\mathbb{X}^2$ defined by $(f_{\#}\mu)(A)=\mu(f^{-1}(A))$. Then an equivalent way of defining a coupling $\gamma\in C(P^1,P^2)$ is as a probability measure on $\mathbb{X}\times\mathbb{X}$ such that $\pi_{i\#}\gamma = P^i$ for $i=1,2$. Thus,

$$\mathscr{W}_p(P^1,P^2)^p = \inf_{\gamma\in\Gamma(P^1,P^2)} \int_{\mathbb{X}\times\mathbb{X}} d(x,y)^p \mathrm{d}\gamma(x,y).$$

We now extend the Wasserstein distance to Lévy measures on $\Omega_d=[0,+\infty)^d\setminus\{\mathbf{0}\}$ with p-th finite moment, that is, the set of positive Borel measures ν such that

$$\int_{\Omega_d} \|s\|^p \mathrm{d}\nu(s) < +\infty.$$

We observe that the finite moment condition does not prevent the mass from being infinite. Indeed, there can be an accumulation of infinite mass around the origin, as is the case with infinitely active Lévy measures.
Let ν^1,ν^2 two Lévy measures. An extended coupling γ between ν^1 and ν^2 is a Lévy measure on Ω_{2d} such that $\pi_{1\#}\gamma|_{\Omega_d}=\nu^1$ and $\pi_{2\#}\gamma|_{\Omega_d}=\nu^2$. We denote $\Gamma_*(\nu^1,\nu^2)$ the set of all extended couplings.

Definition 2. Let ν^1 and ν^2 be Lévy measures on $(\Omega_d,\|\cdot\|)$ with finite p-th moment. The *extended Wasserstein distance* of order p between ν^1 and ν^2 is

$$\mathscr{W}_{*,p}(\nu^1,\nu^2)^p = \inf_{\gamma\in\Gamma_*(\nu^1,\nu^2)} \iint_{\Omega_{2d}} \|\boldsymbol{s}-\boldsymbol{s}'\|^p \,\mathrm{d}\gamma(\boldsymbol{s},\boldsymbol{s}'). \tag{1}$$

The main difference between a coupling and an extended coupling is that extended couplings are not defined on $\Omega_d \times \Omega_d$ but rather on Ω_{2d}, which is strictly larger since it also contains the axis $\{\mathbf{0}\} \times \Omega_d$ and $\Omega_d \times \{\mathbf{0}\}$. This is fundamental to ensure the compactness of extended couplings which, in turn, guarantees the existence of an *optimal extended coupling*, that is, a coupling that attains the minimum in the definition of extended Wasserstein distance.

3 Completely random vectors

The key component of Bayesian nonparametric models is the use of random measures to learn complex distributional patterns in the data. Among these, completely random measures stand out for combining inferential flexibility with analytical tractability. Quantifying the discrepancy between two completely random measures or their multivariate extension - completely random vectors - is key to establishing a principled approach to the comparisons between the induced Bayesian nonparametric models. This can be done through the Wasserstein distance at two different levels: either at the level of the random measures or at the level of their underlying Lévy intensity. We here describe these notion and extend a result by [1], which shows that these two levels are intimately connected.
Let $\mathscr{M}(\mathbb{X})$ denote the set of boundedly finite measures on $\mathbb{X}$ endowed with the weak$^\sharp$ topology [6]. We recall that a sequence of measures μ_n converges weakly$^\sharp$ to μ if and only if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for every bounded continuous f vanishing outside a bounded set. A random measure is a random element on $\mathscr{M}(\mathbb{X})$. Similarly, a d-dimensional random vector of measures $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a random element on the product space $\mathscr{M}(\mathbb{X})^d$.

Definition 3. A random vector of measures is a *completely random vector* (CRV) if for every finite collection of pairwise disjoint bounded sets $\{A_1, \cdots, A_n\}$ in $\mathscr{B}(\mathbb{X})$, $\{\tilde{\mu}(A_1), \cdots, \tilde{\mu}(A_n)\}$ are mutually independent.

The definition of completely random vector was first given by [2] and it extends the notion of completely random measure (CRM [10]) in a natural way. In particular, a CRM is a 1-dimensional CRV. Every CRV can be decomposed as the sum of three independent components, $\mu + \tilde{\mu}_f + \tilde{\mu}$ in distribution, where μ is a deterministic vector of measures, $\tilde{\mu}_f$ is a random vector of measures with fixed atoms and $\tilde{\mu}$ is a random vector of measures without fixed atoms. In the rest of the contribution we will focus on CRVs without fixed atoms. Remarkably, for every such CRV there exists a measure ρ on $\Omega_d \times \mathbb{X}$ such that

$$\mathrm{d}\tilde{\mu}(x) = \int_{\Omega_d} s \, \mathrm{d}\mathscr{N}(s,x)$$

in distribution, where $\mathscr{N}$ is a Poisson random measure with Lévy intensity ρ on $\Omega_d \times \mathbb{X}$. In particular, $\tilde{\mu}(A)$ is an infinite divisible distribution with Lévy measure

$d\nu_A(s) = \int_A d\rho(s,x)$ on Ω_d. Thus, to assign a probability distribution to a CRV it suffices to provide a set of Lévy measures $(\nu_A)_A$.
It is then clear that the comparison between CRVs can be based either at the level of the random measures or at the level of the Lévy measures. In the first case, given two CRVs $\tilde{\mu}^1$ and $\tilde{\mu}^2$ one defines

$$d_{\mathscr{W},p}(\tilde{\mu}^1,\tilde{\mu}^2) = \sup_A \mathscr{W}_p(\mathscr{L}(\tilde{\mu}^1(A)),\mathscr{L}(\tilde{\mu}^2(A))), \tag{2}$$

where $\mathscr{L}$ denotes the law of a random object. We observe that $\mathscr{L}(\tilde{\mu}^i(A))$ is a probability measure and thus $\mathscr{W}_p$ is the classical Wasserstein distance in Definition 1. In the second case one defines

$$d_{\mathscr{W}_*,p}(\tilde{\mu}^1,\tilde{\mu}^2) = \sup_A \mathscr{W}_{*,p}(\nu_A^1,\nu_A^2). \tag{3}$$

Here, ν_A^i is a Lévy measure on Ω_d and thus $\mathscr{W}_{*,p}$ is the extended Wasserstein distance in Definition 2.
The advantage for (2) is that, being defined directly on the the random object, it has a very interpretable definition. On the other hand the advantage of (3) is that, being defined directly on the Lévy measure, it is easier to compute. The next result generalizes the findings in [1] and can be proved following the same technique therein. We provide a universal relation between the two distances that allows to keep the best of the two worlds: we can now base the computations on (3) though maintaining the interpretability of (2).

Theorem 1. *Let $\tilde{\mu}^1,\tilde{\mu}^2$ be CRVs with Lévy measures $(\nu_A^1)_A$ and $(\nu_A^2)_A$, respectively, with finite p-th moment. Then for every Borel set A,*

$$\mathscr{W}_p(\mathscr{L}(\tilde{\mu}^1(A)),\mathscr{L}(\tilde{\mu}^2(A))) \leq \mathscr{W}_{*,p}(\nu_A^1,\nu_A^2).$$

References

1. Marta Catalano, Hugo Lavenant, Antonio Lijoi, and Igor Prünster. A Wasserstein index of dependence for random measures. *arXiv:2109.06646*, 2022+.
2. Marta Catalano, Antonio Lijoi, and Igor Prünster. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Statist.*, 49(5):2916–2947, 2021.
3. Marta Catalano, Antonio Lijoi, and Igor Prünster. Approximation of Bayesian models for time-to-event data. *Electron. J. Statist.*, 14(2):3366–3395, 2020.
4. Donato Michele Cifarelli and Eugenio Regazzini. Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III*, 12:1–36, 1978.
5. Donato Michele Cifarelli and Eugenio Regazzini. On the centennial anniversary of Gini's theory of statistical relations. *Metron*, 75(2):227–242, August 2017.
6. Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, 2002.

7. Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de Mathèmatiques Pures et Appliquèes*, 94(2):107–130, 2010.
8. Corrado Gini. Di una misura delle relazioni tra le graduatorie di due caratteri. *Saggi monografici del Comune di Roma, Tip. Cecchini*, 1914.
9. Nestor Guillen, Chenchen Mou, and Andrzej Święch. Coupling Lévy measures and comparison principles for viscosity solutions. *Transactions of the American Mathematical Society*, 372(10):7327–70, 2019.
10. John F. C. Kingman. *Pacific J. Math.*, (21):59–78, 1967.
11. Svetlozar Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.
12. Cedric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

Network Analysis and community detection

Community detection in networks: a heuristic version of Girvan Newman algorithm

Identificazione delle comunità nei grafi: una versione euristica dell'algoritmo di Girvan Newman

Ilaria Bombelli and Lorenzo Di Rocco

Abstract Complex problems in different fields can be modeled by graph data structures. An important and challenging task is the community detection, i.e. the identification of highly connected components. The Girvan Newman algorithm identifies high-quality communities but is not computationally suitable for huge graph analysis. Therefore, we propose a heuristic version of this algorithm, by considering an approximated measure of the edge betweenness. We evaluate the performances of our proposal on benchmark networks.

Abstract *I grafi permettono di modellare problemi complessi di diversa natura. Una delle questioni più stimolanti e importanti è il rilevamento delle comunità, cioè l'identificazione di componenti altamente connesse. L'algoritmo di Girvan Newman è in grado di identificare tali componenti con successo, ma non è computazionalmente adatto ad un'analisi di grafi di grandi dimensioni. Pertanto, proponiamo una versione euristica di questo algoritmo, approssimando la misura di betweenness degli archi. Analizziamo le prestazioni della nostra proposta utilizzando reti benchmark.*

Key words: Network, Communities Detection, Edge Betweenness

Ilaria Bombelli
University of Rome La Sapienza, Statistical Sciences Department, Rome, Italy
e-mail: ilaria.bombelli@uniroma1.it

Lorenzo Di Rocco
University of Rome La Sapienza, Statistical Sciences Department, Rome, Italy
e-mail: lorenzo.dirocco@uniroma1.it

1 Introduction

Networks are interesting tools to represent complex problems that arise in different situations: for example, in a society, networks can represent how class students interact with one another; in technological field, networks can represent flights routes between cities.

In a network framework, a really challenging task is the detection of *communities* (or clusters of nodes). Plenty of literature deals with this task and, among the hierarchical algorithms, the most famous ones are Louvain [2] and Girvan Newman [5]. The former follows an *agglomerative* approach, while the latter is characterized by a *divisive* one.

The Girvan Newman (GN) algorithm suffers from an important drawback: it is unfeasible for community detection problems on large-scale networks, as also the authors highlight [5]. Indeed, the GN algorithm is based on the computation of the edge betweenness, which is practically an extremely expensive task.

For this reason, we propose a Heuristic Girvan Newman (HGN) algorithm, considering an approximation of the edge betweenness. In this way, the computational effort of each iteration of the GN algorithm is reduced.

In Section 2, we explore deeply the methodology underlying our proposal. In Section 3, we assess the performances of our proposal on benchmark networks. Finally, in Section 4 final remarks and future research directions are discussed.

2 Methods

Complex networks of highly connected data are commonly represented by graphs. A graph $G = (V,E)$ is a data structure built on top of a set of vertices (V) and a set of edges (E). The vertices are considered to be the entities of the network, while the edges describe the occurring interactions.

In network science, hubs identification is an important task and aims to identify pivotal vertices. Different measures have been proposed in the literature to quantify the importance of a vertex. Freeman's *betweenness* [4] is a widely used index that evaluates the centrality of a vertex considering the number of shortest paths passing through the vertex.

The definition of betweenness can be extended to the edges [5]. The so-called *edge betweenness* for an edge $e \in E$ is defined as follows :

$$ebw(e) := \sum_{\substack{j,k=1,\dots,|V| \\ j \neq k}} \frac{\sigma_{jk}(e)}{\sigma_{jk}} \tag{1}$$

where σ_{jk} is the number of shortest paths connecting node v_j and node v_k, and $\sigma_{jk}(e)$ is the number of shortest paths connecting node v_j and node v_k that run along the edge $e \in E$.

The greatest values of this index are usually associated with *bridge-like* edges, i.e. those connecting different communities of the network. The GN algorithm is one of the main methods for communities detection and it is based on a recursive procedure that at each iteration deletes the edge(s) with the greatest betweenness.

This divisive procedure provides high-quality clustering performances but is unpractical for large-scale graph processing. This drawback is due to the shortest paths search, which is particularly expensive from a computational viewpoint. A possible solution consists in parallelizing the elaborations on a High Performance Computing architecture, obtaining in this way a significant performance speed-up. However, it is worth considering also high-quality approximations of the edges betweenness to reduce the computational effort.

In this work, we propose a heuristic version of the GN algorithm that estimates the edges betweenness through an accurate sampling technique. More precisely, we take into account the approximation proposed by [8], which we call DIAM, as suggested by [1]. This approach evaluates an appropriate sample size r according to the graph structure, it randomly selects a sample of paths $S = \{p_i, i = 1, \ldots, r\}$ and it returns for each edge $e \in E$ the following estimator:

$$\widehat{ebw} = \frac{1}{r} \sum_{p \in S} \mathbb{I}_p(e) \tag{2}$$

where $\mathbb{I}_p(e)$ is 1 if e belongs to p, 0 otherwise. In the following, we provide more details about the procedure.

Firstly, the estimates of the edge betweenness $\{\widehat{ebw}_i, i = 1, \ldots, |E|\}$ are all set to 0. Then, r pairs of vertices are sampled uniformly from V. For each sampled pair (u, v): (a) all the shortest paths $S_{u,v}$ between u and v are computed, (b) a path p is selected uniformly at random from $S_{u,v}$, (c) the betweenness estimates of the edges belonging to p are increased by $1/r$.

The sample size formula is obtained considering an application of the Vapnik-Chervonenkis dimension [9] and it is defined as follows:

$$r := \frac{c}{\varepsilon^2} \left(\lfloor \log_2(\tilde{D}(G) - 1) \rfloor + 1 + \ln \frac{1}{\delta} \right) \tag{3}$$

where c is a positive constant (typically set to 0.5, as [7] suggested), $\varepsilon \in [0, 1]$ is the additive error, $1 - \delta$ (with $\delta \in [0, 1]$) is the probability of the error and $\tilde{D}(G)$ is the 2-approximation of the diameter of the graph $D(G)$ [8].

Summing up, the HGN algorithm is based on the following steps:

1. Calculate the approximated betweenness for all edges using DIAM.
2. Remove the edge(s) with the highest betweenness.
3. Recalculate the approximated betweenness for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

It is worth mentioning that in the worst-case scenario at each iteration GN computes all the shortest paths between $n(n-1)$ pairs of vertices, while HGN only between r

pairs. Moreover, our implementation allows to stop the algorithm when the required number of communities K is reached.

3 Experimental Analysis

We evaluate the performances of our proposal on benchmark networks in terms of *modularity* [3] and *Adjusted Rand Index* (ARI) [6]. We briefly describe the networks and then we show the main results.

3.1 Dataset

The benchmark networks that we consider are both characterized by an underlying known community structure. *Zachary karate club network* ([10]) represents social ties among the members of a university karate club collected by Wayne Zachary in 1977. *American college football network* ([5]) represents American football games between Division IA colleges during the regular season of Fall 2000. The former contains 78 edges, 34 vertices, and $K = 2$ known communities; the latter contains 616 edges, 115 vertices, and $K = 12$ known communities.

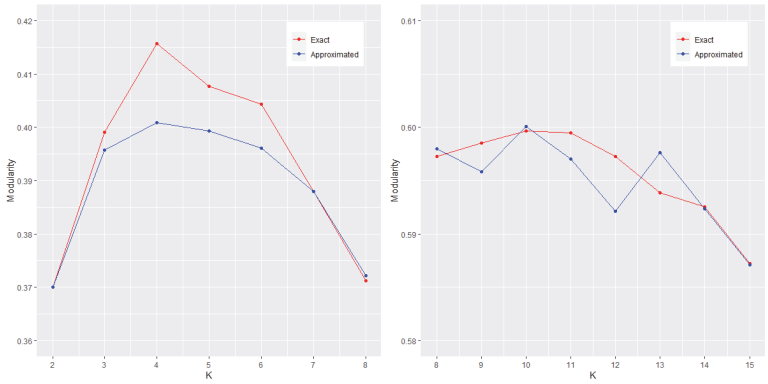
3.2 Results

As a preliminary analysis, we assess the accuracy of DIAM, since, to the best of our knowledge, there are no benchmarks in the literature. We notice that DIAM recognizes the highest-scoring edges, but, by sorting the edges in a decreasing order w.r.t the betweenness, mismatches between approximated ranking and exact one can occur. This implies that the edge deleted by HGN at i-th iteration does not always correspond to the one deleted at the same iteration by the GN algorithm. Anyway, HGN provides good performances in solving community detection problems.

Figure 1 (a) shows that the modularity trends of HGN and GN, when applied to Zachary karate club dataset, are similar: therefore, $K = 4$ is the number of communities that maximizes the modularity in both cases. Moreover, by fixing $K = 2$ as the true partition suggests, both methods induce the same partition of the network, misclassifying only two units. Indeed, when comparing the two partitions against the true one, we got the same value of the ARI.

When applied to Football network, HGN leads to an oscillatory behaviour of the modularity (see Figure 1 (b)). For $K \in \{8, 10, 14, 15\}$ the values of the modularities corresponding to GN and HGN are almost the same, while greater differences are observed for $K \in \{9, 11, 12, 13\}$. However, the greatest difference, observed for $K = 12$, is equal to 0.00513. Therefore we can conclude that all the values are close to

the exact ones. Fixing $K = 12$, HGN leads to an ARI equal to 0.883, while the ARI of GN is equal to 0.885. Hence, HGN provides good results in terms of the ARI index. Indeed, comparing Figure 2 (a) and (b), we observe that the number of misclassifications in both the partitions is almost the same and the communities are made up of almost the same units.



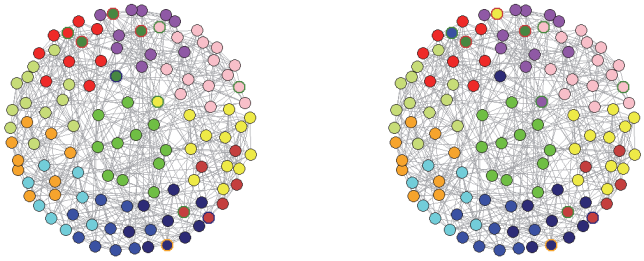
(a) Karate (b) Football

Fig. 1: Modularity values, as a function of the number of communities K, obtained by applying the exact (GN) and the approximated (HGN) algorithm with $\varepsilon = 0.05$ and $\delta = 0.1$.

4 Conclusion

In this work, we proposed a heuristic reformulation of the GN algorithm to face the problem of community detection. Since nowadays many complex problems can be represented by large-scale graphs, this approach can be a useful tool to avoid computationally heavy procedures.

Considering the good results obtained on small benchmark networks, future direction may involve the application of our method on a larger graph to accurately assess the computation time gain. Moreover, we are considering also other approaches to improve the edge betweenness approximation.



(a) GN (b) HGN

Fig. 2: Partitions obtained by applying GN and HGN on Football network with $K = 12$, $\varepsilon = 0.05$ and $\delta = 0.1$. Nodes belonging to the same community are identified by the same color. Nodes border colors compare the true partition with the obtained one: if the border is black, the units are correctly classified; otherwise, the color border identifies the belonging community in the true partition.

References

1. AlGhamdi, Z., Jamour, F., Skiadopoulos, S., Kalnis, P.: A benchmark for betweenness centrality approximation algorithms on large graphs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pp. 1–12 (2017)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
3. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE transactions on knowledge and data engineering **20**(2), 172–188 (2007)
4. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
5. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences **99**(12), 7821–7826 (2002)
6. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
7. Löffler, M., Phillips, J.M.: Shape fitting on point sets with probability distributions. In: European symposium on algorithms, pp. 313–324. Springer (2009)
8. Riondato, M., Kornaropoulos, E.M.: Fast approximation of betweenness centrality through sampling. Data Mining and Knowledge Discovery **30**(2), 438–475 (2016)
9. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. In: Measures of complexity, pp. 11–30. Springer (2015)
10. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of anthropological research **33**(4), 452–473 (1977)

Geographically weighted regression for spatial network data: an application to traffic volumes estimation

La regressione geografica ponderata per dati su network spaziali: un'applicazione alla stima dei volumi di traffico

Andrea Gilardi, Riccardo Borgoni and Jorge Mateu

Abstract Estimating traffic volumes on road networks represents a critical issue in various areas of research such as transport studies and road safety analyses. In these cases, the traffic figures are usually recorded via sparse manual counts or expensive automatic tools (e.g. cameras or inductive loops). However, given the increasing availability of mobile sensors (e.g. smartphones and GPS sat-nav), in the last years several methods were developed to extract traffic information from geo-referenced mobile devices. This paper proposes a geographically weighted regression (GWR) approach to combine fixed counts and GPS data to estimate traffic flows, re-adapting the appropriate statistical methods to the spatial network context. The suggested methodology is exemplified using data collected in the City of Leeds (UK).
Abstract La stima dei volumi di traffico rappresenta un problema rilevante in diversi ambiti come la mobilità urbana e le analisi sulla sicurezza stradale. I dati sul traffico vengono solitamente ottenuti da conteggi manuali o costosi strumenti automatici (e.g. telecamere o spire induttivi). Tuttavia, data la sempre maggiore disponibilità di sensori mobili (come smartphone e dispositivi GPS), negli ultimi anni sono stati sviluppati diversi approcci per ricavare stime di traffico da devices portatili. In questo articolo si propone un modello di regressione geografica ponderata (GWR) per la stima dei volumi di traffico che unisce dati GPS a conteggi manuali, riadattando la metodologia alla rete stradale. Il lavoro viene testato analizzando i conteggi stradali registrati nella città di Leeds (UK).

Key words: Geographically weighted regression, Linear networks, Network analysis, Traffic volumes estimation

Andrea Gilardi, Riccardo Borgoni
Department of Economics, Management and Statistics, University of Milano - Bicocca, e-mail: andrea.gilardi@unimib.it; e-mail: riccardo.borgoni@unimib.it

Jorge Mateu
Department of Mathematics, Universitat Jaume I , e-mail: mateu@uji.es

1 Introduction

Estimating traffic volumes on urban networks represents a critical issue in several areas of research such as transport studies [3], road safety analyses [6], and investigations on street networks efficiencies [5]. In these cases, the traffic flows can be used to quantify transportation demand, simulate driving and commuting behaviours, or approximate road risk exposure.

The traditional ways to compute traffic figures involve manual counts with ad-hoc cameras or automatic counts with road-fixed sensors (e.g. inductive loops and spirals). Unfortunately, both techniques have several limitations linked to their limited spatial coverage, high economical costs of installation and maintenance, and error proneness. For these reasons, in the last years several authors explored different approaches to extract traffic information from geo-referenced mobile sensors (e.g. smartphones and sat-navs), creating a complementary way to estimate the road counts. The mobile sensors have several benefits, such as extremely detailed spatial resolution and (usually) extensive coverage. However, since not all vehicles driving in a road network are equipped with GPS devices, the figures inferred from the data may actually underestimate the real traffic flows.

Hence, in this paper we propose a geographically weighted regression (GWR) approach to combine the two data sources (i.e. classic road counts and GPS figures) into a unique traffic estimate. Moreover, considering that traffic flows measurement from fixed and mobile data represents a classical example of a phenomenon occurring in a spatial network, we re-adapt the suggested statistical technique to this particular spatial domain.

2 Geographical weighted regression for network data

GWR is a local form of spatial analysis that allows the estimation of relationships between a dependent variable and a set of predictors that vary over space [4]. More precisely, given a sample of n units in a region S observed at locations $\mathbf{s}_i$, $i = 1,\dots,n$ according to a given coordinate reference system, the GWR model writes as

$$y(\mathbf{s}_i) = \alpha(\mathbf{s}_i) + \mathbf{x}'(\mathbf{s}_i)\beta(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1,\dots,n, \tag{1}$$

where $y(\mathbf{s}_i)$ denotes the response variable, $\alpha(\mathbf{s}_i)$ is the intercept, $\mathbf{x}(\mathbf{s}_i)$ a q-column vector of explanatory covariates, $\beta(\mathbf{s}_i)$ are the corresponding spatially-varying coefficients, and $\varepsilon(\mathbf{s}_i)$ is a zero mean random error. Since the parameters depend upon the spatial locations, this approach permits one to map the variation in the regression coefficients, gaining understandings of the spatial patterns between the predictor and response variables.

Parameter estimation at a selected location $\mathbf{s}_j \in S$ can be carried out using locally weighted least squares

$$\hat{\beta}(\mathbf{s}_j) = \left[\mathbf{X}'(\mathbf{s})W(\mathbf{s}_j)\mathbf{X}(\mathbf{s})\right]^{-1}\mathbf{X}(\mathbf{s})'W(\mathbf{s}_j)y(\mathbf{s}_j), \tag{2}$$

where $W(\mathbf{s}_j) = diag(w_{1j},\ldots,w_{nj})$ is a local weighting square matrix, w_{ij} is the weight associated to unit i when the regression is estimated at location $\mathbf{s}_j$, and $\mathbf{X}(s)$ represents the design matrix. The weights are defined in terms of a kernel function K that decays gradually with d_{ij}, i.e. the distance between the ith observation and the point $\mathbf{s}_j$. In particular, a Gaussian kernel function is adopted in this paper

$$K(d_{ij}) = \exp\{-d_{ij}^2/2h\}, \tag{3}$$

where the bandwidth parameter h determines the spatial range of the kernel. In the case study presented in the next section, the value of h is selected using cross-validation by minimising the mean square error of traffic flows predictions.

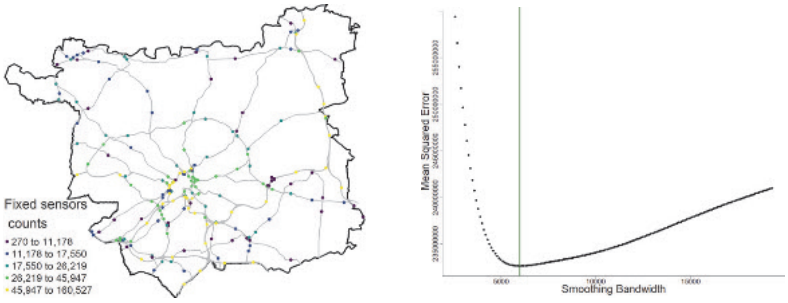
Usually, the inputs in Equation (3) are Euclidean distances in a planar setting, e.g. $d_{ij} = ||\mathbf{s}_j - \mathbf{s}_i||$. However, in our context, the sample units are observations recorded at a set of n road segments represented by their centroid locations $\mathbf{s}_i, i = 1,\ldots,n$. Hence, the distance d_{ij} should be calculated preserving the graph structure of the road network. In the rest of the paper we refer to the shortest path distance in order to take into account the spatial domain of the data. More precisely, indicating by $L = (V,E)$ the one-dimension graph object generated by the street network (where V and E denote the sets of vertices and edges, respectively), a path ρ_{ij} connecting any two generic locations $\mathbf{s}_i$ and $\mathbf{s}_j$ on the network is defined as a finite sequence $\{\mathbf{p}_m\}_{m=1}^{M}$ of adjacencent vertices in V such that the edges with endpoints $[\mathbf{s}_i, \mathbf{p}_1]$ and $[\mathbf{p}_\mathbf{M}, \mathbf{s}_j]$ belong to E. The length of ρ_{ij} can be computed as

$$||\mathbf{s}_i - \mathbf{p}_1|| + \sum_{m=1}^{M-1} ||\mathbf{p}_{m+1} - \mathbf{p}_m|| + ||\mathbf{p}_M - \mathbf{s}_j||,$$

and we define d_{ij} as the minimum length of all paths connecting $\mathbf{s}_i$ and $\mathbf{s}_j$ [1, 2].

3 Estimation of traffic flows in Leeds: data and results

The case study considered in this section is based on fixed and mobile daily traffic volumes recorded in the road network of Leeds (UK) from January to December 2019. The network and the GPS counts, which represent the spatial domain and the covariate used in our model (see Equation (1)), were obtained from TomTom Move service (`https://move.tomtom.com/`). The spatial network is composed by 8959 geo-referenced segments that are associated to traffic volumes estimated using mobile devices connected to cars and anonymous GPS-equipped smart-



(a) (b)

Fig. 1 Leeds road network and locations of fixed cameras used to detect traffic counts by the Department for Transport (a); MSE curve for bandwidth cross validation (b).

phones. Hence, the TomTom data have a reasonable spatial coverage, although they are known to underestimate the real flows.

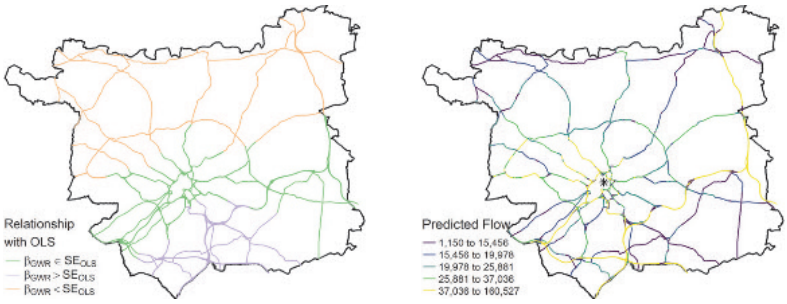
The regular counts, which are derived from fixed traffic cameras, were downloaded from the section Road Bulk Downloads of the platform Road Traffic Statistics developed by the UK Department for Transport (`https://roadtraffic.dft.gov.uk/downloads`). The 197 available count points locations were projected onto the road network, and, for each fixed camera, we assigned the corresponding traffic estimate to the overlapping road segment. The GWR was implemented assuming the actual frequencies observed at each count point as the response variable and the traffic flows measured by the mobile devices as predictor. Figure 1(a) displays the road network and the count points locations, which are distributed in several parts of the municipality.

As already mentioned, the smoothing parameter h in Equation (3) was estimated using leave-one-out cross-validation by minimising the mean squared error of traffic flows predictions. More precisely, we selected a series of bandwidth values in the range of the observed shortest path distances among all road cameras, and we calculated $\sum_{i=1}^{197}[y_i - \hat{y}_{\neq i}(h)]^2$ for all possible values of h. The quantity y_i denotes the observed road count, while $\hat{y}_{\neq i}(h)$ is the predicted flow obtained using bandwidth h and all observations but the ith one. The MSE estimates are reported in Figure 1(b) which suggests an optimal bandwidth as large as 5500 approximately.

Considering the smoothing parameter associated to the minimum MSE, we estimated the geographically weighted regression and predicted the traffic counts for

Table 1 Comparison between values detected by GPS sensors (first row), fixed cameras (second row) and predicted counts (third row) according to the model described in Equation (1).

	Minimum	1st Quartile	Mean	3rd Quartile	Maximum
GPS counts	0	865	1452	2272	12420
Fixed sensors counts	1150	12320	20999	38177	160527
Predictions	2820	16636	22965	32830	147694



(a) (b)

Fig. 2 Predicted daily traffic flows according to the model described in Equation (1) (a); Comparison with OLS estimates (b).

all segments in the network. As we can see from the equations reported above, the GWR is a local approach, which implies that each prediction requires the estimation of a different set of parameters. We explored this aspect more precisely by developing a comparison between the estimator detailed in Equation (2) and a classical OLS. The results are reported in Figure 2(a), where a segment is coloured in violet if the corresponding GWR estimate lays above the 95% confidence interval (CI) of the overall OLS estimate, in orange if the GWR estimate lays below the CI, and green otherwise. A clear spatial pattern has been found in the estimated coefficients. This points out that the relationship between point and GPS traffic measurements is not stationary in space and suggests the adoption of the local approach.

We report in Figure 2(b) a choropleth map of the estimated traffic counts at the road segment level. The figure clearly highlights several roads corresponding to a motorway (i.e. the yellow segments connecting the south area with the north/north-east) and the most important arterial thoroughfares reaching the city centre (i.e. the black star in the middle of the map). Roads in the north-west suburbs are found to be exposed to lower traffic flows as compared to the rest of the city.

As already mentioned, the TomTom figures underestimate the real flows, while the fixed cameras are too sparse to provide useful traffic estimates. The GWR approach, integrating the two data sources, combines their benefits. Table 1 details a convenient summary of the road counts employed in this paper, highlighting the merits of the GWR estimates. More precisely, the first and second rows summarise the GPS and camera data, respectively, and clearly point out that the mobile counts underestimate the real flows. The last row reports a summary of the predicted counts according to the GWR introduced in the previous paragraphs. We can observe that the predicted flows have a similar scale than real traffic data from fixed cameras while preserving a global coverage of the entire network. To conclude, we calculated the pseudo R^2 of the GWR finding values that ranged from 0.67 to 0.99 with a median value equal to 0.91. These quantities indicate a good performance for the estimated model.

4 Conclusions

This study demonstrates that GWR is a powerful tool to predict traffic flows at the road network level, combining data detected with fixed devices and GPS sensors. The classical approach was adjusted to take into account the particular spatial domain, substituting the Euclidean distances with the more appropriate shortest path distances. Our case study focused on daily traffic flows in the road network of Leeds from January to December 2019, and we showed that the suggested methodology allows a realistic estimation of traffic counts in all segments of a street network, combining the main benefits of the two data sources. In fact, the results detailed in Section 3 prove that the proposed model is suitable for the problem at hand.

We plan to extend the analysis presented in this work in several directions. First, the spatio-temporal dynamics of traffic flows could be explored, developing a traffic counts estimate for different hours of the day or different days of the week. Furthermore, we will enhance the geographically weighted regression including a few external covariates (e.g. road types, speed limit and road curvature) that could improve the model's fit.

References

1. Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M.: Analysing point patterns on networks—A review. Spatial Statistics **42**, p. 100435 (2021)
2. Barthélemy, M.: Spatial networks. Physics reports **499(1-3)**, pp. 1-101 (2011)
3. Caceres, N., Romero, L.M., Benitez, F.G. and del Castillo, J.M.: Traffic flow estimation models using cellular phone data. IEEE Transactions on Intelligent Transportation Systems **13(3)**, pp. 1430-1441 (2012)
4. Fotheringham, A.S., Brunsdon, C. and Charlton, M.: Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons (2003)
5. Zadeh, A.S.M. and Rajabi, M.A.: Analyzing the effect of the street network configuration on the efficiency of an urban transportation system. Cities **31**, pp. 285-297 (2013)
6. Zeng, J., Qian, Y., Wang, B., Wang, T. and Wei, X.: The impact of traffic crashes on urban network traffic flow. Sustainability **11(14)**, p. 3956 (2019)

Asymmetric Spectral Clustering: a comparison between symmetrizations

Spectral clustering asimmetrico: un confronto tra simmetrizzazioni

Cinzia Di Nuzzo and Donatella Vicari

Abstract In this work, spectral clustering of asymmetric data is addressed. In particular, two different methods to perform clustering have been compared: the application of spectral clustering on directed graphs represented by an asymmetric matrix; and the application of the classical spectral clustering algorithm once transformed the directed graph into an undirected one. To this end, some symmetrizations are described to convert the directed graph to an undirected one.

Abstract *In questo lavoro viene trattato lo spectral clustering di dati asimmetrici. In particolare, sono state confrontate due diverse metodologie per effettuare il clustering: l'applicazione dello spectral clustering su grafi diretti rappresentati da una matrice asimmetrica; e l'applicazione del classico algoritmo di spectral clustering dopo la trasformazione del grafo diretto in uno non diretto. A tal fine, sono stati descritte alcune simmetrizzazioni per convertire il grafo diretto in uno non diretto.*

Key words: spectral clustering, directed graph, symmetrizations

1 Introduction

Clustering on directed graphs has several useful applications, just think of network and social network analysis. So far, few spectral methods have been proposed for clustering this type of data, see [8], [3] and [2]. Many clustering methods on directed graphs rely on the symmetrization of the weighted matrix of the data by taking into account the out-degrees between the vertices of the graph. Once such a symmetrization has been computed, an undirected graph can be associated with the resulting symmetric matrix, and the standard clustering methods on undirected graphs can be applied. Here, we propose to apply the spectral clustering method proposed by [5]. In particular, three different ways to symmetrize an asymmetric matrix are presented: the first method is based on the symmetrization of the asymmetric weighted matrix associated with the directed graph; the second one is based on the symmetrization of the weighted matrix considering the out-degrees of the graph, while the last symmetrization takes into account both the in-degrees and out-degrees of

Cinzia Di Nuzzo
Dipartimento di Scienze Statistiche - Sapienza Università di Roma e-mail: cinzia.dinuzzo@uniroma1.it

Donatella Vicari
Dipartimento di Scienze Statistiche - Sapienza Università di Roma e-mail: donatella.vicari@uniroma1.it

the directed graph. The spectral clustering, proposed in [5] and applied to these three symmetrizations, has been compared with the asymmetric spectral clustering proposed by [8]. The rest of the paper is structured as follows: in Section 2, some definitions on directed graphs are given. In Section 3, the symmetrizations of the asymmetric matrix associated with the directed graph, and the undirected spectral clustering method are summarized. In Section 4 the asymmetric spectral clustering proposed by [8] is summarized. Section 5 provides a numerical example where a comparison is carried out.

2 Background theory

Let $\mathscr{G} = (V,E)$ be a directed graph, where $V = \{v_1,\ldots,v_n\}$ is the set of n vertices or nodes, $E \subseteq V \times V$ is the set of the edges or arcs. In a directed graph, an edge is an ordered pair (v_i,v_j), where $v_i,v_j \in V$ for $i,j \in \{1,\ldots,n\}$. A tuple of vertices $(v_1,v_2,\ldots,v_p)$ with $(v_h,v_{h+1}) \in E$ for $1 \leq h \leq p-1$, is a *path*. If for each pair of vertices v_i and v_j there is a path where $v_1 = v_i$ and $v_p = v_j$, then the directed graph is *strongly connected*. When $\mathscr{G}$ is a strongly connected graph, there is an integer $l \geq 1$ and a unique partition $V = V_0 \cup V_1 \cup \cdots \cup V_{l-1}$ such that for all $0 \leq r \leq l-1$ each edge $(v_i,v_j) \in E$ with $v_i \in V_r$ and $v_j \in V_{r+1}$, and where $V_l = V_0$. The graph $\mathscr{G}$ is *weighted* when there is a weight function $w : E \to \mathbb{R}^+$ which associates a positive value $w((v_i,v_j))$ to each edge $(v_i,v_j) \in E$. Let $\mathscr{G} = (V,E,W)$ be a weighted directed graph, where W is the *asymmetric weighted matrix* associated to $\mathscr{G}$. The out-degree and the in-degree of a vertex v_i are defined by $d_i^+ = \sum_{j=1}^n w_{ij}$ and $d_i^- = \sum_{j=1}^n w_{ji}$, respectively. Let $D_o = \text{diag}(d^+)$ be the diagonal matrix of the vertex-out degrees. Given a weighted directed graph $\mathscr{G}$, there is a natural random walk on the graph with *transition probability matrix of the Markov chain* equal to matrix $P = D_o^{-1}W$ with entries $p_{ij} = w_{ij}/d_i^+$. When $\mathscr{G}$ is assumed to be a strongly connected graph, the Markov chain P is *irreducible* and has a unique stationary distribution π, i.e. a unique probability distribution π (with n non-negative entries summing to one) that satisfies the following balance equations

$$P'\pi = \pi, \tag{1}$$

where $\pi = (\pi_1,\ldots,\pi_n)'$. Let us denote Π the diagonal matrix having π as main diagonal.

3 Symmetrizations of the asymmetric matrix W and spectral clustering algorithm

Several attempts have been made to generalize the spectral graph theory of undirected graphs to directed graphs and they have been mainly addressed towards the symmetrization of the weight matrix associated to the graph. The graph symmetrizations proposed in literature are summarized below (see [6] for further details).

First symmetrization: $W + W'$. The simplest symmetrization is obtained as

$$U = W + W'. \tag{2}$$

Note that this symmetrization does not take into account the in- or out- links between the vertices of the graph.

Second symmetrization: Random walk. The random walk symmetrization proposed by [1] is defined as

$$U_{rw} = \frac{\Pi P + P'\Pi}{2}. \tag{3}$$

This symmetrization is based on the stationary distribution Π of the graph and does not take into account the in-degrees in each vertex.

Third symmetrization: Degree-discounted. In order to symmetrize the graph, both in- and out- degrees are taken into account in [6].The idea behind this is to look for a similarity measure that takes into account the in- and out-degrees between the vertices of the directed graph. In fact, nodes with high degrees will share many common edges (in- or out-) with other nodes simply by virtue of their higher degrees. The idea is to weigh the edges so that vertices with high (in- or out-) degree are weighted less in the symmetrization process. The reason is very simple: vertices with high (in- or out-) degrees are actually less informative in describing the links between the vertices of the graph than nodes with lower degrees. The symmetrization proposed in [6] is given by the following expression

$$U_d = D_o^{-1/2} W D_i^{-1/2} W' D_o^{-1/2} + D_i^{-1/2} W' D_o^{-1/2} W D_i^{-1/2} \tag{4}$$

where D_i =diag(d^-) is the diagonal matrix of the in-degrees.

Once a directed graph has been transformed into an undirected graph, the spectral clustering algorithm proposed in [5] can be can be run as summarized in Algorithm 1.

Algorithm 1 Spectral Clustering algorithm (see [5])

Let $\mathscr{G}_u = (V_u, E_u, W_u)$ be a undirected weighted graph, where $V_u = \{v_1^u, \ldots, v_n^u\}$ is the set of the vertices, E_u is the set of the edges, and $W_u = \left(w_{ij}^u\right)$ is the symmetric weighted matrix associated to $\mathscr{G}_u$.

Input: Symmetric weighted matrix W_u; number of clusters K.

1. Compute the degree matrix D_u =diag(d_i^u) associated to the undirected graph $\mathscr{G}_u$, where d_i^u is the degree of the vertex v_i^u associated to the undirected graph $\mathscr{G}_u$ and it is computed as $d_i^u = \sum_{i \neq j} w_{ij}^u$
2. Calculate the normalized symmetric Laplacian matrix $L_{\text{sym}} = D_u^{-1/2} W_u D_u^{-1/2}$.
3. Consider matrix Y having the eigenvectors associated to the K largest eigenvalues of L_{sym} as columns
4. Re-normalize the rows of Y to have unit length yielding $X \in \mathbb{R}^{n \times K}$.
5. Perform k-means algorithm on matrix X.

4 Asymmetric Spectral Clustering

Let us suppose to divide the graph into K subgraphs. In order to find the best partition of the graph $\mathscr{G}$, in the standard undirected spectral clustering method a normalized cut is introduced and the clustering is obtained thanks to a minimization of this cut (see [7] and [4]). An extension to directed graphs has been introduced by [8], where the directed *Ncut* of a cluster C is defined as

$$Ncut_{dir}(C) = \frac{\sum_{i \in C, j \in \bar{C}} \pi_i p_{ij}}{\sum_{i \in C} \pi_i} + \frac{\sum_{i \in C, j \in \bar{C}} \pi_j p_{ji}}{\sum_{j \in \bar{C}} \pi_j}. \tag{5}$$

In [8], the $Ncut_{dir}$ is minimized by the eigenvectors of the following matrix

$$\Theta = \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P' \Pi^{1/2}}{2}. \tag{6}$$

The Asymmetric Spectral Clustering is summarized in Algorithm 2.

Algorithm 2 Asymmetric Spectral Clustering algorithm (see [8])

Input: Asymmetric weighted matrix W associated to directed graph $\mathcal{G}$; number of clusters K.

1. Compute the transition matrix P and the diagonal matrix Π from its unique stationary distribution P.
2. Compute the (symmetric) Laplacian matrix Θ (6).
3. Consider the eigenvectors of Θ and introduce matrix Z with columns equal to the eigenvectors $\gamma_1, \ldots, \gamma_K$ associated with the K largest eigenvalues of Θ.
4. Perform k-means algorithm on matrix Z.

5 Numerical experiments

In order to compare the methods described in Sections 3 and 4, we present an example with artificial data.

The artificial data are described in Figure 1, the graph $\mathcal{G} = (V, E, W)$ is composed by three groups: $C_1 = \{v_1, v_2, v_3\}$, $C_2 = \{v_4, v_5\}$ and $C_3 = \{v_6, v_7, v_8\}$. Specifically, in Figure 1-*a*) the edges of the graph $\mathcal{G}$ are displayed with different thickness according to their weights (i.e., solid , dashed and dotted lines represent weights in descending order, respectively).

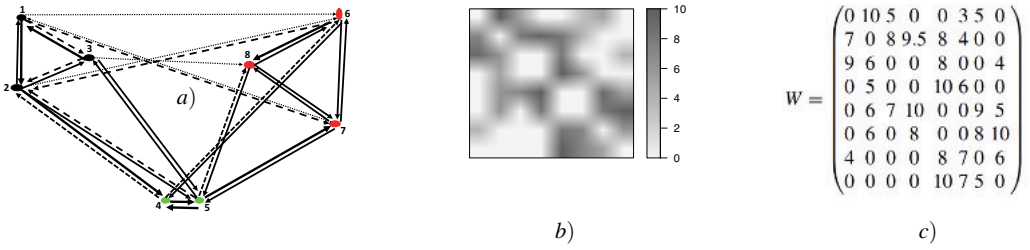


Fig. 1: *Artificial data*. *a*) Representation of the graph $\mathcal{G}$. *b*) Heat map of the asymmetric matrix W. *c*) Asymmetric weighted matrix W associated to graph $\mathcal{G}$.

In Figure 2, the geometric features of the spectral clustering algorithm (Algorithm 1) applied on the three symmetrizations described in Section 3 have been represented. The correct clustering results are attained from the U and U_{rw} symmetrizations; conversely, the spectral clustering applied on the U_d symmetrization does not provide the correct classification, because the accuracy is equal to 0.5 and the confusion matrix is

$$\begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The asymmetric spectral clustering described in Algorithm 2 has been also carried out, the geometric features of the algorithm are shown in Figure 3, and the correct clustering is obtained.

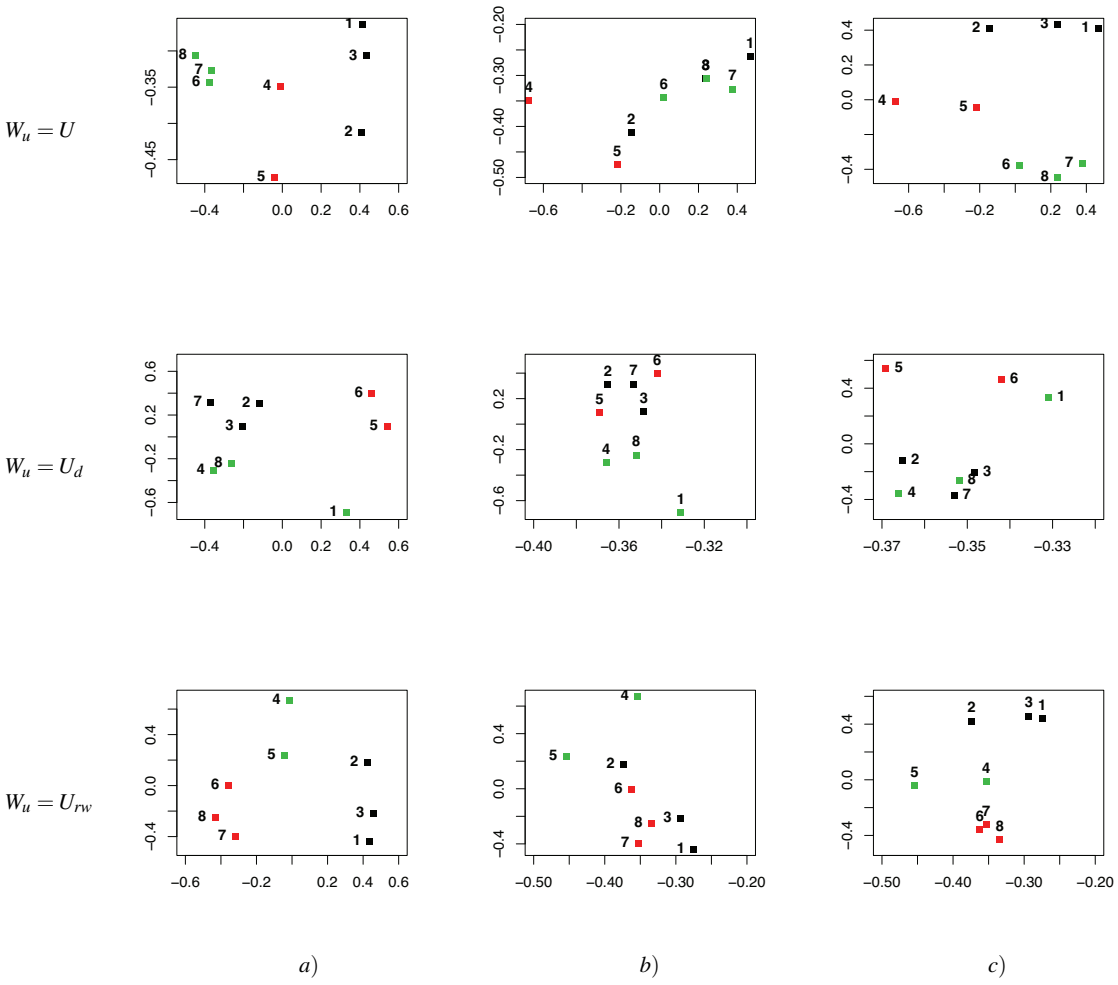
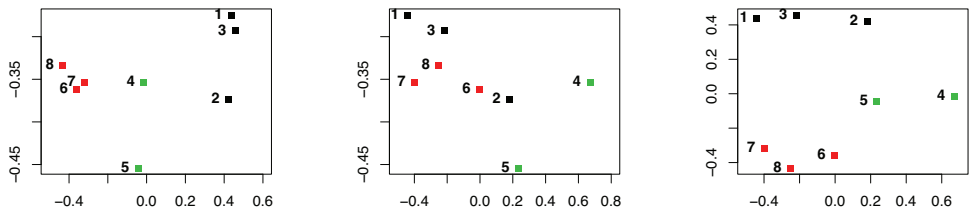


Fig. 2: *Spectral clustering after symmetrizations*. Embedded data associated to L_{sym} according to: a) the first two eigenvectors; c) the first and third eigenvector; d) the last two eigenvectors (different colors denote different clusters).



a) *b*) *c*)

Fig. 3: *Asymmetric spectral clustering (Algorithm 2)*. Embedded data associated to Θ according to: *a*) the first two eigenvectors; *c*) the first and third eigenvector; *d*) the last two eigenvectors (different colors denote different clusters).

To conclude, the asymmetric spectral clustering described in Algorithm 2 provides the correct classification, while the spectral clustering applied after the symmetrizations provides the correct results only when the U_{rw} and U symmetrizations are used. Finally, the Laplacian embedding structure provided by the U_d symmetrization, shown in Figure 2, is different from the others, because U_d also takes into account and incorporates the information from the in-degrees of the nodes of the directed graph $\mathscr{G}$. Therefore, this lays the groundwork for further studies on the information extracted from the different symmetrizations by carrying out studies on both artificial and real data.

References

1. Gleich, D.: Hierarchical Directed Spectral Graph Partitioning. MS&E 337 - Information Networks. (2006)
2. Lejay, A.: Asymmetric Spectral clustering. Technical Report. Inria Nancy - Grand Est. url: https://hal.inria.fr/hal-02372570. (2019)
3. Meila, M. and Pentney, W. : Clustering by weighted cuts in directed graphs. doi: 10.1137/1.9781611972771.13. (2007)
4. Meila, M., Shi, J. : A random walks view of spectral segmentation. In: 8th International Workshop on Artificial Intelligence and Statistics (AISTATS). (2001)
5. Ng, A., Jordan, M., and Weiss, Y.: On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, **14**. (2002).
6. Satuluri, V., Parthasarathy, S.: Symmetrizations for Clustering Directed Graphs. In: EDBT/ICDT'11: Proceedings of the 14th International Conference on Extending Database Technology, pp 343–354. (2011)
7. Shi, J.,Malik, J.: Normalized cuts and image segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888–905. (2000)
8. Zhou, D., Huang, J., and Schölkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: ICML '05, pp.1036-1043. (2005)

Community detection of seismic point processes

Identificazione di comunità in processi di punto sismici

Valeria Policastro, Nicoletta D'Angelo, and Giada Adelfio

Abstract In this paper, we combine `robin` and Local Indicators of Spatio-Temporal Association (LISTA) functions. `robin` is an R package to assess the robustness of the community structure of a network found by one or more methods to give indications about their reliability. We use it to propose a classification algorithm of events in a spatio-temporal point pattern, by means of the local second-order characteristics and the community detection procedure in network analysis. We demonstrate the proposed procedure on a real data analysis on seismic data.

Abstract *In questo articolo, uniamo* `robin` *con le funzioni LISTA (Indicatori spazio-temporali locali di associazione).* `robin` *è un pacchetto R per valutare la robustezza della struttura in comunità di una rete ottenuta da uno o più metodi, per dare indicazioni sulla loro affidabilità. In particolare, proponiamo un algoritmo di classificazione degli eventi in un pattern di punti spazio-temporale, attraverso le caratteristiche locali del secondo ordine, e la procedura di identificazione delle comunità nell'analisi di rete. La bontà della procedura proposta viene valutata a partire da un catalogo sismico.*

Key words: network analysis, community detection algorithm, second-order characteristics, spatio-temporal point processes, statistical validation, earthquakes

Valeria Policastro
Institute for applied mathematics "Mauro Picone" (National Research Council) e-mail: `valeria.policastro@unicampania.it`

Nicoletta D'Angelo and Giada Adelfio
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, Italy e-mail: `nicoletta.dangelo@unipa.it;giada.adelfio@unipa.it`

1 Introduction

In network analysis, many community detection algorithms have been developed. However, their implementation leaves unaddressed the question of the statistical validation of the results. `robin` (Policastro et al., 2021) (ROBustness In Network) is an R Core Team (2022) package which answers this question, assessing the robustness of the community structure of a network found by one or more methods to give indications about their reliability. The procedures implemented in `robin` detect whether the community structure found by a set of algorithms is statistically significant and compare two selected detection algorithms on the same graph to choose the one that better fits the network of interest. We propose the use of the package on an application to the spatio-temporal point process theory, with the use of Local Indicators of Spatio-Temporal Association (LISTA) functions (Siino et al., 2018). In particular, our aim is to analyse seismic data, in order to identify different types of events, which could represent the background and the triggered events (Siino et al., 2020).

The structure of the paper is as follows. Section 2 presents the `robin` package of the statistical software R Core Team (2022). Section 3 recalls the LISTA functions and proposed a distance for building the adjacency matrix to built the network. Section 4 contains the application to seismic data. Conclusions come in Section 5.

2 Comparison and validation of communities with `robin`

The R Core Team (2022) package `robin` assesses the robustness of the community structure of a network found by one or more methods to give indications about their reliability. The aim is twofold: to compare two selected detection algorithms on the same graph to choose the one that better fits the network of interest, and to detect whether the community structure found by an algorithm is statistically significant.

For the first aim, the second workflow described in Policastro et al. (2021) can be used. It helps to choose, among different community detection algorithms, the one that best fits the network of interest, comparing their robustness two at a time.

Once the best community detection algorithm is found, the stability of the partitions can be tested against random perturbations of the original graph structure through the first workflow (also in Policastro et al. (2021)). To address this issue, a null model and a stability measure must be specified. For the construction of the null model, we explored the *dk* null random model provided in Orsini et al. (2015), using the parameter *dk* 2 in this paper. See Policastro et al. (2021) for further methodological details.

The first workflow finds a partition C_1 for the real network and a partition C_2 for the null network, it perturbs both networks, it retrieves two new partitions $C_{1(p)}$ and $C_{2(p)}$, and it calculates two clustering distances (for the real network and the null network) between the original partitions and the ones obtained from the perturbed network as $M(C_{1(p)}, C_1)$ and $M(C_{2(p)}, C_2)$. This process is carried out at different

perturbation levels $p \in [0 : 0.05 : 0.6]$ to create two curves, one for the real network and one for the null model, and to test their similarity by two functional statistical tests. The comparison between the two curves enables to reconsider the problem regarding the significance of the retrieved community structure in the context of stability/robustness of the recovered partition against perturbations. The basic idea is that if small changes in the network cause a completely different grouping of the data, the detected communities are not reliable. The second workflow has the same structure of the first one, but instead of comparing the real network with the null network, it takes the real network and apply two different community detection algorithms on it to compare them. `robin` embeds all the community detection algorithms present in `igraph`. They can be classified as modularity based methods, divisive algorithms, methods based on statistical inference, dynamic algorithms, and some alternative methods. Moreover, it provides the possibility to input a custom external function to detect the communities.

3 Classification of events using LISTA functions

We consider a spatio-temporal point process with no multiple points as a random countable subset X of $\mathbb{R}^2 \times \mathbb{R}$, where a point $(\mathbf{u},t) \in X$ corresponds to an event at $\mathbf{u} \in \mathbb{R}^2$ occurring at time $t \in \mathbb{R}$. The Local Indicators of Spatio-Temporal Association (LISTA) functions are a set of functions that are individually associated with each of the points of the point pattern and can provide information about the local behaviour of the pattern (Siino et al., 2018). Several second-order characteristics of point processes are considered for building LISTA functions, such as the K-function, the pair correlation function, and the product density function. In this paper, we consider the local spatio-temporal pair correlation function $\hat{g}^{(i)}_{\varepsilon,\delta}(r,h) = \frac{1}{4\pi r|W\times T|\hat{\lambda}^2}\sum_{j\neq i}\frac{\kappa_\varepsilon(||\mathbf{u}_i-\mathbf{u}_j||-r)\kappa_\delta(|t_i-t_j|-h)}{\omega(\mathbf{u}_i,\mathbf{u}_j)\omega(t_i,t_j)}$, where r and h are some spatial and temporal distances, κ_ε and κ_δ are kernel functions with bandwidths ε and δ, respectively, and ω is the edge correction factor. Both kernels are computed using the Epanechnikov kernel and the bandwidths are estimated with a direct plug-in method. Extending González et al. (2021) to space and time, LISTA functions consist of a functional dataset $\{\hat{g}^{(i)}_{\varepsilon,\delta}\}_{i=1}^n$, on some compact region $W \times T = [\varepsilon, r_{max}] \times [\delta, h_{max}]$, where r_{max} and t_{max} have to be prior chosen (by using Diggle's rule for instance). We start by defining, for any r and h, the following transformation $g^*_i(r,h) = \hat{g}^{(i)}_{\varepsilon,\delta}(r,h) + \hat{\mathbb{E}}_!\left[\hat{g}^{(i)}_{\varepsilon,\delta}(r,h)\right]$, where $\mathbb{E}_!\left[\hat{g}^{(i)}_{\varepsilon,\delta}(r,h)\right]$ is the Palm expectation. We follow Siino et al. (2018) and consider $\hat{\mathbb{E}}_!\left[\hat{g}^{(i)}_{\varepsilon,\delta}(r,h)\right]_{Poisson} = \lambda^2 + \frac{\lambda}{|W\times T|}$. These translated LISTA functions are above one for any $r \in W$ and $h \in T$ and hold that property for some fixed r and h, and individual surfaces are interpretable as long as they are not isolated from the others. The higher transformed LISTA functions have associated points with the most crowded neighbourhoods. We can simplify that relative information by standardising $\rho_i(r,h) = \log\frac{g^*_i(r,h)}{\sum_i g^*_i(r,h)}$. Each component $\rho_i(r,h)$

represents a relative weight as part of a whole, which in this case is a linear transformation of the pair correlation. Some multivariate methods map the observations in the Euclidean space while accounting for LISTA functions geometry. We employ the Euclidean metric in L_2 by letting $\rho_i(r,h), \rho_j(r,h)$ be two transformed LISTA functions, with distance given by

$$d^2(\rho_i, \rho_j) = \int_T \int_W [\rho_i(r,h) - \rho_j(r,h)]^2 \mathrm{d}r\mathrm{d}t. \tag{1}$$

After applying this distance, all the standard methods of multivariate statistics that rely on dissimilarity measures can be employed.

4 Application

The analysed data concern 1111 earthquakes occurred in Greece between 2005 and 2014. They come from the Hellenic Unified Seismic Network (H.U.S.N.), and they have been recently analysed in D'Angelo et al. (2022) by means of local point process models. First, a network is built by computing the LISTA functions $\hat{g}^{(i)}_{\varepsilon,\delta}(r,h)$ on each point of the pattern, and then, an adjacency matrix is provided, taking as links only the pair of points whose distance (1) is less than the 5% percentile of the resulting distribution of distances. We then apply the second workflow (`robinCompare()` function) on the network, to choose the community detection algorithm that better fits it. We found that the `infomap` algorithm was the most stable algorithm for our network (Figure 1). This belongs to the community detection methods based on statistical inference, and it is based on information-theoretic principles: basically, it finds the community detection structure that minimizes the description length for a random walk on the graph given by the map equation over possible network partitions (Rosvall and Bergstrom, 2008). Therefore we proceed testing also the stability of the partitions found by `infomap` against random perturbations of the original graph structure (`robinRobust()` function). So we construct the curves in Figure 2(a) that represent the measure between the partition of the original unperturbed graph and the partition of each perturbed graph (blue curve). The red curve is obtained in the same way but considering the random graph as the original one. Constructed the curves, we can proceed testing with the Interval Testing Procedure (`robinFDATest()` function) in Figure 2(b-c), which lead to most of the p-values being significant. We also run the Gaussian Process test (`robinGPTest()` function) that, with a Bayes Factor equal to 126, indicates an extreme evidence that the two curves come from two different processes. Therefore the communities identified on the proposed network by `infomap` are statistically significant. The results of the community detection and a possible interpretation follow. Figure 3 depicts the points belonging to the largest cluster found (left panel), and the remaining ones. This means that the points in the left panel are the most similar in terms of local second-order structure, belonging to the same cluster, while the

remaining clusters exhibit more variability (right panel). This is a promising result as it gives us more insight into the local structure of the analysed point pattern, basically splitting it into two sub-patterns. Indeed, separating point patterns into clutter and feature is a common problem in point process theory (e.g. Siino et al. (2020)). In particular, the two regions with the highest density in the left panel of Figure 3, are the same identified by D'Angelo et al. (2022), where points exhibit a different behaviour in terms of local structure, and by Siino et al. (2018) as the background.

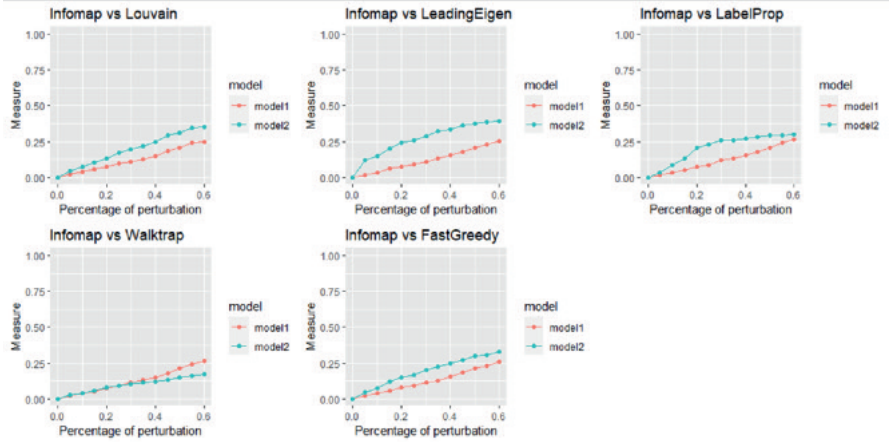


Fig. 1: Curves comparing different algorithms against `infomap` algorithm (red curve)

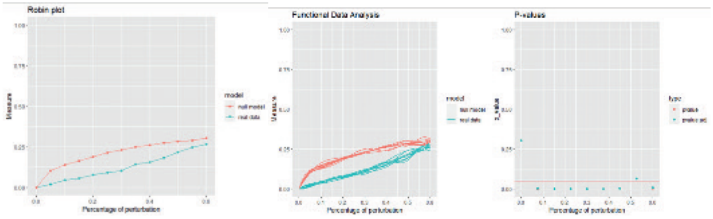


Fig. 2: (a) Curves of the null model and the real data applying `infomap` algorithm. (b) Curves for the testing procedure. (c) Corresponding p-values and adjusted p-values for all the intervals with the horizontal red line on to the critical value 0.05.

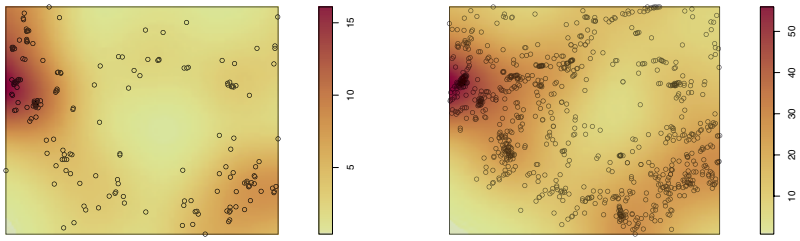


Fig. 3: Kernel smoothed intensity functions from the two point patterns, resulting from the largest cluster with 189 points (left panel) and the remaining ones belonging to smaller clusters (right panels). Overlaid in black, the observed point pattern.

5 Conclusions

In this paper, we have proposed a classification algorithm of events generated by a spatio-temporal point pattern, by means of the local second-order characteristics (LISTA functions) and the community detection procedure in network analysis. The R Core Team (2022) package `robin` helped to choose, among different community detection algorithms, the one that best fits the network of interest, comparing their robustness two at a time. Doing so, we found that `infomap` was the most stable clustering for our network. We then tested the stability of the partitions found by `infomap` against random perturbations of the original graph structure, finding that the identified communities on the proposed network build on distances among LISTA functions are significant. This paper represents a preliminary work on this proposal. In future we aim at first, changing the number of points "linked" into the network. This could be done by selecting a distance with some specific meaning into the context of application. Then, the proposed procedure could be tested on a large number of simulation studies.

References

D'Angelo, N., Siino, M., D'Alessandro, A., and Adelfio, G. (2022). Local spatial log-gaussian cox processes for seismic data. *Advances in Statistical Analysis. Forthcoming*.

González, J. A., Rodríguez-Cortés, F. J., Romano, E., and Mateu, J. (2021). Classification of events using local pair correlation functions for spatial point patterns. *Journal of Agricultural, Biological and Environmental Statistics*, 26(4):538–559.

Orsini, C., Dankulov, M. M., Colomer-de Simón, P., Jamakovic, A., Mahadevan, P., Vahdat, A., Bassler, K. E., Toroczkai, Z., Boguná, M., Caldarelli, G., et al. (2015). Quantifying randomness in real networks. *Nature communications*, 6(1):1–10.

Policastro, V., Righelli, D., Carissimo, A., Cutillo, L., and Feis, I. D. (2021). ROBustness In Network (robin): an R Package for Comparison and Validation of Communities. *The R Journal*, 13(1):292–309.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Siino, M., Rodríguez-Cortés, F. J., Mateu, J., and Adelfio, G. (2018). Testing for local structure in spatiotemporal point pattern data. *Environmetrics*, 29(5-6):e2463.

Siino, M., Rodríguez-Cortés, F. J., Mateu, J., and Adelfio, G. (2020). Spatio-temporal classification in point patterns under the presence of clutter. *Environmetrics*, 31(2):e2599.

An Explorative analysis of Different Distance Metrics to Compare Unweighted Undirected Networks

Analisi Esplorativa di Differenti Metriche di Distanza per Confrontare Reti Non Orientate e Non Pesate

Anna Simonetto[1], Matteo Ventura[2], Gianni Gilioli[1]

Abstract Networks are mathematical structures that make it possible to represent complex systems by characterising the relationships existing between the various elements of the network. In order to understand the differences between two different systems, tools are needed to quantify these differences. The purpose of this work is an exploratory analysis of the impacts that different distance metrics have on the comparative evaluation of two networks, at increasing levels of network perturbation. Preliminary results show that, depending on the chosen metric, it is possible to amplify or reduce the effect of perturbation. As the disturbance increases, the differences between the distance assessment systems are reduced.

Abstract *Le reti sono strutture matematiche che permettono di rappresentare sistemi complessi caratterizzando le relazioni esistenti tra i vari elementi della rete. Al fine di comprendere le differenze tra due sistemi diversi, sono necessari strumenti per quantificare queste differenze. Lo scopo di questo lavoro è un'analisi esplorativa degli impatti che diverse metriche di distanza hanno sulla valutazione comparativa di due reti, a livelli crescenti di perturbazione della rete. I risultati preliminari mostrano che, a seconda della metrica scelta, è possibile amplificare o ridurre l'effetto della perturbazione. All'aumentare della perturbazione, le differenze tra i sistemi di valutazione della distanza si riducono.*

Key words: Undirect graph, Euclidean distance, Spectral method, adjacency matrix, ecological networks.

---

[1] Dip. di Ingegneria Civile, Architettura, Territorio, Ambiente e di Matematica, Università di Brescia; anna.simonetto@unibs.it, gianni.gilioli@unibs.it;

[2] Dip. di Economia e Management, Università di Brescia; m.ventura007@unibs.it.

1 Comparing networks

A network is a data structure describing the interaction pattern existing between a set of elements (e.g., in an ecological perspective the elements could be species, individuals, habitats) by the properties of the links that connect each pair of elements, defined as arcs. In unweighted networks the arc is characterized only by the presence/absence (its value is 1 or 0, respectively). In weighted networks, the weight is proportional to the intensity of the relationship existing between the two elements, usually varying in the range [0, 1] or [-1, 1] in the case of signed networks. The comparison of two networks allows the differences in the two systems to be assessed in terms of the differences in the relationships between the elements.

In research on complex systems, the problem of comparing networks is ubiquitous and not trivial. A tradeoff among interpretability, effectiveness of the results and computational efficiency is needed [8]. In fact, the method applied to measure differences between networks influences the result of network comparison. The choice of the correct method must therefore firstly be based on a clear identification of which differences are to be measured and then on an understanding of how the different methods considered represent these differences.

Some authors have already explored several methods to compare networks. For instance, Tantardini et al. [8] presented a review on methods for comparing networks, both for networks with node correspondence and for networks with unknown node correspondence and tested them on synthetic random networks under some types of perturbations. Wilson and Zhu [9] examined the performance of spectra as graph representation. Specifically, the authors investigated the cospectrality of various matrix representations and compared the Euclidean distance between spectra and the edit distance between graphs.

The aim of the work is to understand how the different used distance metrics describe situations of diversity between networks. Specifically, we explored the performances of three distance measures and three spectral methods applied to compare networks created with the same nodes.

A simulation study was conducted to understand the specific response characteristics of the investigated methods. Specifically, we considered as perturbation the deletion of an edge and with these simulations we aimed at understanding: i) how these metrics react to an increasing number of perturbations, ii) what is the reaction when a node is completely disconnected, and iii) the impact of node degree on the value of the distance.

2 Distance metrics

Mathematically, the network can be represented by a graph. A graph is described by $G = (V, E)$, where $V = \{1, \dots, n\}$ is the set of vertex, i.e. the nodes or elements of the network, and E is the edge set, that is a subset of the set $V \times V$ of ordered pairs of distinct vertices or nodes. The graph is said to be undirected if the edge (i, j) and its

opposite (j, i) are both in E [4]. A graph can be represented by a squared $N \times N$ adjacency matrix $\boldsymbol{A}$ where each row/column corresponds to a node and each cell represents an edge, set to one if the edge exists and zero otherwise. The adjacency matrix of an undirected graph is symmetric [2].

Two classes of methods for comparing networks are exposed hereafter: (1) distances between adjacency matrices, (2) spectral methods.

2.1 Distances between adjacency matrices

Given two $N \times N$ adjacency matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, four distances are identified: Minkowski, Manhattan, Euclidean and Chebyshev distance.

Minkowski Distance Also known as $L_p -$ metric. It is identified by the following expression [6]:

$$d_{MK}(\boldsymbol{A}_1, \boldsymbol{A}_2) = \left[\sum_{i,j=1}^{N} |a_{1,ij} - a_{2,ij}|^p\right]^{\frac{1}{p}} \qquad p \geq 1. \tag{1}$$

Manhattan Distance Also known as $L_1 -$ metric. Its definition is based on Minkowski distance, using $p = 1$ [7]. It corresponds to the sum of the absolute differences, i.e., all the matrices' elements are equally weighted.

Euclidean Distance Also known as $L_2 -$ metric. It is a specific case of Minkowski distance, using $p = 2$ [3]. With this distance, greater weight is attached to longer distances.

Chebyshev Distance It is also called $L_\infty -$ metric and it is a special case of Minkowski distance where p goes to infinity [1], i.e., only the highest difference influences the metric value:

$$d_{CH}(\boldsymbol{A}_1, \boldsymbol{A}_2) = \lim_{p \to \infty} \left[\sum_{i,j=1}^{N} |a_{1,ij} - a_{2,ij}|^p\right]^{\frac{1}{p}} = \max_{i,j} |a_{1,ij} - a_{2,ij}|. \tag{2}$$

2.2 Spectral methods

These approaches are based on spectral theory, that allows to describe networks' structural properties through eigenvalues and eigenvector of a matrix.

The spectrum $\boldsymbol{\Lambda}$ is the sorted sequence of matrix eigenvalues λ and the spectral distance between two graphs is the Euclidean distance between the two correspondent matrices' eigenvalues [9]:

$$d_\lambda(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2) = \sqrt{\sum_{i=1}^{N} \left(\lambda_{1,i} - \lambda_{2,i}\right)^2}. \quad (3)$$

In addition to the adjacency matrix, there are two other possible matrix representations of a graph: the Laplacian and Normalised Laplacian matrix.

The Laplacian matrix $\boldsymbol{L}$ is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, where is the adjacency matrix of the graph and $\boldsymbol{D}$ is the diagonal degree matrix [9].

The Normalised Laplacian matrix, instead, is defined as:

$$\boldsymbol{\mathcal{L}}(i,j) = \begin{cases} 1 & if\ i = j \text{ and } d_i \neq 0 \\ -\dfrac{1}{d_i d_j} & if\ i \neq j \text{ and } i \text{ is adiacent to } j, \\ 0 & \text{otherwise} \end{cases}$$

where d_i and d_j are respectively the degree of the node i and the degree of the node j. The Normalised Laplacian can be also written as $\boldsymbol{\mathcal{L}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$.

3 Simulation study

In this section we test the responses of the above proposed metrics through a simulation experiment. All the analyses have been implemented in R [5].

We started from a full connected, unweighted network composed by twelve nodes. At each iteration the distance between the fully connected network and a disturbed network will be calculated. At step 0 the disturbed network is equal to the fully connected network and itself. At each subsequent iteration the comparison network is perturbed by removing a connection arc. After each perturbation the distance of the disturbed network from the full connected one was measured. This is aimed at checking the distances response with increasing number of removed edges, and thus, with decreasing graph density.

To investigate the distances response to the decrease of the node degree and to the complete disconnection of a node, an ordered edge removal criterion was applied. Considering the upper triangular half of the completely connected graph's adjacency matrix, the ordered removal consists in turn into zero the elements of the first row and proceed with the next row only when the previous is full of zeroes; namely, if we consider a graph, the ordered removal consists in removing all the edges from a node and proceed with another node only when the previous one is completely disconnected from the network. This edge removal criterion leads to the generation of a set of nested networks, i.e., the edges removed from the network with n perturbations do not exist in all the subsequent networks with $n + m$ removed edges.

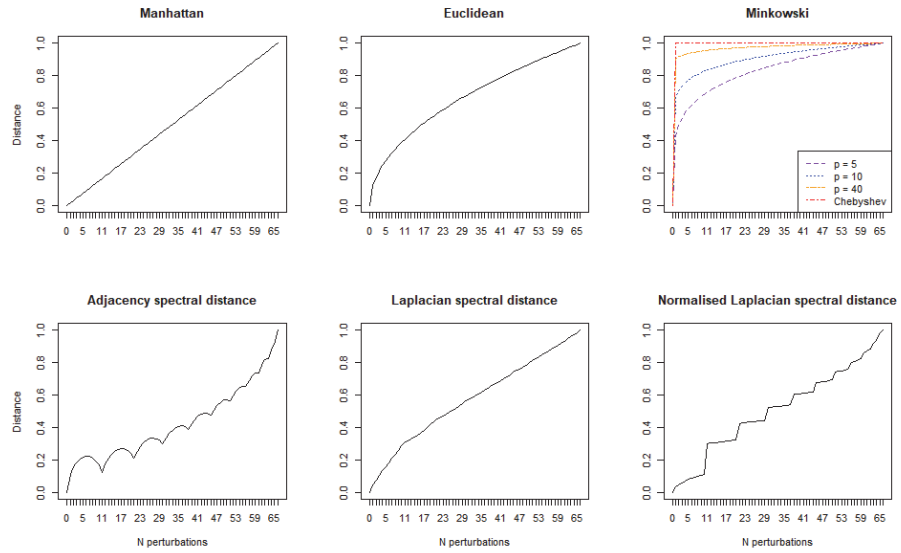
From the results reported in Figure 1, it appears that all the distance metrics have a growing monotonous pattern. We note that, considering the Minkowski distance,

when p increases, the metric distances tend to weight equally the perturbations independently of the number of deleted edges. Namely, the Manhattan distance weight equally all the perturbations, up to the extreme case represented by Chebyshev distance that weights in the same way one perturbation or infinite perturbations.

Instead, spectral methods show some interesting patterns: both the adjacency and the Normalised Laplacian spectral distance show a non-monotonous pattern. We observe level-shifts that seem to be dependent on the position of the removed edges, indeed, spectral methods present a step pattern where the level change happens when, once a node is completely disconnected, the first edge of another node is removed.

This first study shows that, particularly for disturbances of less than 50% of the number of arcs in the fully connected network, the choice of comparison metric influences the assessment of the distance between the compared networks.

Figure 1:Results of the ordered edge removal



4 Conclusions

In this work we compare different distance metrics to estimate the differences between two or more networks.

The simulation study highlights the differences in the distances estimation using four distance metrics and three spectral methods. The networks comparison is performed according to an increasing levels of perturbation. The Manhattan distance estimation resulted not influenced by the disturbance levels, Euclidean distance weighs more heavily on differences at low disturbance levels, Normalised Laplacian spectral

distance weighs more heavily on disturbances that lead to completely disconnecting a network node.
Further developments of the study will be devoted to deep the possible dependance of the distance estimation from the order in which edges are removed.

References

1. Coghetto, R. (2016). Chebyshev Distance. *Formalized Mathematics*, *24*(2), 121–141. https://doi.org/10.1515/forma-2016-0010
2. Coscia, M. (2021). The Atlas for the Aspiring Network Scientist. *ArXiv:2101.00863 [Physics]*. http://arxiv.org/abs/2101.00863
3. Dattorro, J. (2008). *Convex optimization & Euclidean distance geometry* (Version 2008.02.29). Meboo.
4. Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.
5. R Core Team (2021). R: *A language and environment for statistical computing.* Vienna, Austria. https://www.R-project.org/.
6. Richardson, G. D. (1981). *The appropriateness of using various Minkowskian metrics for representing cognitive configurations*. *13*, 475–485.
7. Sowell, K. O. (1989). Taxicab Geometry—A New Slant. *Mathematics Magazine*, *62*(4), 238–248. https://doi.org/10.1080/0025570X.1989.11977445
8. Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Scientific Reports*, *9*(1), 17557. https://doi.org/10.1038/s41598-019-53708-y
9. Wilson, R. C., & Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, *41*(9), 2833–2841. https://doi.org/10.1016/j.patcog.2008.03.011

Gender, attitudes and family ties

Parents of a disabled child in Italy: less healthy but more civically engaged

Avere un figlio disabile in Italia: Genitori meno in salute ma più socialmente impegnati

Nicoletta Balbo and Danilo Bolano

Abstract This article provides novel evidence of spillover effects of children disability on different life domains of their parents. We focus on health and social and civic engagement of the parents. Using data from a nationally representative household survey in Italy (Aspetti di Vita Quotidiana), we show that parents of a disabled child report lower levels of health both in terms of general health and mental health. At the same time these parents show a higher propensity to be socially engaged.

Abstract *Questo articolo mostra come la disabilità nei minori sia un fenomeno che trascende il minore stesso ma ha diramazioni importanti sui genitori. Gli autori, utilizzando dati dall'indagine ISTAT - Aspetti di Vita Quotidiana, mostrano la presenza di effetti negativi sulla salute (salute generale e salute mentale) dell'avere un figlio disabile ma anche una propensione piu' elevata di questi genitori ad essere impegnati socialmente*.

Key words: Disability in family, Italy, Parental effect, Spillover effets.

1 Introduction

According to the most recent Eurostat estimates (data from 2016), approximately 4% among individuals under age 16 have some disabilities and over 15 million

[1] Nicoletta Balbo, Bocconi University – Dondena Center. email: nicoletta.balbo@unibocconi.it
Danilo Bolano, Bocconi University – Dondena Center; email: danilo.bolano@unibocconi.it

school-age children have been identified as having special educational needs (estimates from the European Commission). These disabilities limit children in their everyday activities and impact the lives of their families in myriad ways. There is indeed some evidence suggesting that being confronted with the disability of a child may substantially affect the life and well-being of family members, shaping their social, demographic, and economic trajectories. Parents may experience negative physiological and psychological effects, such as higher levels of anxiety, stress, and uncertainty, and they may have less resources, time and energy to pursue employment [2]. A relevant weakness of the existing literature is that many studies rely on data from small convenience samples, often focusing on one specific type of child disability and rarely have sufficient data to consider other, crucial factors. All this makes any generalizations to the population level, as well as to a broader scientific setting, unfeasible.

This paper investigates the relationship between having a disabled child and parental health and social participation. We do so, by studying different health outcomes, namely self-rated health, and emotional well-being, to better understand the mechanisms through which child disability affects parents' health. Next to this, we also examine the unexplored association between child disability and parental civil engagement to uncover whether parents of disabled children show more pro-social and altruistic behaviours.
The proposed research contributes to the existing literature in several ways. For the first time, the analysis of how child disability can affect parental health and behaviours will be carried out within the Italian context. Italy is an extremely interesting test-bed because it is a country with a strong familialistic welfare regime, where the family is the main responsible for the care of children or non-independent/sick family members. Therefore, we can expect Italian parents to be especially affected by the disability of a child. Second, this paper adopts a population approach using an Italian nationally representative survey (Aspects of Daily Life Italy - AVQ). This will allow us to compare parents with a disabled child with their counterpart with a healthy child, and enable us to control for relevant socio-economic factors, that may act as confounders in the relationship under study. Third, we explore a novel and very much understudied life dimensions of parents' life, that is their social participation and civil engagement and how that is shaped by having a disabled child. Specifically, we aim at investigating whether parents with a disabled child are more likely to engage in associationism, volunteering and political activities, as a signal of higher altruism.

2 Background

The existing literature on families with a disabled child is scattered, with evidence based on small convenience samples, mostly focusing on a specific disease. The resulting evidentiary basis is inconclusive with respect to many outcomes (e.g., parents' fertility, union dissolution and mental health) and results are hardly

generalizable. A population approach is appropriate for adjudicating whether contradictory findings are driven by sampling biases, selection processes, or suggestive of legitimately differential effects by population subgroup.

2.1 Health

Studies of children's disability and family members' health show mixed findings. Most report adverse effects on parents' health [3, 9]. However, there is also evidence that mothers who care for a disabled child have even better health compared to mothers with a healthy child [6]. Along the same lines, other studies suggest that having a child with a disability is associated with a higher frailty in terms of mental health, and emotional stress for parents [1]; at the same time, a few studies highlight the ability of such families to adjust and develop coping strategies and resilience [4]. We therefore aim at studying how child disability influences parents' health considering different aspects of their well-being, to shed further light on which health dimensions are affected the most and how by the disability of a child. We specifically focus on self-rated health, which has been shown to be a good proxy also for physiological issues. We also examine the relationship between child disability and emotional well-being to uncover whether we observe higher distress deriving from the heightened care burden or, rather, we find evidence that the "disability paradox" [10] – according to which disabled individuals tend to perceive higher quality of life than non-disabled individuals because of their lower expectations and structural limitations - applies to the other family members as well.

2.2 Civil Engagement

The existing literature on disability and social participation focuses almost exclusively on disabled individuals themselves. Social participation in school activities, social and physical activities is associated with higher physical, emotional and social well-being for children with disabilities [18], but we know very little about how child disability affects social participation of parents. The evidence we have suggests that family members experience higher barriers to social participation [5] due to scarcity of time and resources but, conversely, have larger support networks. In this paper we aim at investigating potential positive effects of children disability on parents' civil engagement. The existing literature tends to focus on negative consequences and inequalities associated with children disability. However, the positive spillover effects also have a crucial societal relevance: these families create valuable social capital for civil society, broadly speaking. Some studies show that affected families develop higher levels of altruism and pro-social behaviours [7], but we do not know whether that translates into higher probability of engaging in volunteering activities, associationism, and political participation. In this paper we aim at addressing this issue empirically.

3 Data and Methods

For our analyses we use the nationally representative cross-sectional survey "Aspects of Daily Life - AVQ" administered every year since 1993 to around 25,000 households residing in Italy. AVQ covers a wide range of topics, namely schooling, employment, household characteristics, family life, health, political and social participation, and leisure time. In this survey each member living in the household is a respondent.

Independent variable. A measure of (children's) limitation in daily activities was introduced in 2013 using the Global Activity Limitation Indicator (GALI) based on the question: "*For at least the past six months, to what extent has your child been limited because of a health problem in activities that people usually do?*". That is the question we use to identify the presence of a disabled child in the household.

We pool together the last two available waves of AVQ, administered in 2018 and 2019, and we restrict our sample to mothers and fathers who have the oldest child aged 17 or younger. The total sample consists of 12,988 individuals, but due to the presence of missing observations in some of the variables included in the models, the working samples vary across models. While the prevalence of disabled children within the age group 0-17 in 2019 is equal to 5,04%, in our sample we reach a share of 6,17% children whose daily activities are limited or severely limited.

Outcome variables. As aforementioned, this study considered several health outcomes and social behaviour indicators. We considered the self-rated health (dichotomised in being in a poor health or not); emotional wellbeing, a 0-100 scale constructed from the corresponding five items of the Short Form SF-36 questionnaire [11]; and three indicators if the respondent was participating in any kind of associations, active in politics, and active in volunteering activities.

Control variables. We control for a series of socio-demographic characteristics that according to the literature might be associated with health condition and/or social engagement: gender, age, level of education (below high school diploma, diploma, degree or above), number of children in the household, employment condition (employed, unemployed, inactive), and region of residence (NUTS 1 level). Since we pooled together the last two waves of data, we include the year of interview as well.

To test our research hypotheses, a series of multivariate regression models is employed (results reported in Table 1). We run a logit model while analysing the SRH, and the three social and political engagement indicators. A linear regression model has been used when the outcome considered was emotional wellbeing. Since within a couple we might have a mutual influence across partners, the standard errors were clustered at household level.

4 Results and Conclusions

Table 1 shows the findings of our multivariate analyses. Results suggest that parents with a disabled child under the age of 17 are more likely to declare poor health both in terms of general health (self-rated health - Model 1) or mental health (emotional wellbeing - Model 5).
Model 2 to 4 instead show that parents with a disabled child are more likely to be engaged in associationism or volunteering, whereas no difference with their counterpart parents with a healthy child is observable in case of political participation.
In terms of the effect of our socio-demographic controls, the results are consistent with the literature. Women are less engaged and reported a lower level of health probably because of being primary caregiver of the disabled children. Older adults, those with lower levels of education and those unemployed are less healthy and less engaged. We observe as well interesting regional differences with respondents living in the South of Italy reporting lower levels of health and lower propensity to be socially and politically engaged.

Overall, our analyses suggest that parents of a disabled child are more likely to report poor health and lower subjective well-being, but at the same time they are more civically engaged. The present study looked at each outcome separately, but it is possible to look at the interdependence across them as well. These novel results point out that disability must be considered as a family level "condition" with strong effects that crossover family members, parents in this case, and spillover multiple life domains (wellbeing, social and civic engagement for instance). Further research can investigate whether there are relevant heterogeneous effects by mothers and fathers in these as well as other life dimensions. Similarly, studies on siblings or other individuals related to the disabled children (peers, grandparents to cite some) are welcome.

Of course, no causal inference can be made, because many unobservable factors may confound the relationships under study. Nevertheless, the provided evidence suggests that these families are frailer from an emotional and health-related point of view, but the societal contribution they provide in terms of creation of social capital is relevant. Therefore, by further supporting these parents, to improve their well-being, we might expect societies can further gain in terms of social participation and cohesion.

Table 1: Multivariate regression models

	Logit model – OR reported				OLS
	(1)	(2)	(3)	(4)	(5)
VARIABLES	Poor SRH	Active in association	Active in Politics	Active in volunteer.	Emotional wellbeing
Having a disabled child	**1.535****	**1.370*****	**1.210**	**1.283****	**-0.114*****
	(0.332)	**(0.148)**	**(0.142)**	**(0.139)**	**(0.04)**
Being a woman	1.142	0.890**	0.719***	0.863***	-0.178***
	(0.177)	(0.049)	(0.044)	(0.045)	(0.018)

Level of education (Ref. Below diploma)					
Diploma	0.663***	2.041***	1.558***	1.742***	0.066***
	(0.092)	(0.165)	(0.123)	(0.125)	(0.022)
Degree or above	0.401***	3.972***	2.052***	2.727***	0.120***
	(0.085)	(0.341)	(0.178)	(0.219)	(0.026)
Region of residency (Ref. North west)					
North east	0.951	1.589***	1.111	1.394***	0.035
	(0.181)	(0.128)	(0.103)	(0.110)	(0.028)
Center	0.851	0.947	1.055	0.824**	-0.006
	(0.167)	(0.087)	(0.103)	(0.073)	(0.030)
South	0.708*	0.762***	1.246**	0.626***	-0.007
	(0.131)	(0.068)	(0.116)	(0.054)	(0.028)
Islands	1.111	0.841	1.329**	0.632***	0.059
	(0.239)	(0.101)	(0.159)	(0.076)	(0.037)
Control variables	✓	✓	✓	✓	✓
Constant	0.013***	0.046***	0.053***	0.074***	0.062
	(0.004)	(0.006)	(0.008)	(0.01)	(0.041)
R-squared/Pseudo R2	0.057	0.061	0.049	0.043	0.018
Observations	12,988	12,988	12,988	12,988	12,988

Note: robust standard errors in parentheses. Clustered at household level. Control variables not included in the table for readability: age of the respondent, employment status, year of interview, total number of children.
*** p<0.01, ** p<0.05, * p<0.1

References

1. Boyd, B. A. Examining the relationship between stress and lack of social support in mothers of children with autism. Focus on autism and other developmental disabilities, 17(4), 208-215. (2002)
2. Busch, S. H., Barry, C. L. Mental health disorders in childhood: Assessing the burden on families. Health affairs, 26(4), 1088-1095. (2007)
3. Burton, P. Lethbridge, L.N., Phipps, S. Mothering Children with Disabilities and Chronic Conditions: Long-Term Implications for Self-Reported Health. Canadian Public Policy, 34(3), 359-378. (2008)
4. Mitchell, W. Research review: The role of grandparents in intergenerational support for families with disabled children: A review of the literature. Child & Family Social Work, 12(1), 94-101. (2207)
5. Murphy, N. A., Carbone, P. S. Promoting the participation of children with disabilities in sports, recreation, and physical activities. Pediatrics, 121(5), 1057-1061. (2008)
6. Olsson, M. B., Hwang, C. P. Well-being, involvement in paid work and division of child-care in parents of children with intellectual disabilities in Sweden. Journal of Intellectual Disability Research, 50(12), 963-969. (2006)
7. Orm, S., Haukeland, Y., Vatne, T., Silverman, W. K., Fjermestad, K. Prosocial behavior is a relative strength in siblings of children with physical disabilities or autism spectrum disorder. Journal of Developmental and Physical Disabilities, 1-18. (2021)
8. Rigles, B. The Development of Health Lifestyles in Families Experiencing Disability. Journal of Family Issues, 40(7), 929-953. (2019)
9. Seligman, M., Darling, R. B. Ordinary families, special children: A systems approach to childhood disability. Guilford Publications. (2017)
10. Walker, B. Resilience: what it *is* and is *not*. Ecology and Society 25(2), 11. (2020)
11. Ware, J.E., Kosinski, M., Keller, S.D.. SF-36 Physical and Mental Health Summary. Scales: A User's Manual. Boston, MA: Health Assessment Lab (1994)

Searching the nexus between women's empowerment and female genital cutting (FGC)

Potenziamento femminile e FGC: Quale relazione?

Patrizia Farina, Liva Ortensi, Thomas Pettinato, and Enrico Ripamonti[1]

Abstract Female genital mutilation/cutting (FGC) is still present in many African countries, as well as a few others. We used DHS data from seven African countries and considered both the empowerment and the FGC modules. We selected empowerment variables based on Kabeer's conceptual framework and used multilevel logistic models to evaluate the effect of empowerment in support for discontinuing the practice. We found a protective role of education. Other variables, including justification of intimate partner violence and having experienced FGC, were associated with FGC support. The relationship between decision-making and FGC support appears complex, while the unmet need for contraception and job conditions do not seem to play a role.

Abstract *La mutilazione genitale femminile (FGC) è ancora presente in molti paesi dell'Africa, così come in pochi altri. Abbiamo utilizzato i dati del DHS relativi a sette paesi africani e abbiamo considerato sia i moduli di potenziamento femminile che quelli di FGC. Abbiamo selezionato le variabili di potenziamento in base alla concettualizzazione proposta da Kabeer. Quindi, abbiamo utilizzato modelli logistici multilivello per valutare il possibile ruolo del potenziamento a sostegno dell'interruzione della pratica nella generazione successiva. I modelli multilivello hanno evidenziato l'effetto protettivo dell'istruzione. Altre variabili, tra cui la giustificazione della violenza da parte del partner e l'aver subito FGC, sono associate al supporto al FGC. La relazione tra il processo decisionale e il sostegno al FGC appare complessa, mentre, almeno dalle nostre analisi, il bisogno insoddisfatto di contraccezione e le condizioni di lavoro non sembrano svolgere un ruolo.*

Keywords: female genital cutting, empowerment, decision-making, education

[1] Patrizia Farina, Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy; email: patrizia.farina@unimib.it
Livia Ortensi, Department of Statistical Sciences "Paolo Fortunati", "Alma Mater Studiorum", University of Bologna, Bologna, Italy; email: livia.ortensi@unibo.it
Thomas Pettinato, Department of Sociology and Social Research, University of Milan-Bicocca, Milan, Italy; email: thomas.pettinato@unimib.it
Enrico Ripamonti, Department of Economics and Management, University of Brescia, Brescia, Italy, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: enrico.ripamonti@unibs.it

1. Introduction

Female Genital Cutting (FGC) is internationally recognised as an extreme violation of the rights of women and girls since these practices contravene the principle of equality and non-discrimination based on gender, as well as the right of not being exposed to torture or cruel, inhuman punishment [1–4]. FGC is normally performed over girls before the age of puberty, causing short- and long-term health complications, including infections, increased risk of HIV transmission, chronic pain, birth complications, infertility, and, in worst cases, death [5–7]. Four main types of FGC have been described, ranging from total removal of the clitoris to infibulation, and including other types of modification like stretching, cauterisation and piercing. FGC is currently quite prevalent in Africa and the Middle East, spanning at least 31 countries and affecting over 200 million girls and women (UN estimates). Decades of actions of International Agencies, governments, civil society, communities, and individuals accelerated the secular decline of FGC. Herein, we aim to assess the relevance of women empowerment on individual support to FGC continuation. In particular, we targeted the putative protective effect of mothers' empowerment on the next generation of girls, controlling for the background and the socio-economic conditions of adult women.

2. Data and methods

We selected seven African countries in which information on FGC, gender violence and empowerment modules were simultaneously recently collected: Burkina Faso, Ivory Coast, Egypt, Ethiopia, Mali, Nigeria and Togo. We used data from the Demographic and Health Surveys (DHS) carried out between 2008 and 2016 [8]. The binary dependent variable is represented by the women's support to the continuation of the practice. Covariates include an indicator of empowerment, women's age, FGC personal experience (being cut or not), the wealth index, unmet need for contraception, employment status, educational level, children ever born, religion, and FGC regional prevalence. Operationalisation of women's empowerment has been done following Kabeer's [9,10] theoretical framework. To carry out the analysis, we fitted multilevel logistic models.

3. Results

We found that the education level was negatively correlated with FGC support, along with age at first cohabitation. Experience of FGC, on the contrary, was strongly correlated with support for the continuation of FGC. The ideal number of children, the justification of intimate partner violence (IPV), and whether decision-making was mainly left to the respondent or other persons (compared with equal

decision-making within the couple) were associated with more substantial FGC support. Interestingly, the unmet need for contraception and job conditions were not significantly correlated with FGC support.

Predicted probabilities based on the model (Figure 1) confirmed the relevance of the link between indicators of empowerment and FGC. Support for continuing FGC decreases when the educational level increases, suggesting that the higher a woman's education level, the easier is for her to make empowered choices. It is worth observing that predicted probabilities seem to be similar for three clusters: {Burkina Faso}, {Egypt, Mali, Nigeria}, {Ivory Coast, Ethiopia, Togo}. This may indicate underlying similar cultural contexts that have not been investigated in the present context.

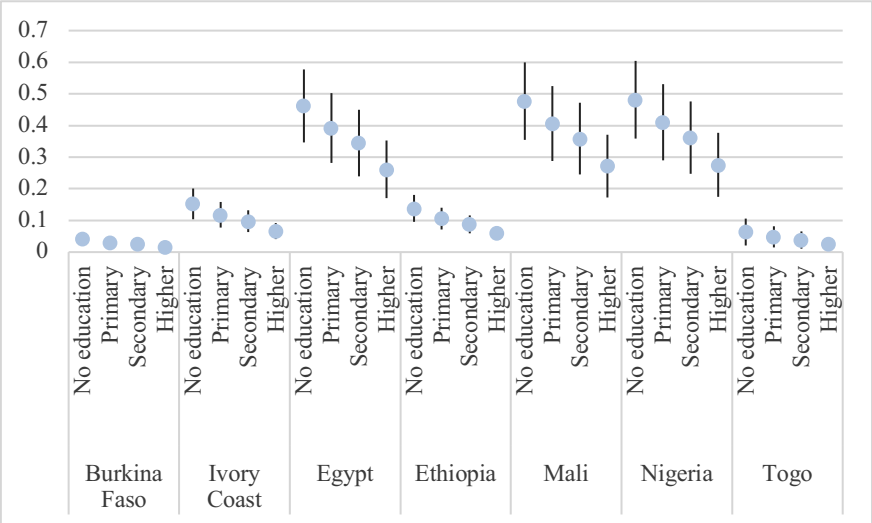


Figure 1. Predicted probabilities by country and education (other variables kept at their average level)

4. Final remarks

Our analyses underline the importance of linking FGC support and empowerment. In this regard, we found a role played by education did emerge. Other relevant factors were personal experience with FGC, and justification of IPV. Such dimensions describe different levels of empowerment, reflecting existing social and gender norms that act upon resources and agency and are internalized by individual women (e.g., IPV justification).

In the measurement of empowerment what has come to be more influential is the dimension of agreement among women on the subordinate role they have in the couple. This is explicitly linked to the legitimization of men's violence. Being it an important empowerment component, policy actions to enforce empowerment must

go firstly in the direction of changing this legitimization, breaking the crystallized unbalance of gender gap shared by women, erasing the role of a punitive husband and a "disrespectful" wife who fully aim to be able and entitled to truly make decisions [11] with no fear of violent consequences.

As from the international organization literature, policies must go not only in the direction of human rights enforcement but also and most notably in the elimination of gender discrimination (as from WHO [4]). As such, gender role changes reflect the definition of empowerment as reported in Kabeer [9,10].

Specific cultural characteristics of different contexts must be taken into consideration to understand the mechanisms operating in one defined territory. It is a limitation of the present analysis both to have a definition of empowerment linked to the available data and to reflect on applying a standardized definition to different cultures as well social contexts. Thus, the influence of empowerment on FGC should be further studied, as well as put in relation to other regional and local contexts.

References

1. UNICEF. Female Genital Mutilation/Cutting: A statistical overview and exploration of the dynamics of change. Reprod Health Matters. 2013;184–90.
2. UNICEF. Joint program on the elimination of female genital mutilation. Annual report 2018 [Internet]. 2019. Available from: https://www.unfpa.org/publications/accelerating-change
3. WHO. Violence prevention: The evidence [Internet]. World Health Organization; 2010. Available from: https://apps.who.int/iris/bitstream/handle/10665/77936/9789241500845_eng.pdf
4. WHO. Elimination of female genital mutilation. An interagency statement. 2008.
5. Berg RC, Denison E. Does female genital mutilation/cutting (FGM/C) affect women's sexual functioning? A systematic review of the sexual consequences of FGM/C. Sex Res Soc Policy. 2012;9(1):41–56.
6. Berg RC, Underland V. The obstetric consequences of female genital mutilation/cutting: a systematic review and meta-analysis. Obstet Gynecol Int. 2013;496564.
7. Berg RC, Underland V, Odgaard-Jensen J, Fretheim A, Vist GE. Effects of female genital cutting on physical health outcomes: a systematic review and meta-analysis. BMJ Open. 2014;4(11).
8. DHS. DHS Overview [Internet]. Online resource. [cited 2020 Jan 24]. Available from: https://dhsprogram.com/what-we-do/survey-Types/dHs.cfm
9. Kabeer N. Resources, agency, achievements: Reflections on the measurement of women's empowerment. Dev Change. 1999;30(3):435–64.
10. Kabeer N. Contextualising the economic pathways of women's empowerment: findings from a multi-country research programme. Brighton: Pathways of Women's Empowerment; 2011.
11. Ewerling F, Lynch JW, Victora CG, van Eerdewijk A, Tyszler M, Barros AJD. The SWPER index for women's empowerment in Africa: development and validation of an index based on survey data. Lancet Glob Heal. 2017;5(9):e916--e923.

Social stratification, gender, and attitudes towards voluntary childlessness in Europe: A double machine learning approach

Stratificazione sociale, genere, e atteggiamenti verso la scelta di essere senza figli: un approccio double machine learning

Danilo Bolano and Francesco C. Billari

Abstract Childnessness, and its links with educational change and inequality, is emerging as a key issue in the study of population change. In this paper we study the intersection between social stratification and gender in shaping the gendered attitudes towards voluntary childlessness in Europe, using data from two rounds of the European Social Survey (ESS). The ESS includes a randomised design that allows to derive experimental results on gender attitudes. We apply a double machine learning approach and find that, in general, higher education has an impact on more favourable attitudes. Our findings are consistent with educational expansion having a causal role in the increase of favourable attitudes towards childlessness in Europe.
Abstract *La scelta di essere senza figli è una decisione sempre piu' comune nelle società occidentali seppure ancora osteggiata in quanto non conforme al concetto che uomo/donna adulto è da considerasi tale se genitore. Studi sul legame tra una tale scelta e stratificazione sociale sono quindi fondamentali per comprendere l'evoluzione della struttura della popolazione. Questo studio si focalizza sull'interesezionalità tra stratificazione sociale e genere nel definire gli atteggiamente verso la scelta di essere senza figli in Europa. Utilizzando un approccio causale di tipo double machine learning su dati dell'European Social Survey, questo studio dimostra come l'espansione del livello di istruzione abbia reso piu' favorevole il giudizio circa la scelta di un individuo di non avere figli.*

Key words: Social Norms, Social Stratification, Lasso, Average Treatment Effect, Gender

Danilo Bolano
Dondena Centre for Research on Social Dynamics and Public Policy, and Department of Social and Political Sciences, Bocconi University, e-mail: danilo.bolano@unibocconi.it

Francesco C. Billari
Dondena Centre for Research on Social Dynamics and Public Policy, and Department of Social and Political Sciences, e-mail: francesco.billari@unibocconi.it

1 Introduction

Europe has experienced, over time, a U-shaped trend in childlessness over time, with data pointing to a general rise in the recent decades both from a period and a cohort perspective [8]. This shift is linked to a higher acceptance of voluntary childlessness, which for a long time, and in many parts of Europe, was running against established social norms. In line with the rise in its prevalence, the acceptance of childlessness has generally risen over the last decades [7].

In the recent literature, gender and social stratification have figured prominently as key forces that shape choices. It has, for instance, been argued that "Future fertility trends are shaped at the intersection of gender and social stratification" [6]. We focus on this issue, and study the intersection of gender and social stratification on attitudes towards voluntary childlessness. We do so by exploiting, on the gender side, an experimental design that was implemented in two waves of the European Social Survey (ESS), and by applying, on the social stratification side, a double machine learning approach.

2 Data and survey design

We exploit data from two rounds of the European Social Survey: round 3 (2005–06), and 9 (2017–18). The ESS is a biennial, cross-sectional, multi-country survey that collects information on attitudes and behavior in almost 30 European countries. It is a representative survey of all persons aged 15 and over living in each country. In rounds three and nine, a specific rotating module on the "Timing of life", was included [3]. Importantly, the module included questions on attitudes towards voluntary childlessness, and a split-ballot design that allowed to randomly ask questions about women or about men. For our analyses, we pool together the two rounds, ending up with almost 80,000 respondents across Europe. Here we shortly describe the main measures used in our analyses.

Outcome variable: Attitude towards voluntary childlessness. We use the question "How much do you approve or disapprove if a woman/man chooses never to have children?", with answers on a scale from 1 (strongly disapprove) to 5 (strongly approve). Answers are standardised for ease of interpretation.

Randomized design on gender. As we mentioned, to (about) half of the respondents were asked about male behavior (i.e., how much the respondent approve or disapprove if a man chooses never to have a children), and to the remaining half were asked about female behavior (i.e., how much the respondent approve or disapprove if a woman chooses never to have a children). This split-ballot design on the gender of the "target" allows to integrate the advantage of having a (randomized) experimental design (internal validity) and information from a nationally representative survey (external validity) [1]. We create a treatment variable with four values, interacting the gender of the respondent and the gender of the target.

Social stratification: educational level of the respondent. As a social stratification factor, we include a dummy variable indicating whether the respondent has completed at least secondary education (ISCED 3 or above, which we code as high level of education), or not.

Control variables. We include a series of variables which, according to the literature, might be associated with either respondent's educational attainment or attitudes toward childlessness. In particular, age, marital status, current employment status, having children, level of religiosity, being native or migrant, citizenship, educational level of both parents, if the parents were born in the same country the respondent currently live in. Finally we consider the year of interview and country of living. For parsimony, we cluster countries considering family policies and social values simultaneously as follows: egalitarian (Sweden, the Netherlands, Denmark, United Kingdom, Norway, Finland), pro-natalist (France and Belgium), Eastern European (Bulgaria, Czech Republic, Poland, Hungary, Slovenia, Estonia, Croatia, Romania, Latvia, Russian Federation, Slovakia), traditional (Austria, Germany, Switzerland, Ireland), and Southern European countries (Spain, Italy, Greece, Portugal, Cyprus).

3 Methods

While the gender of the respondent is exogenously given, and the gender of the target has been randomised, in order to estimate the causal impact of education on attitudes towards voluntary childlessness, we need a specific empirical strategy. We here apply a strategy that makes an optimal use of the set of available control variables, i.e. an Average Treatment Effect (ATE) model with lasso.

More specifically, being y_i the outcome variable (the approval of childlessness), and T the treatment binary variable (having a high level of education or not), and $\mathbf{X}$ a vector of controls, broadly speaking the outcome model can we written as $y_i = g_0(T_i, \mathbf{X}_i) + u_i$ with $E(u_i|T_i,_i) = 0$.

Where $g_0(T_i = 0, \mathbf{x}_i) + u_i$ is then the expected outcome when the treatment T takes value 0 (when the respondent is low educated in our study). Conversely $g_0(T_i = 1, \mathbf{x}_i) + u_i$ when the respondent is highly educated). Of course, we can only observe one of the two expected outcomes. The treatment model can be written as $T_i = m_0(\mathbf{Z}_i) + v_i$ with $E(v_i|\mathbf{z}_i) = 0$.

where $\mathbf{Z}$ is another vector of controls and $m_0(\mathbf{Z}_i)$ is the probability of being treated given the vector of controls Z. It is possible to use different controls for the treatment model ($\mathbf{Z}$) and for the outcome model ($\mathbf{X}$). The objective is to estimate the ATE: $E(g_0(T = 1, \mathbf{x}_i) - g_0(T = 0, \mathbf{x}_i))$

Given the nature of our outcome variable (standardised count variable), and of our treatment (dummy variable), we estimate the $g_0(\cdot)$ using a linear model and $m_0(\cdot)$ using a logistic model.

Crucially, instead of pre-selecting the controls to be included in the model, we use lasso techniques within an ATE approach [5]. This strategy allows to estimate $g_0(1, \mathbf{x_i})$, $g_0(0, \mathbf{x_i})$ and $m_0(0, \mathbf{z_i})$, selecting only relevant controls over a potential

large set of covariates. More specifically, this strategy allows at the same time to include several different variables to potentially satisfy the conditional independence assumption (i.e., that that given the controls, the outcome is independent of the treatment), while reducing the risk of violating the overlap assumption (i.e., that given the controls there must be always a positive probability of each unit of being not treated). The estimated effect is then robust to model selection, and as well as (doubly) robust to functional form miss-specification either in the outcome or the treatment model. Moreover, we report the estimates from a "double machine learning estimator" [5]. Using both cross-fitting and re-sampling, the double machine learning estimator relaxes the sparsity assumption allowing to include (i.e., to control for) more confounders than a traditional lasso approach. The estimated ATEs are then an average across the different cross-folds and resamples.

Leveraging the randomised design of ESS on questions about attitude, we estimate ATEs separately by gender of respondent and gender of the target.

4 Results

Attitudes toward childlessness have become more favorable over time. The level of approval is increased from 26.6% of approval (approve and strongly approve) in 2005 (ESS round 3) to 36.09% in 2017 (ESS round 9). Simply controlling for ESS round and country of residence, we observe a clear gender difference in approval. The highest predicted approval is observed among women when the question is about a woman, the lowest among women about a man. We now report our estimates separately for each of the four groups.

The ATE approach allows to have a separate set of controls for the treatment (educational attainment) model and for the outcome (approval). Potentially confounders of childlessness approval are age of respondent (5 levels), the marital status (3 levels), level of religiosity (4 levels), whether the respondent had a child, country group (4 levels) and year of interview. Level of education (treatment model) is instead expected to be influenced by the age of respondent (that in our case being a cross-sectional analysis also corresponds to cohort), whether the respondent is born/citizen of that country, as well as characteristics of family of origin (level of education of both parents, whether the parents were born in the same country of the respondent). Also for the treatment model we include country group and and year of interview as fixed effects. Using lasso for model selection in both treatment and outcome model allows us to include more controls than in a traditional ATE approach. We include then also interaction effects among each of the variables aforementioned and country group. As conventional, we use one-hot encoding for categorical variables. We end up then with more than 200 controls. In order to obtain robust estimates, we cluster the standard error by country of residence (28 clusters) and use a double machine learning approach [5] with a 5-fold cross-validation over three split samples.

Comparing the results obtained with and without double machine learning (i.e., applying lasso within an ATE estimate, but without cross-validation and re-sampling, results not shown), the number of controls selected by lasso almost doubled while applying the double machine approach, bringing to potentially more robust results.

We found (Table1) that being highly educated has a significantly positive effect on level of childlessness approval of about 10% of a standard deviation (ATE of .096 on the standardised outcome). These effects are heterogeneous by gender and target. While education does not have a significant impact on childlessness men when the respondent is a man, a positive effect of education on approval is observed in all other cases, with the highest value when to a woman is asked about the approval of be a childless women (increase of 11.5% of a standard deviation).

In terms of controls selected by lasso, and using as an example the last model (when women reply on women behavior)[1], lasso for the treatment model selected the age of respondent, and the level of education of the mother and the father, conventional indicators of parental socio-economic status. The country of origin seems instead is not selected. In the outcome model, lasso selected the age of respondent and religiosity in particular for those living either in an "egalitarian" country, or in an Eastern European country.

Table 1 ATE of having high level of education. Estimate with lasso double machine learning

Model	Effect	p-value	95% CI	N of controls retained
Overall sample[a]	.096 (.034)	.005	[.029, .163]	54
Male respondent on men's behavior	.036 (.042)	.391	[-.047, .119]	56
Male respondent on women's behavior	.111 (.041)	.007	[.0298, .192]	52
Female respondent on men's behavior	.104 (.037)	.004	[.032, .177]	48
Female respondent on women's behavior	.115 (.041)	.004	[.036, .1947]	58

Note: Robust standard errors reported in parentheses clustered by country. Outcome is the standardised level of childlessness approval.

[a] We control also for gender of respondent and gender of target in the outcome model; and we control for respondent's gender in the treatment model.

5 Discussion

Leveraging the experimental design of the module Timing of the Life of the European Social Survey, we investigated the combining causal effect of gender and social

[1] The results refer to the first cross-fold and first re-sample.

stratification on the approval of voluntary childlessness in Europe. Overall, our results are consistent with the idea that achieving a high level of education *causes* a more positive attitude towards childlessness, with the only exception of men answering about approval if a man chooses never to have children. Moving from the individual to the population level, our findings are therefore consistent with the hypothesis that educational expansion played a specific causal role in decreasing the attitudes towards voluntary childbearing in Europe.

This work can be extended in several directions. Firstly, in terms of the effect of the institutional setting. The lasso variable selection model suggests that country typology matters in explaining variations in attitudes towards childlessness. Secondly, further analyses might shed light at the role of social origin and parental background in accepting behavior that is not "traditional", as the decision of not having children over the life course.

The general contribution of this work is twofold. On demography, it provides robust evidence on a large number of European countries and individuals of the role of the intersection of gender and social stratification in potential explanations of childlessness. From a methods point of view, it contributes to a growing literature that uses machine learning approaches to answer questions related to demography and life-course trajectories [2, 4].

Acknowledgements The authors gratefully acknowledge financial support from EU H2020 project "DisCont - Discontinuities in Household and Family Formation" (PI: Billari – ERC Grant no. 694262).

References

1. Arpino, B. Matching pre-processing of split-ballot survey data for the analysis of double standards. Soc. Psychol. Q. **79**(4), 397-407 (2016)
2. Arpino B, Le Moglie M, and Mencarini L. What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany. Demography **59**(1), 161-186 (2022).
3. Billari, F. C., Badolato, L. Hagestad, G., Liefbroer, A.C., Settersten Jr, R. and Speder, Z. and Van Bavel, J. The Timing of Life: Topline results from Round 9 of the European Social Survey. ESS Topline Results Series **11** (2021).
4. Bolano, D., and Studer, M. The link between previous life trajectories and a later life outcome: A feature selection approach. LIVES Working Paper (2020) doi: 10.12682/lives.2296-1658.2020.82
5. Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal **21**(1), C1–C68 (2018).
6. Lappegård, T. Future fertility trends are shaped at the intersection of gender and social stratification. Vienna Yearbook of Population Research **18**:43-48 (2020).
7. Merz, E.-M. and Liefbroer, A.C., The Attitude Toward Voluntary Childlessness in Europe: Cultural and Institutional Explanations. Journal of Marriage and Family, **74**: 587-600 (2012)
8. Sobotka, T. Childlessness in Europe: Reconstructing long-term trends among women born in 1900–1972. In Childlessness in Europe: Contexts, causes, and consequences, pp. 17-53. Springer (2017).

Integrating structuralism and diffusionism to explain the new Italian emigration

Interpretazioni strutturaliste e diffusioniste della nuova emigrazione italiana

Francesca Bitonti

Abstract Diffusionism gained momentum as a viable approach to uncover the dynamics underlying the decline in marital fertility that occurred across Europe during the past century. This theory challenges the traditional adaptative explanations to fertility transition posed by the structural approach. This work aims at borrowing and integrating diffusionist and structuralist interpretations of fertility transition to give meaningful insights into the recent new wave of Italian emigration. An extended version of the Bass model, extensively employed to study the diffusion of new products and behaviours across a population, is proposed. The preliminary results highlight the underlying dynamics governing the new Italian emigration flow and provide a novel interpretation of a recent international emigration process.

Abstract *Il diffusionismo è emerso in demografia nello studio delle dinamiche alla base del declino della fecondità avvenuto in Europa nel secolo scorso. Questa teoria sfida le tradizionali spiegazioni "adattative" del crollo della fecondità proposte dall'approccio strutturale. L'obiettivo di questo lavoro è quello di prendere in prestito e integrare diffusionismo e strutturalismo dagli studi sulla transizione della fecondità per dare una spiegazione significativa alla recente nuova emigrazione italiana. Si propone una versione estesa del modello Bass, ampiamente utilizzato per studiare la diffusione di nuovi prodotti e comportamenti in una popolazione. I risultati preliminari evidenziano le dinamiche e forniscono una nuova interpretazione del recente processo di emigrazione internazionale italiana.*

Key words: diffusionism, international migration, Bass diffusion model, new Italian emigration

---

[1] Francesca Bitonti, Department of Economics and Business, University of Catania, Corso Italia 55, 95129; email: francesca.bitonti@phd.unict.it

1 Introduction

The diffusionist theory finds its roots in the field of social sciences to explain the spread of new ideas and practices among the members of a population [26,29]. The ground idea of diffusionism is that the interplay of social influence mechanisms, as interpersonal information exchange, social norms, and emulative processes, shape individual conduct. The diffusionist approach has found wide application also in demographic studies. From a demographic standpoint, an innovation may be intended as any idea or practice having implications on the population dynamics, namely fertility, mortality, and migrations [35]. In this sense, examples are contraceptive adoption, childbearing postponements, smoking behaviour, novel migration practices, and the like. Diffusionism gained momentum with the necessity to give a plausible explanation to the decline in marital fertility that occurred across Europe during the past century. This approach challenged the classical demographic transition theory, which argued that modernisation and structural changes in society were the main driving forces for fertility variations. Roughly speaking, the rising cost and declining economic value of children is the key factor leading to fertility change [5,7]. The diffusionist considerations instead argued that new ideas and norms about fertility have diffused through social interaction modifying the macro-level fertility dynamics [10]. In the 80s, the Princeton European Fertility Project and the World Fertility Survey were the two seminal works suggesting that the fertility decline could have been influenced by the diffusion of modern contraceptive practices, new values, and new ideas of family among people [11,12,33]. After this introduction in the demographic debate, the diffusionist vision has been the theoretical background for several studies of fertility choices and family planning, and for the promotion of public health campaigns aimed at contrasting the spread of diseases and health-damaging behaviours [1,2,3,6,20,24,25,31,34]. Recently, modernisation and diffusionist arguments have no longer been considered distinct or in opposition. Rather, they are conceptualised as strictly embedded explanations to demographic changes [27]. In adherence to this strand of thought, the objective of this work is to analyse and convey meaningful insights of international migration processes in terms of both diffusionist and structuralist approach. In particular, the analysis focuses on a novel emigration flow involving Italy in recent times. The work is organised as follows: the next section illustrates the so-called new Italian emigration; the third section introduces the methodology implemented and describes the data analysed; the fourth one discusses the main results of our applications, while the last reports the conclusions of the work.

2 The new Italian emigration

In recent years, Italian emigration had considerably grown compared to the 80s and 90s, when the phenomenon was ongoing but at a smaller scale. The current

emigration presents specific peculiarities, configuring a situation different from the past, which in literature is known as “the new Italian emigration”.

Among the novelties, an increased share of women and well-educated individuals leaving the country has been recorded [13,15,30]. Furthermore, although the Southern regions continue, as in the past, to contribute to the outgoing movements, recently, the more advanced northern Italy has become the main outflow area [8,30]. Another relevant feature is that the destinations preferred by Italian migrants mainly converge to the EU and EFTA countries [8]. Finally, the new Italian emigration appears to be paired with a change in attitude towards international migration. The globalisation of the world and the availability of new communication technologies could have overcome several of the barriers hindering movements in the past, allowing Italians, especially the youngsters, to approach international mobility in a novel and more confident way [32].

Overall, the interaction among the global economic downturn, the EU integration process [14,21,23], and the change in attitude towards outward mobility appear to have shaped a new migration behaviour that has diffused across Italy. The novel aspects characterising the Italian emigration configure an innovative migrant behaviour that diffuses among the population through particular communication channels and according to established social dynamics, but also in accordance with the economic cycle of origin and destination countries.

3 Methods and data

The Bass model [4] is extensively employed in market research studies to analyse the diffusion process of new ideas, products, and behaviours (called “innovations”) in a given population. According to the model, the pace of diffusion is driven by two forms of communication: the *external* one, such as mass media advertisement or awareness and prevention health campaigns, and the *internal* one, including social influence, word of mouth, and imitation. These two drivers of diffusion shape two categories of adopters: the innovators, mainly influenced by the external source, and the imitators, who adopt in response to their interaction with prior adopters. The Bass model consists of a simple first-order differential equation

$$Y'(t) = \left(p + \frac{q}{m}Y(t)\right)\left(m - Y(t)\right) = p\left(m - Y(t)\right) + q\frac{Y(t)}{m}\left(m - Y(t)\right) \quad (1)$$

where the variation of adoption $Y'(t)$ over time is proportional to the residual susceptible population $(m - Y(t))$, with m being the constant overall susceptible population, and $Y(t)$ the cumulative adoptions at time t. Rearranging the left-hand side equation (1) it is possible to notice that the instantaneous adoptions $Y'(t)$ is the result of the sum between two components, the external one governed by parameter p and the internal one modulated by q. Parameter p is the so-called coefficient of innovation, representing the effect of the external influence to adopt. Parameter q is the coefficient of imitation and reflects the inter-personal influence individuals can exert on each other. The basic Bass model and its extensions have found empirical

applications not only in marketing studies but also in the demographic field: e.g. to analyse the diffusion of oral contraception [28], to clarify the dynamics in vaccination propensity and address public health policy [19,22], and to study the diffusion of disease-related information during an epidemic outbreak [18]. In the present work, we propose an extended version of the traditional Bass model to capture both structural and diffusionist explanations to the new Italian emigration:

$$\frac{dY_{it}}{dt} = p_{it}(m - Y_{it}) + \frac{q_{it}}{m} Y_{it}(m - Y_{it}) \quad (2)$$

where the coefficients p and q depend on the structural characteristics of the destination countries the emigration flow is directed to:

$$p_{it} = \alpha_0 + \alpha_1 U_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + \epsilon_{it} \quad (3)$$

$$q_{it} = \beta_0 + \beta_1 U_{it} + \delta_2 D2_i + \delta_3 D3_i + \cdots + \delta_n Dn_i + \epsilon_{it} \quad (4)$$

$i=1,\ldots,n$ are the destination countries at each time interval t, U_{it} is the unemployment rate for destination country i at time t, $D2_i + \cdots + Dn_i$ are $n-1$ dummy variables identifying the $i=2,\ldots,n$ destination countries respectively, lastly the erratic component is defined as $\epsilon_{it} \sim N(0, \sigma^2)$. The extended version of the Bass model has been fitted using an ad-hoc numerical optimization algorithm minimizing the root relative squared errors (RRSE) in the R software, which automatically estimated all the model parameters. The main advantage of the proposed model is to parsimoniously provide the time-varying parameters p and q for each destination country considered. With the same interpretative rather than forecasting objective found in literature [9,16,17], this work intends to offer a combined diffusionist and structuralist perspective of the new Italian emigration process. The basic reasoning is that the Italian crisis, combined with the spread of mass media information about the more dynamic labour market and the better living conditions in some EU countries, might have acted as external drivers of influence. On the other end, the information conveyed by actual emigrants about the possibility to be better-off could have triggered word-of-mouth effects and emulative behaviours.

Annual counts of Italian citizens who emigrated from 2002 to 2018 towards the main EU destination countries (Austria, Belgium, France, Germany, Spain, and United Kingdom – pre-Brexit) were retrieved from the Italian National Institute of Statistics (Istat) repository. Data on unemployment rates come from the World Bank online database.

4 Results and conclusions

The results of the extended Bass model fit are displayed in Figure 1. During the period 2002-2018, the outgoing flows (black dots) have increased for each destination country, with Germany, France, and United Kingdom accounting for the highest portion of movements. The global model (black lines) shows, in general, a good fit to the real data, loosing goodness of fit when the data present

discontinuities. The external component (light grey dashed lines) accounts for a little proportion of emigrations, while the internal one (dark grey dashed lines) is the main driver of the outgoing trends. Progresses in the EU integration and the 2007 financial crisis have accelerated the spread of novel migratory behaviours in Italy, characterized by an increased attitude towards international mobility and imitation. We proposed an interpretation of the new trends in Italian out-migration flows combining diffusionist and structuralist reasoning in an ad-hoc methodological framework. The proposal is that of a "toy model", namely a simplistic model aimed to provide interpretation to complex real-world phenomena; a typical example is the Lotka–Volterra model in population ecology.

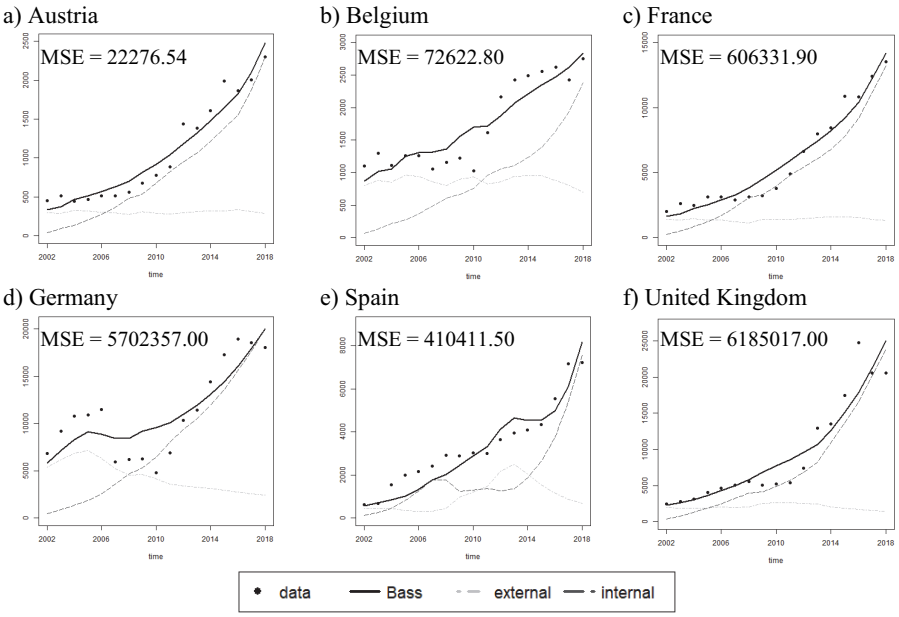


Figure 1: Extended Bass model fitting to the Italian annual emigrations toward the main EU destination countries. Time period: 2002-2018 (different y-axis scales).

References

1. Abraham, A.J., Roman, P.M.: Early adoption of injectable naltrexone for alcohol-use disorders: Findings in the private-treatment sector. J. of Stud. Alcohol Drugs, 71, 3, 460--466 (2010)
2. Alvergne, A., Gibson, M.A., Gurmu, E., Mace, R.: Social Transmission and the Spread of Modern Contraception in Rural Ethiopia. PLoS ONE, 6, 7, 22515 (2011)
3. Backer, T.E., Rogers, E.M.: Diffusion of innovations theory and work-site aids programs. J. Health Commun., 3, 1, 17--28 (1998)
4. Bass, F.M.: A New Product Growth for Model Consumer Durables. Manage. Sci., 15, 5, 215--227 (1969)
5. Becker, G.S.: A treatise on the family. Harvard University Press (1993)

6. Bengtsson, T., Dribe, M.: The historical fertility transition at the micro level: Southern Sweden 1815--1939. Demogr. Res., 30, 17, 493--534 (2014)
7. Bongaarts, J., Watkins, S.C.: Social interactions and contemporary fertility transitions. Popul. Dev. Rev., 639--682 (1996)
8. Bonifazi, C.: Da dove si parte, dove si va. Il Mulino, 6, 49--56 (2018)
9. Bunea, A. M., Della Posta, P., Guidolin, M., Manfredi, P.: What do adoption patterns of solar panels observed so far tell about governments' incentive? Insights from diffusion models. Technol. Forecast. Soc. Chang., 120240 (2020)
10. Carlsson, G.: The Decline of Fertility: Innovation or Adjustment Process. Pop. Stud., 20, 2, 149--174 (1966)
11. Cleland, J.: Potatoes and Pills: An Overview of Innovation Diffusion Contributions to Explanations of Fertility Decline. In: Casterline, J.B. (ed.), Diffusion processes and fertility transition: Selected perspectives. Washington D.C.: National Academy Press (US), 39--65 (2001)
12. Coale, A.J., Watkins, S.C. (eds.): The Decline of Fertility in Europe: the Revised Proceedings of a Conference on the Princeton European Fertility Project. Princeton University Press (1986)
13. Colucci, M.: Formazione e reclutamento degli italiani che emigrano. Il Mulino, 6, 33--40 (2018)
14. De Rose, A., Strozza, S. (eds.): Rapporto sulla popolazione. L'Italia nella crisi economica. Il Mulino (2015)
15. Fondazione Migrantes: Rapporto Italiani nel Mondo 2020 (2020)
16. Furlan, C., Guidolin, M., Guseo, R.: Has the Fukushima accident influenced short-term consumption in the evolution of nuclear energy? An analysis of the world and seven leading countries. Technol. Forecast. Soc. Chang., 107, 37--49 (2016)
17. Guidolin, M., Mortarino, C.: Cross-country diffusion of photovoltaic systems: Modelling choices and forecasts for national adoption patterns. Technol. Forecast. Soc. Chang., 77, 2, 279--296 (2010)
18. Gündüç, S.: A Study on the Effects of Diffusion of Information on Epidemic Spread. Int. J. Model., Simul., Sci. Comp., 10, 3, 1950015 (2019)
19. Kahana, D., Yamin, D.: Accounting for the spread of vaccination behavior to optimize influenza vaccination programs. PLoS ONE, 16, 6, e0252510 (2021)
20. Lesthaeghe, R.: Long-Term Spatial Continuities in Demographic Innovations: Insights from the Belgian Example, 1846--2000. Population Studies Center Research Report, 10--695 (2010)
21. Livi Bacci, M.: Italiani verso la Gran Bretagna. Neodemos (2014)
22. Onofrio, A., Manfredi, P., Poletti, P.: The Interplay of Public Intervention and Private Choices in Determining the Outcome of Vaccination Programmes. PLoS ONE, 7, 10, 45653 (2012)
23. Pugliese, E.: Tutto il mondo è paese: la nuova emigrazione italiana. Il Mulino, 6, 8--22 (2018)
24. Ramseyer Winter, V.: Diffusion of Innovations Theory: A Unifying Framework for HIV Peer Education. Am. J. Sex. Educ., 8, 4, 228--245 (2013)
25. Rogers, E.M.: Diffusion of drug abuse prevention programs: Spontaneous diffusion, agenda setting, and reinvention. NIDA Res. Mg., 155, 90--105 (1995)
26. Rogers, E.M.: Diffusion of innovations. New York: The Free Press (1962)
27. Rosero-Bixby, L., Casterline, J.B.: Modelling Diffusion Effects in Fertility Transition. Popul. Stud., 47, 1, 147--167, DOI: 10.1080/0032472031000146786 (1993)
28. Sharif, M.N., Ramanathan, K.: Binomial innovation diffusion models with dynamic potential adopter population. Technol. Forecast. Soc. Chang., 20, 1, 63--87 (1981)
29. Strang, D., Meyer, J.W.: Institutional Conditions for Diffusion. Theor. Soc., 22, 4, 487--511 (1993)
30. Strozza, S., Tucci, E.: I nuovi caratteri dell'emigrazione italiana. Il Mulino, 6, 41--48 (2018)
31. Svenkerud, P.J., Singhal, A., Papa, M.J.: Diffusion of innovations theory and effective targeting of HIV/AIDS programmes in Thailand. Asian J. Commun., 8, 1, 1--30 (1998)
32. Tirabassi, M.: Migranti da sempre. Il Mulino, 6, 24--32 (2018)
33. Tolnay, S.E.: The spatial diffusion of fertility: a cross-sectional analysis of counties in the American South, 1940. Am. Sociol. Rev., 60, 2, 299--308 (1995)
34. Vitali, A., Aassve, A., Lappegård, T.: Diffusion of Childbearing Within Cohabitation. Demogr., 52, 355--377 (2015)
35. Vitali, A., Billari, F.C.: Changing Determinants of Low Fertility and Diffusion: a Spatial Analysis for Italy. Popul., Space Place, 23, 2 (2017)

On the effects of rooted family ties in business networks: The South of Italy in the 19^{th} century

Un'analisi sugli effetti dei legami familiari radicati nelle reti di affari: il Mezzogiorno nel XIX secolo

Roberto Rondinelli, Giancarlo Ragozini and Maria Carmela Schisani

Abstract In the economics literature, rooted family ties, mainly associated to backward areas, are related to factors disruptive to economic growth. We study the effects of bonding family ties (associated to family-firms) in the entrepreneurial environment of the 19^{th} century South of Italy, to answer some significant questions about its economic delay over the national Unification. Using original data and the analytical approach of the informational content of surname we compute the assortativity of surnames (as proxy of family ties) within firms. We find that the presence of bonding ties, scattered over all the business networks, is highly differentiated by sector, by time span and related to the number of economic actors involved in a firm.

Abstract *Nella letteratura economica, i legami familiari radicati, tipici di aree arretrate, sono legati a fattori avversi alla crescita economica. Noi studiamo gli effetti dei legami familiari "bonding", tipici delle imprese familiari, nell'ambiente imprenditoriale del Mezzogiorno dell'800, a cavallo dell'Unità, per rispondere a domande ancora irrisolte sulle radici del suo ritardo economico. Utilizzando dati originali e l'approccio analitico del "contenuto informativo del cognome" calcoliamo l'assortitività dei cognomi (come proxy dei legami familiari) all'interno delle imprese. Come risultati, osserviamo che la presenza di legami "bonding", sparsi su tutte le reti di imprese, è fortemente differenziata per settore, per tempo e per numero di attori economici coinvolti in un'impresa.*

Key words: Family ties, Informational content of surnames, Assortativity, Permutation test

Roberto Rondinelli, Maria Carmela Schisani
Department of Economics and Statistics, University of Naples Federico II, Via Cintia, 26, Naples (Italy), e-mail: roberto.rondinelli@unina.it, e-mail: schisani@unina.it

Giancarlo Ragozini
Department of Political Science, University of Naples Federico II, Via Rodinò, 22, Naples (Italy), e-mail: giragoz@unina.it

1 Introduction

Rooted family ties in a society may be regarded as being of less of benefit to activating the forces of economic change, modernization and innovation [2, 12, 1]. Particularly for the South of Italy this form of "bonding" solidarity within families or kinship webs has been considered as a determinant of its backwardness [12]. In the literature, claims about the entrenched power of the "logic of surname" [8] – primarily regarding the traditional Southern landowning aristocratic elite and narrow middle-class groups of the late 18^{th} and 19^{th} century rising bourgeoisie [5, 8] – have been largely speculative and backed up with qualitative evidence.

We address the topic by a new different perspective by studying the effects of family ties in the entrepreneurial environment of the Southern largest city, Naples. We use original archival data on the universe of Naples enterprises to build the networks of business relations between economic actors for the twenty-year periods before and after Unification, between 1820 to 1880.

Given the framework of graph theory, we investigate the probability of having family ties in a firm. More specifically, adopting the analytical approach of the informational content of surname [6, 3], we compute the assortativity of surnames (used as proxy of family ties) within firms. We classified these latter by sector of economic activities for three time spans. We test the significance of the observed probabilities by randomly reshuffling the ties and generating their distributions under the hypothesis of random graphs [11]. We find that the presence of bonding ties, even though scattered over all the business networks, is highly differentiated by sector, by time span and is related to the number of economic actors involved in a firm. These results can be explained in connection to historical institutional changes.

Section 2 introduces a brief review of the literature on the role of family ties in economics. Section 3 describes data e methodology, while Section 4 discusses the main results.

2 The role of families in the economics: a brief review

As the earliest and most fundamental institution able to shape human activity, the family has been extensively studied in its capacity to generate major implications in terms of socio-economic outcomes that have, in turn, an influence on economic development [12]. In the economics literature, rooted family ties, mainly associated to backward areas characterized both by weak formal institutions [10] and weak social capital [2, 12], are regarded as being of less of benefit to activating the forces of economic change and modernization. Also in the most recent years, in the empirical literature focused on social capital, family ties are deemed to be related to a number of factors disruptive to economic growth: aversion to change and lower levels of innovation, trust, geographical mobility and economic dynamism [1]. Through qualitative evidence, a similar argument has been made for the South of Italy to explain the origins and the causes of its economic delay. By insisting on the entrenched

power of the "logic of surname" [8], the socio-economic literature has suggested this traditional form of rooted "bonding" solidarity within families or kinship webs as a determinant of the Southern backwardness [2, 12].

Family relationships, when translated into the context of business organizations, can be broadly classified as bonding or bridging [7] with ambivalent contents. For example, the capacity of family firms (bonding ties) to "fit" into the local community can be alternatively defined in the negative terms of paternalism as well as in the more positive view as a component of corporate social performance [4]. On the other hand, considering family relations in large companies, closer to bridging functions, the business literature underlines how this type of control relations are especially good at political rent-seeking in underdeveloped economies, where a lack of general trust, official corruption and family control of large firms are mutually reinforcing factors that combine to crowd-out growth [9]. On the other hand they are seen as facilitating the firm's access to external resources, crucial for enhancing growth [12].

In the present work, we focus on the bonding function of family ties and we adopt a quantitative approach relying on social network analysis.

3 Data and Methodology

Our work is based on a unique and original proprietary database, storing data on the universe of enterprises and companies operating in Southern Italy between 1800 and 1913. The database IFESMez (www.ifesmez@unina.it, acronym for Imprese, Finanza, Economia e Società nel Mezzogiorno) has been conceived as a large multi-source relational database[1], gathering interrelated datasets on economic and socio-political individual and collective actors. Data directly collected from original local, national and international archival sources are organized to show relations among actors, among firms and among firms and actors. At present, the database contains information regarding 3,500 firms – created and/or operating in Naples and its province – and almost 28,000 people.

Given such a database, we built three complex networks, namely three time-varying two-mode networks representing the relationships among economic actors and firms over three twenty-year spans between 1820 and 1880. These networks are also attributed and weighted. Formally, such time-varying two-mode attributed networks can be represented by a set of three two-mode networks $\{\mathscr{G}_t\}_{t=1,\ldots,3}$. The t index refers to different time spans. Each network $\mathscr{G}_t$ consists of a 6-tuple $\mathscr{G}_t(\mathscr{A}_t,\mathscr{F}_t,\mathscr{R}_t,\mathscr{W},\mathscr{S},\mathscr{C})$, where $\mathscr{A}_t = \{a_{1t},a_{2t},\ldots,a_{N_t}\}$ are the sets of N_t economic actors at time t; $\mathscr{F}_t = \{f_{1t},f_{2t},\ldots,f_{J_t}\}$ represents the set of J_t firms in which the actors are involved in each time span; $\mathscr{R}_t \subseteq \mathscr{A}_t \times \mathscr{F}_t$ is the set of the edges $r_{ijt} = (a_{it},f_{jt})$, $r_{ijt} \in \mathscr{R}_t$, i.e. each edge is an ordered couple indicating whether or not an actor a_{it} plays one, or more then one, role in the firms f_{jt}; $\mathscr{W} : \mathscr{R} \to \mathbb{N}$,

[1] The primary source of data is the Mercantile Court archival funds (from 1883 on, Civil Court) available at the Naples State Archive. For a detailed description of the database and of the information therein see [13].

with $\mathscr{W}(r_{ijt}) = w_{ijt} \in \mathbb{N}$, the number of roles of the actor a_{it} in the firm f_{jt}. Both the economic actors and the firms are featured by attributes: $\mathscr{S} : \mathscr{A}_t \rightarrow dom(\mathscr{S})$, where $\mathscr{S}(a_{it}) = s_i \in dom(\mathscr{S})$ is the *surname* of the actor a_{it}, with $dom(\mathscr{S})$ the set of surnames; $\mathscr{C} : \mathscr{F}_t \rightarrow dom(\mathscr{C})$, where $\mathscr{C}(f_{jt}) = c_j \in dom(\mathscr{C})$ is the *ATECO code* of the firm f_{jt}, with $dom(\mathscr{C})$ the list of the first-level classification of the economic sectors.

Given this complex data structure and aiming at analyzing the presence of family ties in the business network, we first adopt the logic of surnames, i.e. as we have missing information on the kinship ties (except for a relative small number of families), we use the surname as proxy of being part of the same family [3]. Hence we assume that two actors a_i and $a_{i'}$ with the same surname, $\mathscr{S}(a_i) = \mathscr{S}(a_{i'})$, participating to the same firms f_j, $r_{ij}, r_{i'j} \in \mathscr{R}$, are relatives. This assumption works well in our context given that the large part of economic actors are males and the transmission of surname follows the paternal line. Then we compute the assortativity by surname of the economic actors, dividing the firms by their *ATECO code*. Considering each twenty-year span, for each firm with more than one actor, the assortativity by surname is defined by the following probability:

$$P_{f_{jt}}(\mathscr{S}|c^*) = P(\ \mathscr{S}(a_{it}) = \mathscr{S}(a_{i't}) \wedge r_{ijt}, r_{i'jt} \in \mathscr{R} \mid \mathscr{C}(f_{jt}) = c^*\ ) \quad \forall\, d(f_{jt}) > 1, \tag{1}$$

where $d(f_{jt})$ is the weighted degree of the $j-th$ firm at time t. The probability that two randomly selected actors linked to the same firm belong to the same family can be interpreted as a measure of the bonding effect of families, i.e. the propensity of actors to form business ties with members of the same family and to be embedded in family firms. In order to test the hypothesis that such bonding effect exists, we combine the informational content of surname proposed by [6] and the random permutation test in networks [11]. More precisely, we first compute the observed average probability $\bar{P}_t(\mathscr{S}|c^*)^T$ with true surnames for each time span and *ATECO code*, and then we compute 1000 times the same probability reshuffling the surnames, i.e. under a fake distribution of surnames, $\bar{P}_t(\mathscr{S}|c^*)^F_l, l = 1, \ldots, 1000$. These latter values yield the distribution of the assortativity by surname under the null hypothesis that there is no informational content in the true surname distribution. Note that reshuffling the surnames it is equivalent to reshuffle the edges taking constant the network structure characteristics.

4 Main results

The three networks show an increasing number of vertices both in terms of existing firms and actors involved in them, denoting a more crowded market. Specifically, 286 firms and 1,905 actors in the first time span increase to 1,266 firms and 11,693 actors in the third time span, respectively, passing through 639 firms and 2,803 pre-Unification. We observe such efflorescence since after Unification when the number of limited liability companies increases mainly in the financial and industrial sectors, determining a sharp increase of actors involved. The networks are quite sparse and

the density decreases over the time reaching the value of 0.0012 in the last time span (the initial value was 0.0058). The presence of 127, 280 and 563 components (disconnected subgraphs), for the three time span, suggests that several firms (and their members) are isolated with respect to the main core of the network. Furthermore, a relevant part of these components is represented by the single dyads actor-firm or by the single triads actor-actor-firm. For example the dyads are the 22% of the components in the first period, while the triads are the 49% in the third time span. This mirrors the common and persisting trait of Southern business to be organised in individual or very small family firms. Focusing on the node-level characteristics of the networks, a strong core-periphery structure is suggested by the distributions of the degree centrality of the economic actors. In the first two periods, 67% of the actors has degree equal to 1 - i.e. they take part in one firm -, while only the 0.3% for the first period and the 0.5% for the second one, has degree higher than 10. The same happens in the third period, where the percentage of actors with degree equal to 1 increase to 77%, while the 0.4% (41 actors) shows degree higher than 10.

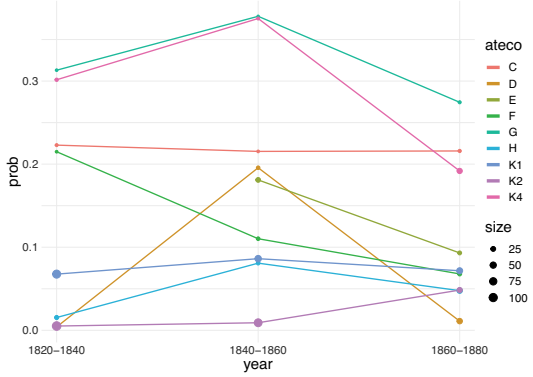
As said, as a second step, we test the probabilities of having family ties within firms with respect to random distribution of the surnames. Comparing the distribution of $\bar{P}_t(\mathscr{S}|c^*)_l^F$ with the observed values of $\bar{P}_t(\mathscr{S}|c^*)^T$ for all ATECO codes and all the time spans, we have that for 28 over 45 cases $\bar{P}_t(\mathscr{S}|c^*)^T >> \max_l(\bar{P}_t(\mathscr{S}|c^*)_l^F)$. Then all probabilities are significantly greater than the fake ones: the largest $p-value$ equal to 0.011 corresponds to the public utilities for the electricity and gas supply (D) in the 1860-1880. This entails that there is a tendency to have bonding ties in all sectors.

Given this result, in Figure 1 we report the values of the average probabilities for the main sectors for the three time spans. The size of the dots is proportional to the average degree of the firms in that sector. In the first time span, the highest probabilities are in trade (G) and banks (K4), with values greater than 0.3, as well as in the sectors of manufacturing (C) and construction (F). Trade and manufacturing are both characterized by many small economic activities like shops or artisans' workshops mainly run by members of the same family. For different reasons, in the first decades of the 19th century banks are generally family-based private firms, owned by members of the narrow inner circle of the local financial elite. In pre-Unification the trend to establish family ties increases or remains stable, for almost all the sectors. This is especially true for trade and banks. After the Unification, the increasing number of large joint stock companies and large cooperatives, both after the liberal turn of the Italian governments and the introduction of the 1882 Italian Commercial Code, yields a decrease in the tendency to form family ties. This is especially true in the banking sector due to the creation of large cooperatives with thousands shareholders, and in the trade sector due to the increasing trades and the enlarged competition. Marine insurances (K1) and other insurances (K2) show an almost constant and low (but still significant) tendency to have family ties. However such probabilities is lower due to the need of reducing risks by diversifying investments within the same family.

Finally note that in all the time spans there is a negative correlation between the probability to have family ties and the average size of firms: such correlation ranges

from -0.38 in pre-Unification to -0.54 post. This is quite expected: the larger is the number of people in a firm, the lower is the probability of having family ties. This directly entails that the presence of family ties is associated to family firms and hence to the bonding effects.

Fig. 1 *Observed probability of the main economic sector per twenty-year period.*



References

1. Alesina, A., Giuliano, P.: The power of the family. Journal of Economic growth **15.2**, 93-125 (2010)
2. Banfield, E.C.: The Moral Basis of a Backward Society. IL: The Free Press, Glencoe (1958)
3. Clark, G., Cummins, N., Hao, Y., Vidal, D.D.: Surnames: A new source for the history of social mobility. Explorations in Economic History **55**, 3-24 (2015)
4. Colli, A.: Contextualizing Performances of Family Firms: The Perspective of Business History. Family Business Review **25(3)**, 243-257 (2012) doi: 10.1177/0894486511426872
5. Davis, J.A.: Società e imprenditori nel regno borbonico, 1815-1860. Laterza, Bari (1979)
6. Guell, M., Rodriguez Mora, J.V., Telmer C.I.: The Informational Content of Surnames, the Evolution of Intergenerational Mobility, and Assortative Mating. The Review of Economic Studies **82(2)**, 693–735 (2015)
7. Gupta, J.L.: Economic and Business Environment. Vision **17(1)**, 83-85 (2013) doi: 10.1177/0972262912469569
8. Macry, P.: Ottocento: Famiglia, élites e patrimoni a Napoli. Il mulino, Torino (1988)
9. Morck, R., Yeung, B.: Agency Problems in Large Family Business Groups. Entrepreneurship Theory and Practice **27(4)**, 367-382 (2003) doi: 10.1111/1540-8520.t01-1-00015
10. North, D.C.: Institutions, institutional change and economic performance. Cambridge University Press, (1990)
11. Pelechrinis, K., Wei, D.: VA-Index: Quantifying Assortativity Patterns in Networks with Multidimensional Nodal Attributes. Plos One (2016) doi: 10.1371/journal.pone.0146188
12. Putnam, R.D., Leonardi, R., Nanetti, R.Y.: Making Democracy Work Civic Traditions in Modern Italy. Princeton University Press, (1994)
13. Schisani, M.C., Balletta, L., Ragozini, G.: Crowding out the change: business networks and persisting economic elites in the South of Italy over Unification (1840–1880). Cliometrica **15**, 89–131 (2021) doi: 10.1007/s11698-020-00204-3

Methods and Applications in Clustering

A semi-supervised clustering method to extract information from the electronic Word Of Mouth

Un metodo di clustering supervisionato per estrarre informazioni dalle review online.

Giulia Contu, Luca Frigau, Maurizio Romano, Marco Ortu

Abstract In time, electronic Word Of Mouth has become a resource to support the decision-making process. Different techniques have been proposed to extract information from online textual data. We propose a semi-supervised clustering model able to identify clusters homogeneous with respect to the overall sentiment of the analyzed texts. The model is built by combing Sentiment Analysis, and Network-based Semi-supervised Clustering. We apply the model to the Booking.com data related to the Sardinian hotels. The first results highlight the presence of different clusters non-overlapped in terms of the distribution of the overall sentiment.

Abstract *Nel tempo, il passaparola online è diventato una risorsa a supporto del processo decisionale. Sono state proposte diverse tecniche per estrarre informazioni da dati testuali online. Proponiamo un modello di clustering semi supervisionato in grado di identificare cluster omogenei rispetto al sentiment complessivo dei testi analizzati. Il modello è costruito combinando diverse metodologie come Sentiment Analysis e Network-based Semi supervised Clustering. Applichiamo il modello ai dati di Booking.com relativi agli hotel che operano in Sardegna. I primi risultati evidenziano la presenza di diversi cluster non sovrapposti con riferimento alla distribuzione del sentiment complessivo.*

Key words: electronic Word Of Mouth, semisupervised clustering, Naïve Bayes classifier, Booking.com

---

Giulia Contu
University of Cagliari, e-mail: giulia.contu@unica.it

Luca Frigau
University of Cagliari e-mail: frigau@unica.it

Maurizio Romano
University of Cagliari e-mail: romano.maurizio@unica.it

Marco Ortu
University of Cagliari e-mail: marco.ortu@unica.it

1 Introduction

In time, Internet has produced a huge amount of information through social network feed, emails, blogs, online forums, survey responses, corporate documents, news, call center logs, and electronic Word Of Mouth (e-WoM). Particularly relevant in the tourism field is the information extracted by e-WoM. It has been defined by [9] as *any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet* [9, p. 39]; and, by [3] as *any positive, negative or neutral comment, rating, ranking of a product, a service, a brand, or a person supposedly made by a former customer and that is shared with other consumers in an unstructured format* [3, p. 47].
Different techniques have been proposed to extract information from textual data with the aim to convert large volumes of text into information able to support the decision-making process [5]. These techniques are for instance cluster analysis, natural language processing, machine learning, neural networks, predictive modeling, regression models, social network analysis, sentiment analysis, signal processing, and data visualization [2].
We build on the existent literature proposing a new semisupervised clustering method useful to extract information from the e-WoM. Different researchers have previously used semi-supervised clustering to classify texts and documents. Semi-supervised clustering is a variant of the traditional clustering paradigms. It aims to obtain a better partitioning of data considering and incorporating background knowledge [6, 8]. Researchers as [11] and [12] have used this kind of approach to identify clusters inside the data using labeled information. For instance, [11] have proposed a *Semi-Supervised Cluster tree method* (SSC). It is a tree-like semi-supervised classifier built taking into account both labeled and unlabeled data. [12] have theorized the *TExt classification using Semi-supervised Clustering* (TESC) in order to improve text classification. They have used semi-supervised clustering to identify text components and, later, to use text components to predict labels of unlabeled documents. [7] have proposed a semi-supervised learning algorithm for the classification of text documents. It is defined using K-means algorithm for partitioning both labeled and unlabeled data collection. K-means algorithm is applied recursively until each partition contains labeled documents of a single class. Finally, cluster centroids are used for classifying an unknown text document.
In this paper, following the classification proposed by [1], we consider a Semi-supervised clustering model with respect to an outcome variable to identify clusters that are internally homogeneous with respect to a specific text. It involves an outcome variable and previous information about the outcome variable. The algorithm is built combing different methodologies such as Sentiment Analysis (SA), and Network-based Semisupervised Clustering (NeSSC). The proposed method is applied on the online reviews published on Booking.com.
Four sections, besides the introduction, complete this study. In Section 2 we present the real data of online reviews published on *Booking.com*. The methodological

framework is described in Section 3. In Section 4 the results are presented and, finally, in Section 5 concluding remarks are discussed.

2 Data

We have investigated the Booking.com data. Booking.com is an online platform that offers the possibility to book hotels, homes, and transport options. It promotes more than 28 million accommodation listings, including over 6.2 million listings alone of homes, apartments, and other unique places to stay. In fact, it supports the travelers in the tour planning giving the possibility to plane some of the most relevant aspects necessary to make a trip. Moreover, it gives the possibility to write a review of the travel experience to support the potential tourists in the decision process. The tourist that decides to write a review can distinguish his judgment in two different sections: one where it is possible to express the positive aspects of the experience, the other the negative aspects. To guarantee the reliability and the quality of the reviews, only the customers who have bought the service through Booking.com can publish reviews.
We have focused the analysis on the reviews of 619 hotels located on the Sardinia's coast, an Italian island. We have scraped the data from the platform *Booking.com*. We have taken into account the reviews published between 2015 to 2018. We have collected 106,800 sentences: 62,291 are positive reviews, whilst 44,509 are negative reviews.
We have carried out the analysis taking into account specific topics presented in the reviews. We have identified $M = 12$ main categories of words, as shown in Table 1. They are grouped in four main relevant aspects that influence the judgement of the hotel experience. They are the elements related to the room quality, the provided services, the staff, and the hotel characteristics.

Feature	Description	
Provided Services	bar	bar service, minibar or affiliated bars around the structure
	general services	quality of services provided by the hotel
	food	food quality served in the hotel structure or affiliated restaurant
	wifi	Wi-Fi connection quality, such as general coverage and lags
Room quality	cleaning	cleaning of the room and of the hotel structure
	comfort	comfort perceived by guests during their stay
	room	room quality referred to the type of room
	sleep-quality	quality of sleep such as bed, pillows and noises
Hotel characteristics	structure	quality of physical structure
	position	position quality of the hotel structure
	quality-price-rate	quality-price rate
Staff	staff	kindness, courtesy, professionalism of the staff

Table 1 Reference categories of words.

3 Methodology

As evidenced before, we propose a semisupervised clustering method associated with an outcome variable. The clusters are identified considering a specific outcome variable and using the previous information about it [1].
The model is built combining Sentiment Analysis (SA), and Network-based Semisupervised Clustering (NeSSC). This combination allows identifying clusters homogeneous with respect to the overall sentiment of the texts analyzed.
It operates in different steps. More in detail, in the first stage the data has been analyzed and cleaned to homogenize the text and to remove all unnecessary words, stop words, acronyms. Later, Sentiment Analysis (SA) is applied in order to firstly categorize them into positive or negative classes, secondly to identify an overall sentiment score of the texts. The score is estimated through the log-likelihood ratio of the specific categories that can be or not in a sentence. This information allows us to evaluate the relevance of specific categories of words and how this category can influence positive or negative reviews.
In the last stage, the previously estimated score becomes the outcome variable used to define the clusters. In this phase, the Network-based Semisupervised Clustering (NeSSC) is used to partition the instances into K disjoint groups [4]. NeSSC identifies the pairwise affinity of the instances through a classifier and uses this information to organize them into a complex network. It operates in three steps. They are called initialization, training, and agglomeration. The initialization and training steps deal with the estimation of the affinities between observations by using an iterative process. These two steps aim to estimate an affinity matrix Π, where the amount of time that two generic elements i and j are classified together measures the propensity to behave in the same manner to sentiment scores, $\mathbf{y}$.
Finally, the affinity matrix Π has been transformed into a set of complex networks, where the observations correspond to nodes and the entries $\pi_{i,j}$ of the affinity matrix defines the strength of the edge between the two generic observations i and j. Later, a community detection algorithm is trained on the nodes of the networks in order to find, for each network, a certain number of homogeneous communities corresponding to the groups defining a proper partition of the original data. Finally, among all the partitions, the optimal one is that presenting the lowest minimum penalized average overlapping in terms of the distribution of the overall sentiment, the latter computed alternatively as the weighted mean of the pairwise group overlapping index [10]. Broadly speaking, the overlapping index η between two probability density functions $f_A(x)$ and $f_B(x)$ is defined as

$$\eta_{A,B} = \int_{\mathbb{R}} \min\left[f_A(x), f_B(x)\right] dx \tag{1}$$

where the integral can be replaced by summation in the discrete case. Usually, $\eta_{A,B}$ is normalized to one, thus to easily indicate that the two distributions $f_A(x)$ and $f_B(x)$ might be not overlapped ($\eta_{A,B} = 0$) or fully overlapped ($\eta_{A,B} = 1$).

4 Results

The results show the presence of five clusters significantly characterized by a different level of the overall sentiment (Table 2). Each cluster presents specific characteristics and a specific average value of the outcome variable and of the four most important variables.
The first cluster presents a very large and positive overall sentiment. All variables characterized positively and significantly the cluster highlighting how the positive sentiment is influenced by all categories taken into account. The clusters two and three present in a similar way a positive overall sentiment; however, some differences can be observed. Specifically, the sentiment of cluster two is influenced positively by the categories room and staff. On the contrary, only the category staff influences the third cluster. Finally, the last two clusters are characterized by an overall sentiment lower than the mean. The unique element that is significant for cluster four is the staff. On the contrary, cluster five is characterized for a negative impact of the room, the staff, and the provided services. Only the characteristics of the hotel are not able to influence the overall rating of this cluster.

Feature	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
overall sentiment	**1.285**	*******	**0.676**	*******	**0.354**	******	**-0.128**	*****	**-0.801**	*******
room	0,39	***	0,19	***	0,12		-0,03		-0,25	***
Provided services	0,35	**	0,19		0,14		-0,07		-0,20	***
Hotel Characteristics	0,04	**	0,19		0,06		0,00		-0,13	
staff	1.787	***	0,74	***	0,33	**	-0,17	**	-0,96	***

Table 2 The five clusters obtained partitioning *Booking.com*.

5 Conclusions

This paper has proposed Semi-Supervised Sentiment Clustering on natural language Texts framework for clustering textual data according to the overall sentiment expressed in their texts. The method is built by combining different statistical and natural language processing techniques and methodologies. The framework has been applied to *Booking.com* dataset concerning hotel structures of Sardinia (Italy). The results highlight the feasibility and interpretability of the framework with real data: we were able to find well-defined, non-overlapping in terms of the distribution of the overall sentiment, and interpretable clusters among the hotel structures. The implication of the results from the managerial point of view highlighted actionable insights spendable in decision-making support.

References

1. Bair, E. (2013). Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: Computational Statistics, 5(5), 349-361.
2. De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review.
3. Filieri, R. (2016). What makes an online consumer review trustworthy?. Annals of Tourism Research, 58, 46-64.
4. Frigau, L., Contu, G., Mola, F., & Conversano, C. (2021). Network-based semisupervised clustering. Applied Stochastic Models in Business and Industry, 37(2), 182-202.
5. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.
6. Gao, J., Tan, P. N., & Cheng, H. (2006, April). Semi-supervised clustering with partial background information. In Proceedings of the 2006 SIAM International Conference on Data Mining (pp. 489-493). Society for Industrial and Applied Mathematics.
7. Gowda, H. S., Suhil, M., Guru, D. S., & Raju, L. N. (2016, December). Semi-supervised text categorization using recursive K-means clustering. In International Conference on Recent Trends in Image Processing and Pattern Recognition (pp. 217-227). Springer, Singapore.
8. Hang, W., Choi, K. S., Wang, S., & Qian, P. (2017). Semi-supervised learning using hidden feature augmentation. Applied Soft Computing, 59, 448-461.
9. Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?. Journal of interactive marketing, 18(1), 38-52.
10. Inman, H. F., & Bradley Jr, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Communications in Statistics-theory and Methods, 18(10), 3851-3874.
11. Sun, Z., Ye, Y., Zhang, X., Huang, Z., Chen, S., & Liu, Z. (2012, December). Batch-mode active learning with semi-supervised cluster tree for text classification. In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 388-395). IEEE.
12. Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. Knowledge-Based Systems, 75, 152-160.

Spectral approach for clustering three-way data

Un approccio spettrale per il clustering di dati matriciali

Cinzia Di Nuzzo and Salvatore Ingrassia

Abstract Spectral methods are a quite recent approach for both data clustering and dimensionality reduction based on the properties of the Laplacian matrix of the pairwise similarity matrix coming from a suitable kernel function. In this paper, a spectral clustering approach for the three-way data is presented, where three-way data are composed by a $t \times p$ matrix for each statistical observation, so each unit is represented by a matrix. The relationship between the matrices are here described a new kernel function based on the $L_{2,1}$ norm. The performance of this approach is illustrated on the ground of a real three-way data set.

Abstract *I metodi spettrali costituiscono un approccio abbastanza recente di clustering e di riduzione dimensionale di insiemi di dati basato sulle proprietà del Laplaciano di una matrice di similarità derivante da un'opportuna funzione kernel. In questo articolo, si presenta un metodo di clustering spettrale per dati matriciali, che sono costituiti da una matrice di dimensione $t \times p$ per ogni osservazione statistica; in particolare, ogni unità è rappresentata da una matrice. Al fine di descrivere la relazione tra le matrici, viene introdotta una nuova funzione kernel basata sulla norma $L_{2,1}$. La proposta viene illustrata presentando un'applicazione su dati reali.*

Key words: Spectral clustering, Kernel function, Three-way data

1 Introduction

Spectral clustering methods are based on the graph theory, where the units are represented by the vertices of an undirect graph and the edges are weighted by the

Cinzia Di Nuzzo
Department of Statistics - Sapienza University of Roma e-mail: cinzia.dinuzzo@uniroma1.it

Salvatore Ingrassia
Department of Economics and Business - University of Catania, e-mail: s.ingrassia@unict.it

pairwise similarities coming from a suitable kernel function, so the clustering problem is reformulated as a graph partition problem, see e.g. [14], [7].

Here, we present a spectral clustering approach for clustering *three-way data*, where a suitable kernel function based on $L_{2,1}$ norm between two matrices is introduced.

Three-way data derives from the observation of various attributes measured on a set of units in different situations; some examples are longitudinal data on multiple response variables and multivariate spatial data. In particular, this kind of data is composed of three objects: n units (matrices), p variables (columns), and t times (occasions). Clustering of three-way data has attracted a growing interest in literature, see e.g. [12], [1]. Moreover, model-based clustering of three-way data has been introduced by [13]; in particular, in the framework of matrix-variate normal mixtures, [8] handle on parsimonious models for modeling matrix data, [9] consider the elliptical heavy-tailed generalization of the matrix-variate normal distribution, [10] deal with three-way data clustering using matrix-variate cluster-weighted models and [11] present an application to educational data via mixtures of parsimonious matrix-normal distribution.

The rest of the work is organized as follows: in Section 2 the three-way spectral clustering is introduced; in Section 3 an application based on real three-way data is presented. Finally, in Section 4 we provide concluding remarks.

2 Three-way spectral clustering

Three-way data is organized in three modes: n units, p variables, and t occasions. Therefore, given n matrices that represent the vertices of the graph, each matrix is composed by p columns that represent the variables and t rows that represent the time or another feature. So we have a tensor of dimension $n \times t \times p$, thus the data set is a tensor $\{\mathbf{X}\}_{ijk}$ for $i = 1,\dots,n$, $j = 1,\dots,t$, $k = 1,\dots,p$.

Let $V = \{X_{1jk}, X_{2jk}, \dots, X_{njk}\} \subset \mathscr{X}$, for $j = 1,\dots,t$, $k = 1,\dots,p$, be a set of vertices in an undirect graph. In order to group the data V in K cluster, the first step of a spectral clustering algorithm consists of the definition of a symmetric and continuous function $\kappa : \mathscr{X} \times \mathscr{X} \to [0,\infty)$ called the *kernel function*. In the following, we consider the so-called *adaptive density aware kernel* (a two-way version of this kernel is introduced in [6], where the Euclidean distance is considered),

$$\kappa : \mathbf{X} \times \mathbf{X} \to [0,+\infty), \qquad \kappa(i_1,i_2) = \exp\left(-\frac{\sum_{k=1}^{p}\sqrt{\sum_{j=1}^{t}|X_{i_1jk} - X_{i_2jk}|^2}}{\varepsilon_i \varepsilon_j CNN(X_{i_1jk}, X_{i_2jk})}\right). \quad (1)$$

where

- ε_i is the distance between X_{i_1jk} and its h-th neighbor X_{hjk};
- $CNN(X_{i_1jk}, X_{i_2jk}) = |\mathscr{B}(X_{i_1jk}, \tau) \cap \mathscr{B}(X_{i_2jk}, \tau)|$, where $\mathscr{B}(X_{i_1jk}, \tau)$ is the sphere centered in X_{i_1jk} with radius τ.

Afterwards, a *similarity matrix* $W = (w_{i_1 i_2})$ can be assigned by setting $w_{i_1 i_2} = \kappa(i_1, i_2) \geq 0$, for $X_{i_1 jk}, X_{i_1 jk} \in \mathscr{X}$ and finally the *normalized graph Laplacian* matrix $L_{\text{sym}} \in R^{n \times n}$ is introduced as

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}, \tag{2}$$

where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is the *degree matrix* and d_i is the *degree* of the vertex $\mathbf{X}_i$ defined as $d_i = \sum_{j \neq i} w_{ij}$ and I denotes the $n \times n$ identity matrix. The Laplacian matrix L_{sym} is positive semi-definite with n non-negative eigenvalues. For a fixed K, let $\{\gamma_1, \ldots, \gamma_K\}$ be the eigenvectors corresponding to the smallest K eigenvalues of L_{sym}. Then the *normalized Laplacian embedding* is defined as the map $\Phi_\Gamma : \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \to R^K$ given by

$$\Phi_\Gamma(\mathbf{X}_i) = (\gamma_{1i}, \ldots, \gamma_{Ki}), \quad i = 1, \ldots, n,$$

where $\gamma_{1i}, \ldots, \gamma_{Ki}$ are the i-th components of $\gamma_1, \ldots, \gamma_K$, respectively. In other words, the function $\Phi_\Gamma(\cdot)$ maps the data from the input space $\mathscr{X}$ to a feature space defined by the K principal subspace of L_{sym}. Afterwards, let $Y = (\mathbf{y}_1', \ldots, \mathbf{y}_n')$ be the $n \times K$ matrix given by the embedded data in the feature space, where $\mathbf{y}_i = \Phi_\Gamma(\mathbf{X}_i)$ for $i = 1, \ldots, n$.

Remark 1. In the spectral clustering algorithm, the embedded data Y are clustered according to some clustering procedure; usually, the k-means algorithm is taken into account in literature. However, here, the Gaussian mixtures have been proposed because have more flexible cluster shapes with respect to the k-means, see [5] and [2]. Finally, we point out that the performances of other mixture models based on non-Gaussian component densities have been analyzed, but Gaussian mixture models can be considered as a good trade-off between model simplicity and effectiveness, see [3] for details.

Remark 2. The spectral clustering algorithm requires to set the number of clusters K and the kernel function κ. In order to select the number of clusters K, in the following we consider the method proposed in [4]. In practice, K is detected by a joint analysis of the number of blocks in the Laplacian matrix, the maximum eigengap between two consecutive eigenvalues of the Laplacian matrix, and the number of spikes of the embedded data that can be visualized after the Laplacian embedding.

Remark 3. Finally, we point out an important advantage of the three-way spectral clustering approach with respect to others approaches. As matter of fact, differently from approaches based on mixtures of matrix-variate data, the number of variables of the data set is not a critical issue because the spectral clustering algorithm is based on distance measures.

3 Example: Satellite data

Here, we analyze a matrix data set purchased by NASA from the Australian Center for Remote Sensing, it has already been analyzed in [13] and in [8]. The Satellite data set is included in the `MatManlyMix` R package and contains 845 units composed of 4×9 matrices. Thus, we have $n = 845$, $t = 4$ and $p = 9$. In particular, it represents a neighborhood of 9 pixels of satellite images of the same scene using 4 different spectral bands: two belonging to the spectrum of the visible region (green and red) and two of infrared. The units of satellite images correspond to 3 types of soil: gray soil, damp gray soil, and soil with vegetation stubble. Therefore, Satellite data has $K = 3$ labeled classes and, in order to perform the validity of our proposal, we will use ARI and accuracy.

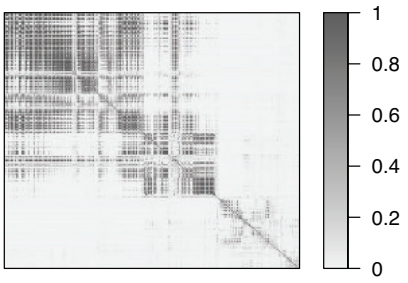
In Figure 1, we present the graphic features of the spectral clustering process setting $h = 7$ and τ equal to the first quartile of the $L_{2,1}$ distribution that is defined by

$$\delta_{2,1}(X_{i_1 jk}, X_{i_2 jk}) = \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{t} |X_{i_1 jk} - X_{i_2 jk}|^2}, \qquad for \quad i_1, i_2 = 1, \ldots, n. \tag{3}$$

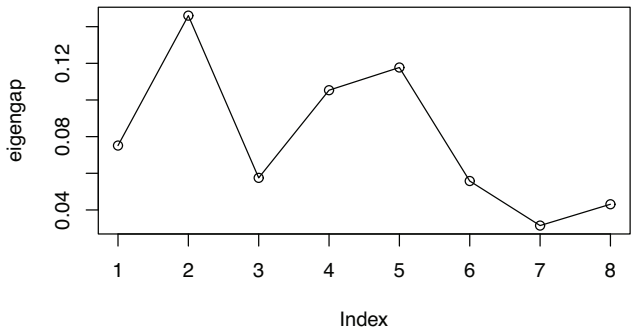
From the number of blocks in Figure 1-a), the maximum eigengap in Figure 1-b), and the number of spikes in Figure 1-c), we can deduce that the number of clusters is 3. The three-way spectral clustering algorithm has been carried out and the resulting accuracy is equal to 0.8722. Finally, comparing our approach with [8], we can state that our method improves clustering performance, because in [8] the best model had an ARI equal to 0.42, while our method yields an ARI of 0.68. On the contrary, comparing our approach with [13], we get a slightly worse result, indeed the classification error rate in [13] is equal to 0.11 while in our approach it is equal to 0.13.

4 Conclusion

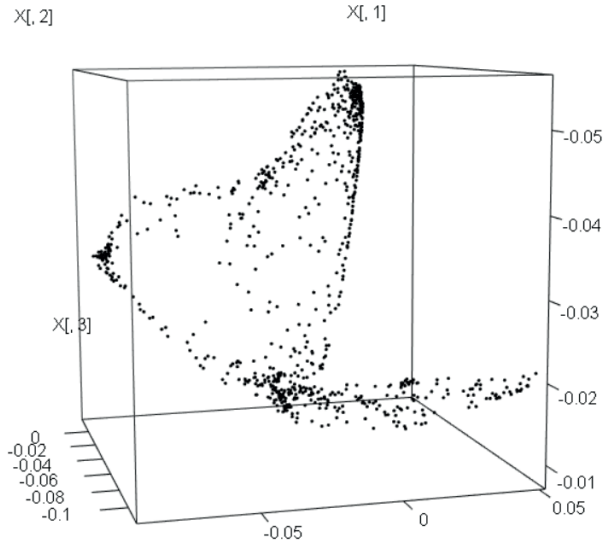
A spectral approach to cluster three-way data has been proposed. The matrix data are organized in a tensor, thus the vertices in the graph are represented by the matrices of dimension $t \times p$. In order to weight the vertices in the graph, a kernel function based on the $L_{2,1}$ norm between the matrix difference has been introduced. The performance of the three-way spectral clustering algorithm has been shown in one real three-way data set. The spectral method is competitive with respect to other clustering methods proposed in the literature to perform matrix-data clustering. Finally, in order to provide suggestions for future research, other kernel functions can be introduced considering different distances with respect to the $L_{2,1}$ norm.



a)



b)



c)

Fig. 1 *Satellite data.* Spectral clustering features: *a*) plot of Laplacian matrix in greyscale; *b*) plot of the first eight eigengap values; *c*) scatterplot of the embedded data along with directions $(\gamma_1, \gamma_2, \gamma_3)$.

References

1. Bocci, L., Vicari, D. (2019). ROOTCLUS: Searching for "ROOT CLUSters" in Three-Way Proximity Data. *Psychometrika*, **84**, 941–985.
2. Di Nuzzo, C. Ingrassia S. (2022). A mixture model approach to spectral clustering and application to textual data, *Statistical Methods and Applications (forthcoming).*.
3. Di Nuzzo, C. (2021). Model selection and mixture approaches in the spectral clustering algorithm,*Ph.D. thesis, Economics, Management and Statistics, University of Messina*.
4. Di Nuzzo, C. Ingrassia S. (2022). A joint graphical approach for model selection in the spectral clustering algorithm, *Tech. rep., submitted for publication*.
5. Di Nuzzo C., Ingrassia S. (2021). Some Issues on the Parameter Selection in the Spectral Methods for Clustering, *Book of Short Papers SIS 2021, Perna C., Salvati N. Schirripa Spagnolo F. eds. ISBN 9788891927361*.
6. John, C. R., Watson, D., Barnes, M. R., Pitzalis, C., and Lewis, M. J. (2019). Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, **36**(4), 1159–1166.
7. Meila, M. (2015). Spectral clustering. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci, editors, *Handbook of Cluster Analysis*. Chapman and Hall/CRC.
8. Sarkar, S., Zhu, X., Melnykov, V., and Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, **142**, 106822.
9. Tomarchio, S. D., Punzo, A., and Bagnato, L. (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics and Data Analysis*, **152**, 107050.
10. Tomarchio, S. D., McNicholas, P., and Punzo, A. (2021a). Matrix normal cluster-weighted models. *Journal of Classification*.
11. Tomarchio, S. D., Ingrassia, S., and Melnykov, V. (2021b). Modeling students' career indicators via mixtures of parsimonious matrix-normal distributions. *Australian & New Zealand Journal of Statistics https://doi.org/10.1111/anzs.12351*.
12. Vichi M., Rocci R., Kiers H.A.L. (2007). Simultaneous Component and Clustering models for three-way data: Within and Between Approaches. *Journal of Classification*, **24**, 71–98.
13. Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**, 511–522.
14. von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**(4), 395–416.

Double clustering with a matrix-variate regression model: finding groups of athletes and disciplines in decathlon's data

Doppio clustering con il modello di regressione matriciale: ricercare gruppi di atleti e discipline in dati di decathlon

Mattia Stival and Mauro Bernardi and Manuela Cattelan and Petros Dellaportas

Abstract Decathlon is a two-day track and field competition in which athletes compete in 10 different disciplines placed in an invariant order. In this work we propose a co-clustering method using a generalization of the Bayesian matrix-variate regression model, in which both athletes and disciplines are grouped according to the observed data. Specifically, we describe the model and briefly outline the inferential strategy, based on a Gibbs sampling algorithm. The model provides powerful insights that can be used by coaches and athletes to improve their performances.
Abstract *Il Decathon è una gara dell'atletica che si svolge in due giorni in cui gli atleti si sfidano in* 10 *discipline diverse che seguono un ordine fisso. In questo lavoro proponiamo un metodo di co-clustering utilizzando una generalizzazione del modello di regressione matrice-variato, in cui sia gli atleti sia le discipline vengono raggruppati sulla base dei dati osservati. In questo articolo viene descritto il modello e si delinia brevemente la strategia inferenziale. Il modello fornisce spunti interessanti che possono essere utilizzati da atleti e allenatori per migliorare i loro risultati.*

Key words: co-clustering, decathlon, sports statistics

Mattia Stival
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova.
e-mail: mattia.stival@unipd.it

Mauro Bernardi
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova.

Manuela Cattelan
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova.

Petros Dellaportas
Department of Statistical Science, University College London, London
Department of Statistics, Athens University of Economics and Business, London
The Alan Turing Institute, London

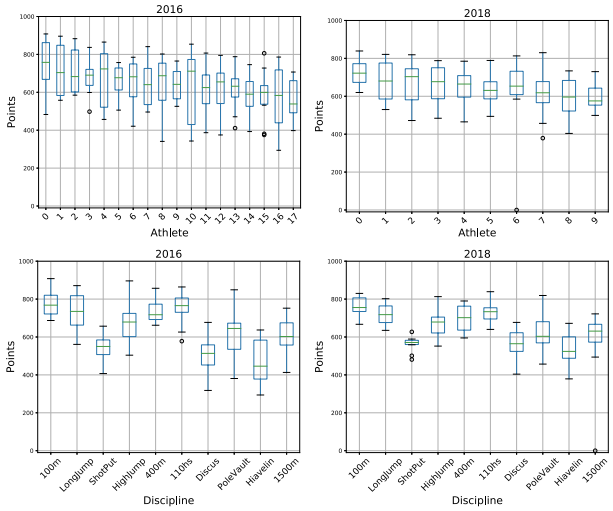


Fig. 1 Results of two different decathlon competitions. In the first row, the boxplots represent the marginal distribution of points earned by each athlete during two decathlon competitions (2016 and 2018 national championships). In the second row, the boxplots represent the marginal distribution of points earned by the same athletes in different disciplines.

1 Introduction

Decathlon is a two-day track and field competition in which athletes compete in 10 different disciplines placed in an invariant order. In the first day, the athletes compete in the following disciplines: 100meters, long jump, shot put, high jump, and 400 meters; in the second day, athletes compete in: 110 meters hurdles, discus, pole vault, javelin and 1500 meters. During the competition, athletes are evaluated according to an international scoring system (see, www.worldathletics.org), useful for evenly evaluating their performances and for deriving the final ranking of a competition. From a sports science point of view, the disciplines are typically grouped in running, jumping and throwing disciplines, however, [4] showed the existence of an alternative grouping structure that highlights different relations among the 10 disciplines. If, one side, clustering disciplines can provide insights on the relations between disciplines that are useful for both training and competition purposes, on the other side, recognizing how one athlete performs is a key aspect from a coaching point of view. For example, knowing that one athlete performs better in a group of disciplines with respect to the others can suggest which are the important disciplines to train, in order to improve the total score. Grouping athletes means recognizing differences between them and, thus, showing strengths and weaknesses of each subject. Differences among athletes are indeed typically expected, with some athletes stronger in some disciplines and other athletes who are better performers in others.

In this work, we propose a double clustering strategy for finding groups of disciplines and athletes in different decathlon Italian championship competitions, yearly

held in the period 2012–2021 (source of data: www.fidal.it). The results observed in two championships are summarized in Figure 1. Double clustering (co-clustering) is a statistical technique that allows to find groups of rows and of columns in a matrix of observations [2]. We tackle the double clustering problem within a matrix-variate regression framework, in which matrix of results in different competitions are explained by means of a low dimensional matrix of means. Matrix-variate regression modelling is a topic which has recently catalyzed increasing attention in statistics, in light of the need to analyze datasets in which observations of different units arise in the form of a matrix [3, 1]. Our contribution differs from standard matrix-variate regression models as we let the left and right design matrices to be random selection matrices, allowing, by means of a compact representation, the clustering of rows, corresponding to different athletes participating in a competition, and the clustering of columns, corresponding to different decathon's disciplines. In the following, Section 2 presents the model and briefly outlines the inferential strategy, Section 3 presents the results of the real data application.

2 The model

Let $\mathbf{Y}_t \in \mathbb{R}^{P_t \times Q}$ be the matrix storing the results of the t–th decathlon competition in the sample, where the rows of $\mathbf{Y}_t$ store different athletes and the columns of $\mathbf{Y}_t$ store different disciplines. Here, P_t denotes the number of athletes participating in the t–th competition, which is different for each race, and $Q = 10$ denotes the number of disciplines in decathlon, and T the total number of independent races in the sample.

We consider the following model:

$$\mathbf{Y}_t = \mathbf{L}_t \mathbf{M} \mathbf{R}^\top + \mathbf{E}_t, \quad \mathbf{E}_t \sim MN_{P_t \times Q}(\mathbf{0}, \Sigma_t).$$

The matrix $\mathbf{E}_t$ is a matrix of errors, and follows a zero-mean matrix-variate normal distribution with covariance $Cov(\varepsilon_t) = \Sigma_t$, with $\varepsilon_t = Vec(\mathbf{E}_t)$ with $Vec(\cdot)$ denoting the vec-operator. For simplicity, in this work we assume $\Sigma_t = \sigma^2_\varepsilon \mathbf{I}_{P_t Q}$, meaning that the errors within $\mathbf{E}_t$ are independent and are characterized by the same scale regardless of the athlete and the discipline we are considering. The matrix $\mathbf{M} \in \mathbb{R}^{J \times G}$ is a matrix of means, storing, on different rows, the means related to each group of athletes, and, on the columns, the means related to different group of disciplines. Typically, J and G are assumed to be $J \ll \sum_{t=1}^{T} P_t$ and $G \ll Q = 10$, allowing for a parsimonious model in which few groups explain data of large dimensions.

The matrices $\mathbf{L}_t$ and $\mathbf{R}$ are left and right random selection matrices of dimensions $P_t \times J$ and $Q \times G$, respectively. The p_t–th row of $\mathbf{L}_t$ is $\mathbf{l}_{p_t} = [I(l_{p_t} = 1), \ldots, I(l_{p_t} = J)]$, where $l_{p_t} \in \{1, \ldots, J\}$ is a random cluster allocation of athlete l_{p_t}, and $I(\cdot)$ is the function such that $I(l_{p_t} = j) = 1$ if athlete p_t belongs to the j–th group of athletes, and 0 otherwise. Thus, the role of $\mathbf{L}_t$ is to indicatite to which of the J groups of athletes each athlete participating to competition t belongs to, as well as to select the specific row of $\mathbf{M}$ associated to each athlete. In a similar way, the q–th row of

the matrix $\mathbf{R}$ is $\mathbf{r}_q = [I(r_q = 1), \ldots, I(r_q = G)]$, where $r_q \in \{1, \ldots, G\}$ is the random cluster allocation of discipline, indicating to which of G groups of disciplines the discipline r_q belongs to. Thus, the role of $\mathbf{R}$ is indicating to which of the G groups of disciplines each discipline belongs to, and to select the corresponding columns of $\mathbf{M}$ for each discipline. Cluster allocations for both athletes and disciplines are assumed to be independent multinomial distributions, such that $l_{p_t} \sim Mult(1, \pi_L)$ and $r_q \sim Mult(1, \pi_R)$, where $\pi_L = (\pi_{L1}, \ldots, \pi_{LJ})$ and $\pi_R = (\pi_{R1}, \ldots, \pi_{RG})$ are vector of probabilities such that $\sum_{j=1}^{J} \pi_{Lj} = 1$ and $\sum_{g=1}^{G} \pi_{Lg} = 1$, with $\pi_{Lg} \geq 0$ and $\pi_{Rj} \geq 0$.

Due to the limited space, we only derive the scalar version of the model, in which the (p_t, q)–th entry of $\mathbf{Y}_t$ is

$$y_{p_t q,t} = \sum_{j=1}^{J} \sum_{g=1}^{G} I(l_{p_t} = j) I(r_q = g) \mu_{jg} + \varepsilon_{p_t q,t},$$

where μ_{jg} indicates the (j, g)–th entry of $\mathbf{M}$. The scalar μ_{jg} represents the conditional mean of the result $y_{p_t q,t}$ when athlete p_t belonging to group j (of athletes) competes in a discipline of group g (of disciplines). We refer to the literature related to the matrix variate regression model for alternative ways of interpreting the model [3, 1].

2.1 Posterior distribution and Gibbs sampling

Let $\mathscr{L} = \{\mathbf{L}_1, \ldots, \mathbf{L}_T\}$, $\mathscr{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$, and also $\Theta = \{\sigma_\varepsilon^2, \mathbf{M}, \mathbf{P}_{R0}, \pi_L, \pi_R\}$. We adopt a Bayesian approach and derive the augmented likelihood, which is

$$p(\mathscr{L}, \mathbf{R}, \mathscr{Y} | \Theta) \propto \prod_{t=1}^{T} \left[p(\mathbf{Y}_t | \mathbf{L}_t, \mathbf{M}, \mathbf{R}, \sigma_\varepsilon^2) p(\mathbf{L}_t | \pi_L) \right] p(\mathbf{R} | \pi_R),$$

where we have assumed

$$p(\mathbf{L}_t | \pi_L) = \prod_{p_t=1}^{P_t} \prod_{j=1}^{J} \pi_{Lj}^{I(l_{p_t}=j)}, \quad p(\mathbf{R} | \pi_R) = \prod_{q=1}^{Q} \prod_{g=1}^{G} \pi_{Rg}^{I(r_q=g)}.$$

We adopt a conditionally conjugate approach, in which the prior probability of Θ factorizes as follows

$$p(\Theta) = p(\mathbf{M} | \mathbf{P}_{R0}) p(\mathbf{P}_{R0}) p(\sigma_\varepsilon^2) p(\pi_R) p(\pi_L).$$

Here, $\pi_L = (\pi_{L1}, \ldots, \pi_{LJ}) \sim Dir_J(\mathbf{1}_J^\top / J)$ and $\pi_R = (\pi_{R1}, \ldots, \pi_{RG}) \sim Dir_G(\mathbf{1}_G^\top / G)$, which are conjugate priors under the multinomial model. We assume also that $\sigma_\varepsilon^2 \sim IG(1, 1)$, and set $\mathbf{M} | \mathbf{P}_{0R} \sim MN_{J \times G}(\mathbf{M}_0, \mathbf{P}_{0R} \otimes \mathbf{I}_J)$, and let $\mathbf{P}_{0R} \sim IW_G(\nu, \mathbf{P}_{00})$ be an inverse Wishart with large variance. Under these assumptions, it is possible to obtain a sample of the posterior distribution

$$p(\Theta, \mathscr{L}, \mathbf{R}|\mathscr{Y}) \propto p(\mathscr{L}, \mathbf{R}, \mathscr{Y}|\Theta)p(\Theta)$$

by deriving a Gibbs sampling algorithm that iteratively simulates from the following full conditionals:

1. $p(\mathbf{M}|\mathscr{L}, \mathbf{R}, \sigma^2_\varepsilon, \mathbf{P}_{R0}, \mathbf{Y}_t) \propto p(\mathbf{M}|\mathbf{P}_{R0}) \prod_{t=1}^{T} p(\mathbf{Y}_t|\mathbf{L}_t, \mathbf{M}, \mathbf{R}, \sigma^2_\varepsilon)$;
2. $p(\mathbf{P}|\mathbf{M}) \propto p(\mathbf{M}|\mathbf{P}_{R0})p(\mathbf{P}_{R0})$;
3. $p(\sigma^2_\varepsilon|\mathscr{L}, \mathbf{M}, \mathbf{R}, \mathscr{Y}) \propto p(\sigma^2_\varepsilon) \prod_{t=1}^{T} \left[p(\mathbf{Y}_t|\mathbf{L}_t, \mathbf{M}, \mathbf{R}, \sigma^2_\varepsilon)\right]$;
4. $p(\mathbf{L}_t|\mathbf{R}, \mathbf{M}, \sigma^2_\varepsilon, \pi_L, \mathbf{Y}_t) \propto p(\mathbf{Y}_t|\mathbf{L}_t, \mathbf{M}, \mathbf{R}, \sigma^2_\varepsilon)p(\mathbf{L}_t|\pi_L)$, for $t = 1, \ldots, T$;
5. $p(\mathbf{R}|\mathbf{M}, \mathscr{L}, \sigma^2_\varepsilon, \pi_R, \mathscr{Y}) \propto \prod_{t=1}^{T} \left[p(\mathbf{Y}_t|\mathbf{L}_t, \mathbf{M}, \mathbf{R}, \sigma^2_\varepsilon)\right] p(\mathbf{R}|\pi_R)$;
6. $p(\pi_L|\mathscr{L}) \propto p(\pi_L) \prod_{t=1}^{T} p(\mathbf{L}_t|\pi_L)$ and $p(\pi_R|\mathbf{R}) \propto p(\pi_R)p(\mathbf{R}|\pi_R)$.

Steps 1-3 are simply updates of a Bayesian linear model, obtained by deriving the full conditional of the model in its vectorized form, since the selection matrices are given. Steps 4-5 are cluster-allocation updates, in which (conditional) independent updates are obtained for each row of $\mathbf{L}_t$ and $\mathbf{R}$. Finally, step 6 are standard Dirichlet updates.

3 Application

We apply our method under three different scenarios, where $J = G = 3$ for all the scenarios. In the first case, we consider the data using the original grading system, in which no transformation of the data is required. In the second case, we center the observations with respect to the discipline-specific mean of each competition, i.e. we apply the following transformation $y^\star_{p_t q,t} = y_{p_t q,t} - \frac{1}{P_t} \sum_{p=1}^{P_t} y_{pq,t}$, for $p_t = 1, \ldots, P_t$, $q = 1, \ldots, Q$, and $t = 1, \ldots, T$. In the third case, we apply the following transformation: we first center the observations with respect to the discipline-specific mean of each competition to obtain $y^\star_{p_t q,t}$, and re-center the transformed data with respect to the athlete-specific mean of the centered data, i.e. $y^{\star\star}_{p_t q,t} = y^\star_{p_t q,t} - \frac{1}{Q} \sum_{q=1}^{Q} y^\star_{p_t q,t}$, for $p_t = 1, \ldots, P_t$, $q = 1, \ldots, Q$, and $t = 1, \ldots, T$. In the first scenario (original data), the model recognizes three groups of athletes (index J) that differ with respect to the group-specific means of all the groups of disciplines (index G). More specifically, we recognize an ordering in the performance of the three groups of athletes (J), with respect to the points earned by the athletes in different disciplines (the higher the points, the better the performance). Group 1 (blue) in Figure 2 is overall better than group 2 (orange) and group 2 is overall better than group 3 (green) for all the groups of disciplines. By looking at how disciplines are grouped we recognize disciplines that allow to earn many points (group $j = 2$: 100 m, long jump, high jump, 400 m, 110 hs), disciplines with few points (group $j = 3$: pole vault and 1500 m), and intermediate disciplines that allows to earn, on average, a number of points in between group 2 and 3. In light of these results, it seems that there are disciplines that allow athletes to score more points due to the ranking system rather than actual differences between disciplines. For this reason, the performances of the athletes in different disciplines were centered with respect to the discipline-specific mean of

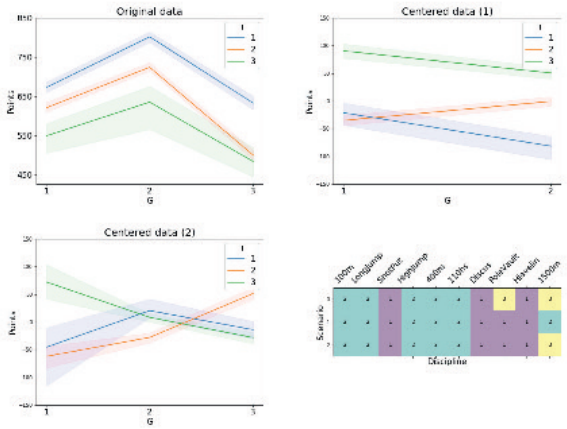


Fig. 2 The first three graphs show the median and the 90% quantile-based credible interval for the posterior distribution of the matrix of means **M** for the considered scenarios. The bottom-right panel shows how different disciplines are grouped in the three considered scenarios according to MAP.

each competition. In this case—i.e. centered data (1)—one discipline-specific group is not filled, leading to only two groups of disciplines. Athletes are clustered in three groups. Group 3 (green) is the group with the highest scoring athletes in both groups of disciplines. On the contrary, groups 1 (orange) and 2 (green) score similarly with respect group $g = 1$ of disciplines, but are characterized by differences between the number of points earned in group $g = 2$ of disciplines (shot put, discus, pole vault, javelin). In the third scenario performances were centered with respect to both the discipline- and athlete-specific means, in order to detect differences between groups that are not influenced by the personal strength and the scoring system. In this case, larger differences in scoring were found between group $j = 3$ and the others, the former characterized by a large positive mean in group of discipline $g = 1$, composed of throwing disciplines and the pole vault.

Aknowledgments: This research was supported by funding from the University of Padova Research Grant 2019-2020 under grant agreement BIRD203991.

References

1. Ding, S., Cook, D. R.: Matrix variate regressions and envelope models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). **80**(2), 387–408 (2018)
2. Govaert, G., Nadif M.: Co-clustering: models, algorithms and applications. John Wiley & Sons (2013)
3. Viroli, C.: On matrix-variate regression analysis. Journal of Multivariate Analysis. **111**, 296–309 (2012)
4. Woolf, A., Ansley, L., Bidgood, P.: Grouping of Decathlon Disciplines. Journal of Quantitative Analysis in Sports. **4**(3) (2007)

Classification of the population dynamics

Classificazione delle dinamiche di popolazione

Federico Bacchi[1] and Laura Neri[2]

Abstract Italy's population is chronically shrinking, especially at local level. An original classification method, based on the demographic trend observed during the last seventy years, is proposed to classify 7914 Municipalities. Considering the population dynamic as a Markov Chain, a division into groups is proposed. As a result of a series of statistical analysis, we observe that the earthquake risk and the quality of internet connection are connected with population dynamics:

Abstract *Questo lavoro propone un metodo di classificazione originale, per classificare 7914 Comuni, basato sulle tendenze demografiche osservate negli ultimi settanta anni. Considerando la dinamica della popolazione come una catena di Markov, si propone una suddivisione in gruppi distinti. Come risultato di una serie di analisi statistiche, si osserva che il rischio sismico e la qualità della connessione a internet sono caratteristiche associate con le dinamiche demografiche.*

Key words: demographic trend, Markov Chain, population dynamics.

1 Introduction

Population declines and depopulation associated with strong out-migration of some specific geographic areas is attracting the attention of both researchers and policy makers. Over the last decades, different demographic trends have been observed across European countries, especially at local level [3].

Italy has shown a gradual growth of its population, but this was not uniformly distributed throughout time and space. Particularly after the Second World War, Municipalities have recorded different patterns of population growth. It might be interesting to understand if there exists a link between those patterns and other specific characteristics. To address this issue, we follow two steps. Firstly, focusing

[1] Federico Bacchi, Alma Mater Studiorum; email: federico.bacchi3@unibo.it

[2] Laura Neri, Università degli Studi di Siena; email: laura.neri@unisi.it

Acknowledgment: F. Bacchi is supported by funds FSE REACT-EU - A.A. 2021/2022.

on demographic data, we classify the Italian Municipalities into groups showing similar demographic trends. None of the classic clustering methods are useful for the purpose, so, it is necessary to come up with original solutions. The second step concerns the study of the characteristics of the groups, complementing the demographic figures with data from other sources: geographical, socio-demographic and institutional variables.
Even though the topic is becoming increasingly popular, there are only few scientific publications about the relationship between population dynamics and other features of the geographic areas and, to the best of our knowledge, most of the contributions available are focused on continents or nations while only few specific analysis are conducted at Municipality level [4].
For these reasons, we get more in-depth on this topic.

2 Data

To analyse the dynamics of the Italian municipalities' population over the past 70 years, census data about population have been downloaded from the National Statistical Institute (ISTAT) website. The target variable to evaluate the phenomenon of depopulation is the resident population at municipality level; such information was historically collected by the Population Census. The first one was in 1951 and, until 2011, the census was conducted every ten years. Since 2018, ISTAT has started the so called "permanent census" which allows to collect data yearly. Thus, data on the resident population at municipality level, is available for the years 1951, 1961, 1971, 1981, 1991, 2001, 2011, 2018, 2019. We construct a dataset composed of 7914 rows and 10 columns: each row stands for each Italian municipality, at 31/12/2019[1]; the first column identifies the municipality code; columns 2 to 10 represent the resident population in each census time. This dataset allows classifying the units into clusters, representing a common observed pattern in demographic trend.

3 Classification methods

To identify clusters of Italian Municipalities based on their population's patterns, we propose two different methods. Both are based on the following idea: eight indicators, specified as dummy variables – one for each census couple of years – whose value is equal to 1 when the resident population of the municipality has increased, 0 otherwise. Thus, the data structure for the next analysis is the 7914 x 9 matrix for which the first three rows are reported in Table 1.

[1] The time series of the resident population (1951-2019) is based on the existing municipalities at 31/12/2019 (https://esploradati.censimentopopolazione.istat.it/databrowser/#/it).

Table 1: Few rows of the dataset containing indicators for the increasing or decreasing population for the census couple of years: Italian Municipalities.

Code	*51_61*	*61_71*	*71_81*	*81_91*	*91_01*	*01_11*	*11_18*	*18_19*
001001	0	1	0	1	0	1	0	0
001002	0	1	1	1	1	1	0	0
001003	0	0	0	1	0	0	0	0

3.1 Sum of dummy variables

The simplest way to induce if a municipality has shown an increasing or a decreasing demographic trend is to compute a new variable (Growth1), counting in how many periods its population has grown. Empirically, this is equivalent to sum all the dummy variables defined previously.
Now, it is possible to classify the 7914 Italian municipalities into different clusters, that can be defined according to two different options: i) deciding the number of clusters; ii) or the dimensions of these ones. Here we choose option (i) considering the classification according to 5 or 3 clusters. Thus, two new variables - Trend1 and Trend1bis - are introduced, to identify respectively the grouping into five or three clusters, as shown in Table 2.

Table 2: Values of "Growth1" and corresponding groups.

Growth1	*Trend1*	*Trend1bis*
0	Decreasing	Decreasing
1	Decreasing	Decreasing
2	Tendentially Decreasing	Decreasing
3	Tendentially Decreasing	Alternate
4	Alternate	Alternate
5	Tendentially Increasing	Alternate
6	Tendentially Increasing	Increasing
7	Increasing	Increasing
8	Increasing	Increasing

According to the previous method, the variable Growth1 assumes value 2 if the population growth in two subsequent or detached periods. It is strong assumption, based on the hypothesis that all the time points have the same importance in the definition of the demographic trends and, consequently, on the clustering process.

3.2 Population dynamics as a Markov Chain

Aiming at removing the overmentioned assumption, we wonder whether it seems reasonable to think that the dynamic of the population in *(t; t+1)* depends only on

the dynamic in *(t-1; t)*. In probability theory, a sequence of possible events in which the probability of each event depends only on the state attained in the previous event is the Markov Chain. Thus, we wonder if the population dynamics observed between 1951 and 2019 on the Italian municipalities could be represented as a Markov Chain. First, it is necessary to verify if the Markov property holds for the variables describing the population trends. The Chi-square statistics is close to 0, thus the dynamic of the population across the nine periods can be treated as a Markov chain. In general, any Markov chain is characterized by the "transition matrix" describing the probabilities of transitioning from one state to another. We address the problem of the estimation of the transition matrix by using the maximum likelihood method. In Table 3 the transition matrix (by rows) estimated and the standard errors.

Table 3: ML estimate of the transition matrix: estimated probabilities and SE (in parenthesis).

		Future State	
		0	***1***
Present	***0***	0.777 (0.005)	0.223 (0.003)
State	***1***	0.297 (0.003)	0.703 (0.005)

The practical meaning of the transition probability is very important for our purpose, in fact, for instance, the figure 0.703 represents the probability of an increase in population in the next period, given an increase in the present period. Now we can compute the vector of the probability to observe "1" in future, given "1" in the present state as

$\boldsymbol{w} = ((0.703)^7,(0.703)^6,(0.703)^5,(0.703)^4,(0.703)^3,(0.703)^2,(0.703)^1,(0.703)^0)$.

Now, let us consider the matrix M obtained from the one shown in Table 1 deleting the first column, M has dimension 7914x8 and each row m_i. *(i = 1,..,7914)* contains the indicators for the increasing/decreasing of population for each census couple of years. The variable "Growth2", for each municipality, can be computed as the vectorial product, $\boldsymbol{t}_i \times \boldsymbol{w}^T$. According to "Growth2" we can define a new classification of the Italian municipalities. In Table 4 the main quantiles of "Growth2" are shown.

Table 4: Main quantiles of "Growth2".

Min	***20%***	***25%***	***33,3%***	***40%***	***Median***	***60%***	***66,7%***	***75%***	***80%***	***Max***
0	0.206	0.348	0.622	0.842	1.117	1.442	1.666	2.047	2.168	3.168

Now, we group the municipalities, again into three or five clusters, considering new thresholds for the groups definition based on Growth2's quantiles (Table 5).

Table 5: Bounds of "Growth2" and groups according to Trend2 and Trend2bis.

Lower	***Upper***	***Trend2***	***Lower***	***Upper***	***Trend2bis***
[0	0.206)	Decreasing	[0	0.622)	Decreasing
[0.206	0.842)	Tendentially Decreasing	[0.622	1.666)	Alternate
[0.842	1,442)	Alternate	[1.666	3.168]	Increasing
[1.442	2.168)	Tendentially Increasing			
[2.168	3.168]	Increasing			

4 Empirical Analysis

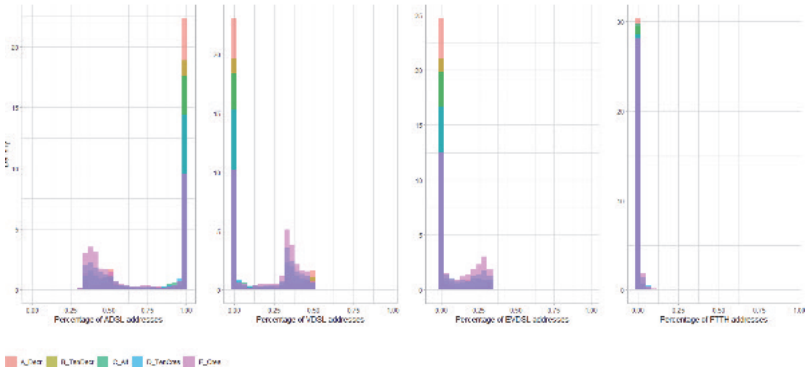
The next empirical analysis will concern the clustering method based on the Markov chain. Following are just a few of the analyses carried out. First of all, we check how the municipality classified by the Dipartimento per lo Sviluppo e la Coesione Economica, as Centro ("Polo", "Polo Intercomunale", "Cintura") or Area Interna ("Intermedio", "Periferia", "Ultraperiferia") are in turn classified according to our groups. The results appear to be reasonable: the percentage of "Decreasing" Municipalities increases as the distance of the Municipality from the "Centro" increases and, on the other hand, the percentage of "Increasing" or "Alternate" ones becomes higher as the distance of the Municipality from the "Centro" decreases.
We can expect that geological and physical characteristics of an area may influence population dynamics, so it seems reasonable to suppose an association between the clustering variable and the seismic risk index. For this reason, a discrete variable representing the Earthquake risk - defined by Protezione Civile – is assigned to each Municipality a seismic risk zone: where "Zone 1" is the most dangerous and "Zone 4" is the least dangerous area. Tables 6 suggests that the seismic risk influences population dynamics. Indeed, most of the statistical units with higher levels of earthquake risk have shown a decreasing (or tendentially decreasing) trend, while most of municipalities with lower levels of earthquake risk have shown an increasing (or tendentially increasing) trend. The results are coherent with the observations depicted by the literature, see for instance [1] that observe a general framework of demographic distress for areas most affected by the seismic events.

Table 6: Cross tabulation of "Trend2" and seismic risk index (percentage).

	Seismic Risk			
Trend2	***Zone 1***	***Zone 2***	***Zone 3***	***Zone 4***
Decr.	39.34	27.93	11.88	18.07
Tend. Dec.	31.33	17.60	13.73	20.58
Alternate	16.74	20.31	25.32	22.22
Tend. Inc.	8.44	16.03	20.94	17.37
Increas.	4.15	18.13	28.14	21.75
Total	100.0	100.0	100.0	100.0

Finally, it seems reasonable to suppose that geographic areas having better infrastructure and services may attract more people [2]. Nowadays, the quality of internet connection could be considered as an indicator of the quality of life. Accordingly, data from AGCOM website (year 2019), on inhabitants classified by type home connection are considered. The four types of connection, listed from the worst to the best, are: ADSL, VDSL, EVDSL, FTTH. The frequency polygon of the percentage of different home connections (Figure 1) show that better connections are more likely to be present for municipalities presenting increasing demographic trends.

Figure 1: Frequency polygon of the % of connections across the level of "Trend2".



5 Conclusions

Original classifications methods have been adopted to classify the Italian municipalities into groups of statistical units showing a similar demographic trend during the last seventy years. To avoid the implicit assumption that all the past time points have the same influence on the future trend, we introduced a vector of weights based on the estimated transition matrix of a Markov Chain and, we divided the 7914 units into five or three groups. This classification seems to be coherent with the official one. Considering several characteristics of the units, we observe that the ones presenting a stronger association with demographic trends are the home connection quality and the earthquake risk.

The analysis presented in this work do not aim to be exhaustive. Indeed, it would be interesting to consider fuzzy indicators, instead of dummy variables for intercensal trends, and to extend the set of variables which may influence depopulation. Moreover, it might be interesting to build a predictive statistical model for future demographic trend based on specific characteristics of the units.

References

1. De Lucia, M., Benassi, F., Meroni, F., Musacchio, G., Pino, M., Strozza S.: Seismic disasters and the demographic perspective: 1968, Belice and 1980, Irpinia-Basilicata (southern Italy) case studies. Annals of geophysics, Vol. 63, (2020).
2. ESPON (2017) Shrinking rural regions in Europe. https://www.espon.eu/sites/default/files/attachments/ESPON%20Policy%20Brief%20on%20Shrinking%20Rural%20Regions.pdf
3. EU-Committee of the Regions (2016) The impact of demographic change on European regions. https://cor.europa.eu/en/engage/studies/Documents/The%20impact%20of%20demographic%20change%20on%20European%20regions/Impact_demographic_change_european_regions.pdf.
4. Reynaud, C., Miccoli, S., Benassi, F., Naccarato, A., & Salvati, L. (2020). Unravelling a demographic 'Mosaic': Spatial patterns and contextual factors of depopulation in Italian Municipalities, 1981–2011. Ecological Indicators, 115, 106356.

Locating γ-Ray Sources on the Celestial Sphere via Modal Clustering

Individuazione di sorgenti di raggi γ sulla sfera mediante clustering non parametrico

Anna Montin, Alessandra R. Brazzale, Giovanna Menardi

Abstract Searching for as yet undetected γ-ray sources is a major target of the Fermi LAT Collaboration. In this paper, we present an algorithm capable to identify such type of sources by non-parametrically clustering the directions of arrival of the high-energy photons detected by the telescope. Using statistical tools from hypothesis testing and classification, we furthermore present an automatic way to skim off sound candidate sources from the γ-ray emitting diffuse background and to quantify their significance. The algorithm was calibrated on simulated data provided by the Fermi LAT collaboration and will be illustrated on a real Fermi LAT case-study.

Abstract *L'individuazione di sorgenti di raggi gamma è uno degli obiettivi dichiarati della Collaborazione Fermi LAT. Presentiamo qui un algoritmo per l'individuazione di queste sorgenti basato sul clustering non parametrico delle direzioni di arrivo dei fotoni ad alta energia rilevate dal telescopio. Sfruttando risultati della teoria dei test statistici e dell'apprendimento supervisionato, presentiamo, inoltre, come scremare le sorgenti candidate dalla componente diffusa della radiazione di fondo e attribuire loro una misura della significatività. L'algoritmo è stato calibrato su dati simulati forniti dalla Collaborazione Fermi LAT e sarà illustrato su un caso di studio reale.*

Key words: directional data, kernel density estimator, man-shift algorithm, tree-based classification

---

Anna Montin
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: anna.montin@studenti.unipd.it

Alessandra R. Brazzale
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, e-mail: alessandra.brazzale@unipd.it

Giovanna Menardi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: giovanna.menardi@unipd.it

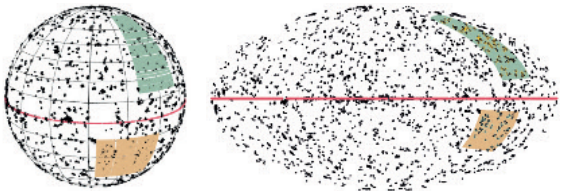


Fig. 1 Fermi-LAT γ-ray photon count maps for a 5-year observation period. Left: in polar coordinates. Right: in Galactic coordinates. Yellow: analysed region. Green: training set for post-processing classifier. Red: Galactic plane.

1 Motivation and rationale

In γ-ray astronomy, the data typically consist of an event list which gives the direction in the sky of each detected photon together with additional information. If the distance to the emitting source is not relevant, the data points are placed on the celestial sphere with Earth at its center and unit radius, as shown in the left panel of Figure 1. Directions are often expressed in *galactic coordinates*, which place the origin of the Cartesian system in the center of our galaxy — the Milky Way — and align the x-axis with the galactic plane (right panel of Figure 1). To overcome mismatches due to projecting data on the 2-dimensional sky map, we rather express directions through *polar coordinates*, that is, co-latitude (θ) and longitude (ϕ) in geographical terms, which can easily be back-transformed to Cartesian coordinates $\mathbf{x} = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)^\top$ on the unit sphere.

Discovering and locating high-energy emitting sources in the whole sky map is a declared target of the Fermi Gamma-ray Space Telescope collaboration. An astronomical source is an object in outer space which, in our case, emits γ-ray photons, that is, quanta of light in the highest energy range. Traditionally, analyses are based on so-called *single-source models* [4, § 7.4], which require the whole sky map to be split into small regions. The presence of a possible new source is assessed on a pixel-by-pixel basis using Poisson regression and likelihood ratio testing. Conversely, *variable-source-number models* address the problem from a more global perspective, as they simultaneously model and locate all sources in a sky map [4, § 7.3]. A most recent example of application to the γ-ray count maps accumulated by the LAT — the principal scientific instrument on board the Fermi spacecraft — is [7].

In this paper, we present a flexible algorithm for the efficient identification of γ-ray sources. In particular, we address the problem from the global perspective of variable-source-number models while working on the sphere. From the modeling point of view, the sources will be represented by highly concentrated clusters. [2] provide an illustration of directional *model-based* clustering of Fermi LAT data using a finite mixture of von Mises-Fisher distributions. Our approach uses *modal clustering*, which combines the advantages of both, model-based clustering and non-parametric methods, to guarantee the required flexibility. The corresponding methodological background is reviewed in Section 2. We will illustrate our proposal through a case-study of Fermi LAT data in Section 3.

2 Modal clustering on the sphere

General framework. Allocating objects to an unknown number of groups according to a set of observed attributes or features is a natural activity of any science. A surge of techniques has been proposed over the years, which differ significantly in their definition of what a "group" is. The non-parametric formulation, referred to as *modal* clustering, associates clusters with the domain of attraction of the modes of the underlying density, which are usually estimated non-parametrically [6]. Modal clustering can be recast into the frame of a standard statistical problem by considering the observed data $(\mathbf{x}_1,\dots,\mathbf{x}_n)$ as a sample of n realisations of a d-dimensional random vector $\mathbf{X}$ from a probability density function $f:\mathscr{X}\subseteq\mathbb{R}^d\to\mathbb{R}^+$. The modes of $f(\cdot)$ represent the archetypes of the clusters, which are in turn described by the surrounding regions.

Mode hunting. In our setting, directions in $\mathbb{R}^3$ are represented as unit vectors $\mathbf{x}$, that is, as points on the sphere $\Omega^2=\{\mathbf{x}\in\mathbb{R}^3:||\mathbf{x}||_2=x_1^2+x_2^2+x_3^2=1\}$ with unit radius and centre at the origin. Density estimation is performed by a suitable extension of a kernel density estimator to the directional setting [1]

$$\hat{f}_n(\mathbf{x})=\frac{c_h(K)}{n}\sum_{i=1}^{n}K\left(\frac{1-\mathbf{x}^\top\mathbf{x}_i}{h^2}\right), \tag{1}$$

where $K(\cdot)$ is a kernel function, $c_h(K)$ the associated normalizing constant, and the bandwidth parameter $h>0$ controls the smoothness of the estimator.

To account for the rugged nature of the data, which exhibit highly heterogeneous levels of concentration over the sphere, the estimator (1) is extended to account for a variable bandwidth $h=h(\mathbf{x}_i)$, selected according to scientific input. In particular, we use the scale parameter of the *point spread function*, which describes the response of the LAT to the point source. Grossly, this amounts to associating smaller values of h with precise events characterized by higher energies and which are usually disclosed around the direction of photon emission. The choice is consistent with the requirement of reducing the amount of smoothing nearby the high-density regions, as it is usually acknowledged by variable kernel density estimators.

A popular choice for the directional kernel is linked to the von Mises-Fisher (vMF) distribution [5]

$$f_{\text{vMF}}(\mathbf{x};\mu,\kappa)=c_3(\kappa)e^{\kappa\mathbf{x}^\top\mu},$$

where c_3 is a normalizing constant, $\mu\in\Omega^2$ is the mean direction, and κ is a concentration parameter around the mean. Here, κ is set to vary inversely with the bandwidth, i.e. $K(\cdot)=f_{\text{vMF}}(\cdot;\mathbf{x}_i,1/h)$. This distribution describes observations which scatter symmetrically around their mean value and can be regarded as the generalization of the normal distribution to spherical data.

Sources are aimed to be identified by pursuing the explicit task of mode detection, and modal regions are formed by sets of points along the steepest ascent path towards a mode. This is achieved by adapting the *mean-shift* algorithm [6, § 2.2] to be used with the directional kernel estimator (1). Starting from a generic point

$\mathbf{x}^{(0)}$, the algorithm recursively shifts it to a local weighted mean, until convergence. Denoted by $w_i(\mathbf{x}^{(s)})$ the vector of weights of the components of $\mathbf{x}_i$ at step s, at the next step

$$\mathbf{x}^{(s+1)} = \sum_{i=1}^{n} w_i(\mathbf{x}^{(s)})\mathbf{x}_i = \mathbf{x}^{(s)} + M(\mathbf{x}^{(s)}),$$

where $M(\mathbf{x}^{(s)}) = \sum_{i=1}^{n} w_i(\mathbf{x}^{(s)})\mathbf{x}_i - \mathbf{x}^{(s)}$ denotes the mean shift. Up to a normalising factor, the weights $w_i(\mathbf{x})$ are specified as $\nabla K\left(h^{-2}(1-\mathbf{x}^{\top}\mathbf{x}_i)\right)$, where $\nabla K(\cdot)$ is the gradient of the kernel function.

Post-processing. To separate the signal of the supposed emitting source from the diffuse γ-ray background, which spreads over the entire area observed by the telescope, we propose a post-processing procedure that combines the findings of two parallel quests.

On one hand, we supervise a suitable classifier, based on a training sample drawn from the available LAT catalogue, for which information on the earlier detected sources is available. The classifier integrates additional information on the photons such as their energy, position, the number present in the same cluster, the density estimates for the signal and the background model and various types of distances to the detected mode. This allows us to skim off the photons emitted from high-energy emitting sources from those which originate from the diffuse background.

In parallel, we evaluate the significance of each candidate mode. Here, we consider an adaptation of [3] who verify whether the maximum eigenvalue of the Hessian matrix of the kernel density, evaluated at the mode, is negative. In the directional setting, due to the constraints induced by working on the surface of the sphere, one eigenvalue is necessarily zero. An $1-\alpha$ level confidence interval for the second largest eigenvalue is hence constructed using bootstrap resampling. The mode is considered as such if the interval includes only negative values. The same allows us to infer the significance of the candidate source.

By super-imposing the findings from the two quests, we can identify candidate sources which are both, statistically significant and qualified as such by the non-parametric classifier. A by-product of our post-processing algorithm is the assignment of the photons to their assumed emitting source.

3 A Fermi LAT case study

Mode hunting. The two yellow regions in Figure 1 show a portion of the Southern sky of size $(l,b) \in [95^\circ, 135^\circ] \times [-40^\circ, -10^\circ]$ for which the LAT accumulated 3,849 photon counts over a five-year period of observation.[1] Of these, about 26% were emitted by the 44 sources present in the area, while the remaining 74% originated from the diffuse γ-ray background. The left panel of Figure 2 plots the estimated kernel density (1) using a von Mises-Fisher kernel. Here, the bandwidth parame-

[1] https://fermi.gsfc.nasa.gov/ssc/data/access/

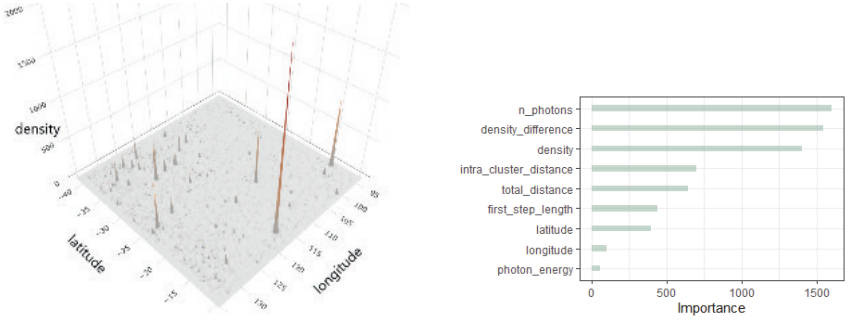


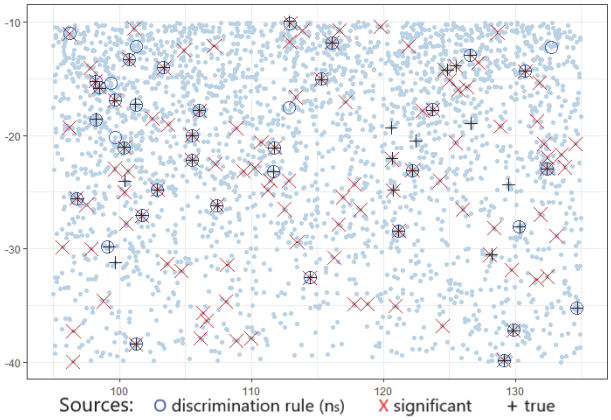
Fig. 2 Left: Kernel density estimate using a von Mises-Fisher kernel of the γ-ray photon counts accumulated by the LAT in a 5-year period. Right: Feature importance plot for the tree-based photon classifier.

ter h was set according to scientific input, as described in Section 2. This choice revealed to be the most performing one in terms of adjusted Rand index (ARI), average distance between the true source direction and the reconstructed one ($\bar{d}(s,\hat{s})$) and number of identified sources (n_s), according to an extensive numerical investigation (results not shown here) we carried out. In all, the mean-shift algorithm identified 876 modes.

Post-processing. To further refine the list of candidate sources we proceeded in two steps as outlined in Section 2.

1. A tree-based classifier to discriminate between source and background photons was trained on the 6,814 photon counts highlighted in green in the right panel of Figure 1 using as predictor variables those listed in the right panel of Figure 2. The most discriminating features are the number of photons assigned to a cluster, the difference between the two photon densities for, respectively, the all sky and background counts only, and the density observed for each photon. This reduces the original 876 modes to 39 candidate sources, which are shown as blue circles in the left panel of Figure 3. The table on the right reports the performance of our classifier in terms of ARI and average distance $\bar{d}(s,\hat{s})$. The true positive rate for single photon classification is 98.5% rate, while the percentage of false positives is 22.9%. Indeed, the five missed sources are the less photon emitting ones.
2. In parallel, we tested all the 555 clusters which contain two or more photons at a significance level of 5% while applying Bonferroni's correction. This skimmed off 448 modes, for a total of 107 remaining candidate sources, shown in the left panel of Figure 3 as red crosses. Here, the true positive rate for single photon classification is 85.0% and the false positive rate is 11.2%.

By super-imposing these two findings, we obtain in all 27 sources which are both, statistically significant and qualified as such by the non-parametric classifier. The global true positive rate for single photon classification is 94.6% while the false positive rate is 14.1%.



Results	
ARI	0.961
$\bar{d}(s,\hat{s})$	0.001
n_s	39
True sources	44

Fig. 3 Left: Fermi-LAT γ-ray photon count map (in Galactic coordinates) for the analysed 5-year observation period with superimposed the true and candidate sources. Right: Performance measures of the tree-based classifier.

To analyse the whole sky map, *consensus clustering* [8] may allow us to aggregate results from multiple runs on portions of the sphere whenever computational costs and limited memory won't allow us to do it in one go.

References

1. Bai Z.D., Rao C.R., Zhao L.C.: Kernel estimators of density function of directional data. Journal of Multivariate Analysis **27**, 24–39 (1988)
2. Costantin D., Menardi G., Brazzale A.R., Bastieri D., Fan J.H.: A novel approach for pre-filtering event sources using the von Mises-Fisher distribution. Astrophysics and Space Science **365** (2020)
3. Genovese C.R., Perone-Pacifico M., Verdinelli I., Wasserman L.: Non-parametric inference for density modes. Journal of the Royal Statistical Society Series B **78**, 99–126 (2016)
4. Hobson M.P., Jaffe A.H., Liddle A.R., Mukherjee P., Parkinson D.: Bayesian Methods in Cosmology. Cambridge University Press (2009).
5. Mardia K.V., Jupp P.E.: Directional Statistics. John Wiley & Sons (2000)
6. Menardi G.: A review on modal clustering. International Statistical Review **84**, 413–433 (2006)
7. Sottosanti A., Bernardi M., Brazzale A.R., Geringer-Sameth A., Stenning D.C., Trotta R., van Dyk D.A: Identification of high-energy astrophysical point sources via hierarchical Bayesian nonparametric clustering. arXiv:2104.11492 (2021)
8. Vega-Pons S., Ruiz-Shulcloper J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence **25**, 337–372 (2011)

Sampling and Official Statistics

Fisher's Noncentral Hypergeometric Distribution for Population Size Estimation

La distribuzione ipergeometrica noncentrale di Fisher per la stima di numerosit di popolazione

Veronica Ballerini and Brunero Liseo

Abstract Fisher's noncentral hypergeometric (FNCH) distribution naturally suits biased sampling processes. Indeed, this distribution describes a biased urn experiment where balls of different colors are associated with different weights. Despite its potentiality, FNCH distribution has never been applied to official statistics problems, such as the size estimation of heterogeneous populations. Such underuse is mainly due to the computational burden given by its probability mass function, which makes the evaluation of the likelihood function challenging. We present a methodology to estimate the posterior distribution of FNCH parameters, exploiting extra-experimental information and the computational efficiency of MCMC methods. We assess the robustness to weights prior specifications via simulation studies.
Abstract *La distribuzione ipergeometrica non centrale di Fisher (FNCH) si adatta naturalmente a situazioni in cui il processo di campionamento è affetto da distorsione. Tale distribuzione descrive un esperimento di urna distorto, in cui palline di colori diversi sono associate a pesi diversi. Nonostante le sue potenzialit, la FNCH non mai stata applicata a problemi di statistica ufficiale, come la stima della numerosit di popolazioni eterogenee. Tale sottoutilizzo principalmente dovuto all'onere computazionale dato dalla funzione di massa di probabilit, che rende non banale la valutazione della funzione di verosimiglianza. In questo lavoro, presentiamo una metodologia per stimare la distribuzione a posteriori dei parametri della FNCH, sfruttando informazioni extra-sperimentali e l'efficienza computazionale dei metodi MCMC. La robustezza alle diverse specificazioni delle a priori sui pesi valutata tramite studi di simulazione.*

Key words: Official statistics, MCMC, MNAR, Biased sampling

Veronica Ballerini
University of Florence, Piazza di San Marco, 4, 50121 Firenze. e-mail: veronica.ballerini@unifi.it

Brunero Liseo
Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Roma. e-mail: brunero.liseo@uniroma1.it

1 Fisher's noncentral hypergeometric distribution

In 2008, Agner Fog clarified the distinction between two distributions, both known in the literature as "the" noncentral hypergeometric distribution (see [7] and [6]). He first solved the nomenclature issue, naming them *Wallenius*' and *Fisher*'s, after the persons who first proposed them (see [5] and [11]).

The main difference between the two distributions resides in the dependence structure of the draws. Assume an urn of size N contains M_1 balls of color 1 and M_2 balls of color 2. The univariate Wallenius' noncentral hypergeometric distribution describes a situation in which the balls are drawn without replacement until n balls are sampled, and the probability to sample X_1 balls of color 1, and X_2 balls of color 2 depends on the colors' relative weights. Such weights can be seen as the probability to retain a ball of that color when drawn (as suggested by [4]). The competing draws make Wallenius' distribution fit perfectly the context of preference or ranking data (see [8]).

Instead, the univariate Fisher's noncentral hypergeometric (FNCH) distribution describes a biased urn experiment when the balls are drawn independently, without replacement, and the sample size n is observed only at the end of the experiment. It is the conditional distribution of two independent Binomial distributions given their sum ([9]):

$$\begin{aligned} X_1 &\sim Binom(M_1, \zeta_1) \\ X_2 &\sim Binom(M_2, \zeta_2) \end{aligned} \tag{1}$$

$$X_1 | X_1 + X_2 = n \sim FNCH(M_1, M_2, n, w_1, w_2) \tag{2}$$

where the weights w_c are the odds $\zeta_c/(1-\zeta_c)$, $c = 1, 2$. The weights w_c are defined up to a positive constant k; then, FNCH distribution is identified by the odds ratio $w = w_1/w_2$. Since $M_2 = N - M_1$ and $x_2 = n - x_1$, the formulation (2) is equivalent to:

$$X_1 | n \sim FNCH(M_1, N, n, w)\ . \tag{3}$$

Consider ζ_1, ζ_2 in (1) to be the two capture probabilities of the respective two population groups; then, the ratio w may be interpreted as the relative "exposure" of group 1 over group 2 in that capture occasion. As an alternative interpretation, we may consider ζ_1, ζ_2 as the two groups' "non-missing probabilities". In this case, w would be a measure of non-randomness in the missing data mechanism.

Given such considerations, it is clear that FNCH distribution has a great potentiality in the official statistics field. It can manage situations where n units belonging to a heterogeneous population of unknown size N are observed, and such units' capture, or non-missing, probabilities vary according to the population's subgroup they belong to. Such a situation is typical of survey data when the response rates of various groups differ according to their characteristics.

2 Bayesian inference for FNCH

We aim to estimate the size of the heterogeneous population N, or the subgroups' sizes M_1 and M_2. Whether both M_1 and M_2 are unknown, even if w is known and fixed, we can only estimate the relative size of the two groups in the population. Nevertheless, we might have some prior information on one of the groups: such a situation is common when dealing with administrative data. Indeed, consider a sample of resident (group 1) and non-resident (group 2) persons living in a city; given the reliable information contained in the municipal registries, including genuine prior information on M_1 would be legitimate. Moreover, assume data about a sample of self-employed individuals to be available, collecting information on their working situation. Self-employed workers may have (group 1) or have not (group 2) a VAT number. Since people with a VAT number must be listed in social security registers, we still may assume a reasonably concentrated prior on M_1. Finally, imagine we want to estimate the number of unemployed (group 1) and employed (group 2) young Italian graduates of a specific cohort; their respective sizes are unknown. Nevertheless, the National Student Register provides the annual total number of graduates, N; since the associated error is very small, we can elicit a concentrated prior for N.

The subjective elicitation is a debated issue since the attribute "subjective" is often perceived as including personal beliefs in a negative sense. Instead, we consider the elicitation process a rational way to incorporate experts' knowledge and take advantage of their experience; for a deep and detailed discussion about the probabilities' elicitation process, see [3] and [10].

Now, let us assume the following hierarchical model:

$$X_1 | X_1 + X_2 = n \sim FNCH(M_1, M_2, n, w) \tag{4}$$

$$M_1, M_2 \overset{\text{ind}}{\sim} \pi(M_c; \boldsymbol{\theta}^{M_c}) \, , c = 1, 2 \tag{5}$$

where $\pi(\cdot)$ denotes a generic distribution depending on some parameters $\boldsymbol{\theta}$.

If also w is unknown, we must elicit a prior on w:

$$w \sim \pi(w; \boldsymbol{\theta}^w) \, . \tag{6}$$

To use a noninformative prior on w is mathematically legitimate. However, we must notice that this parameter plays a crucial role which is often peculiar to the specific data set, and the use of genuine extra-experimental information is strongly suggested here. In the next section, we assess the sensitivity of the posterior to different specifications of w.

We denote with $\pi(M_1, M_2, w)$ the joint prior distribution; it can be factorized into $\pi(M_1)\pi(M_2)\pi(w)$ assuming that the odds ratio for the two groups of being included in the sample is independent on the groups' sizes. We also denote with $L(x_1; M_1, M_2, n, w)$ the likelihood function. Hence, the joint posterior distribution is

$$\pi(M_1, M_2, w | x_1, n) \propto L(x_1; M_1, M_2, n, w)\pi(M_1)\pi(M_2)\pi(w) \, . \tag{7}$$

Table 1: Sensitivity of M_2 to different prior specifications of w. Average variation between the posterior mean of M_2 and the value used to simulate the data, and average standard deviation, estimated on 50 samples for $N = 1000$.

σ_w	Average B_{M2}	Average $sd(\bar{M}_2^*)$
$\sigma_w = 0.2$	0.15	183.2
$\sigma_w = 0.4$	0.17	198.4
$\sigma_w = 1$	0.22	228.3

The posterior in (7) can be easily computed via MCMC methods, e.g., using a Metropolis-within-Gibbs algorithm.

3 Simulation studies

This section shows how the posterior of the subgroups' size changes as the prior specified for the odds becomes wider.

We simulate 50 samples setting $n = 50$ and $w = 2$ for two different population sizes, i.e. $N = 1000$ and $N = 10000$. We assume a very concentrated prior for M_1, namely a Poisson centered on its simulated value, and a discrete Uniform prior distribution for M_2, defined between x_2 and a large upper bound U_{M_2}; we set U_{M_2} equal to 5000 and 15000 for the two population sizes, respectively. Finally, for w we assume a Log-normal prior with parameter $\mu = 2$, and let the standard deviation σ_w vary; see Table 1 and Table 2.

For the two population sizes, Tables 1 and 2 show the average variation between the M_2 posterior and the value we used to simulate data, and the average standard deviation of the M_2 posterior. For each sample, we define our measure of variation as

$$B_{M_2} = \frac{1}{D} \sum_{d=1}^{D} \frac{(M_{2,d} - M_2)}{M_2} \tag{8}$$

where $M_{2,d}$ is d-th draw from the M_2 posterior.

For both population sizes, the mean standard deviation increases as the prior on w becomes wider; moreover, introducing more uncertainty leads to an increasing (upward) mean bias for the posterior mean. This bias is due to the large upper bound we elicit for M_2 prior.

Figures 1 and 2 shows how, however, the posterior is way more concentrated on the true M_2 value than the dashed flat prior (one sample only).

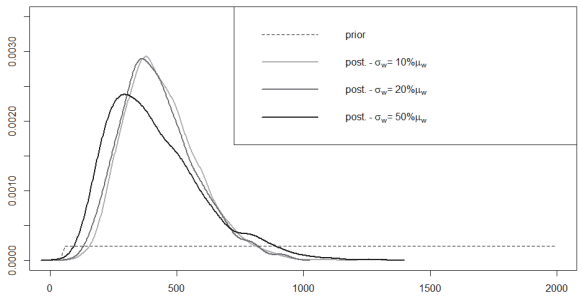


Fig. 1: M_2 prior vs. posteriors with different prior specifications of w. $N = 1000$.

Table 2: Sensitivity of M_2 to different prior specifications of w. Average variation between the posterior mean of M_2 and the value used to simulate the data, and average standard deviation, estimated on 50 samples for $N = 10000$.

σ_w	Average B_{M2}	Average $sd(\bar{M}_2^*)$
$\sigma_w = 0.2$	0.13	1774.2
$\sigma_w = 0.4$	0.17	2051.0
$\sigma_w = 1$	0.29	2818.4

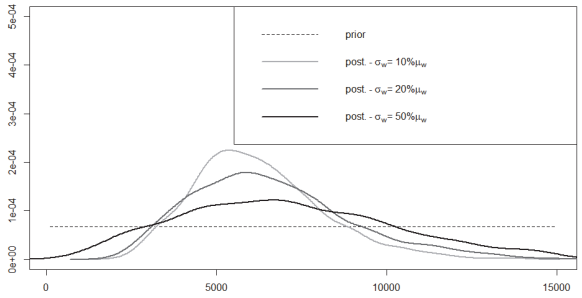


Fig. 2: M_2 prior vs. posteriors with different prior specifications of w. $N = 10000$.

4 Conclusions

Undercoverage is a pervasive issue, mainly addressed in the capture-recapture framework. Although widespread, the single-list case does not boast such vast literature, especially when the target population is elusive — e.g., non-resident inhabitants in a big city, homeless people, irregular migrants. We addressed the problem of

size estimation of a heterogeneous population when a single list is available, or we have multiple lists, but we lack unique identifiers. Our model considers the different sources' reliability and the different units' propensity to be captured. Genuine prior information is needed to manage such sources of uncertainty together; a Bayesian model comes naturally in such a situation.

In this short paper, we only discuss the use of our methodology in a simulation study. In [1], an application to the problem of size estimation of unemployed young graduates is presented. The authors exploit the accurate information on the total number of graduates made available by the National Student Register of the Italian Ministry of University and Research. We are currently working to extend the model to the multivariate case. In such a case, the repeated evaluation of the likelihood is prohibitive; thus, a vanilla MCMC algorithm is computationally intractable. As a result, it is needed to use algorithms that avoid the direct evaluation of the FNCH likelihood function. A viable approach is then based on Approximate Bayesian Computation methods, as discussed in [1] and to a more general extent in [2].

References

1. Ballerini, V., Liseo, B.: Fisher's noncentral hypergeometric distribution for the size estimation of unemployed graduates in Italy. Working paper MEMOTEF Sapienza - available upon request (2021)
2. Beaumont, M. A.: Approximate Bayesian Computation. Annual review of statistics and its application **6**, 379-403 (2019)
3. Berger, J. O.: Statistical decision theory and Bayesian analysis. Second Edition. Springer series in statistics, Springer-Verlag (1985)
4. Chesson, J.: A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. Journal of Applied Probability **13**, 4, 795–797 (1976)
5. Fisher, R. A.: The logic of inductive inference. Journal of the Royal Statistical Society **98**, 1, 39–82 (1935)
6. Fog, A.: Calculation methods for Wallenius' noncentral hypergeometric distribution. Communications in StatisticsSimulation and Computation **37**, 2, 258–273 (2008)
7. Fog, A.: Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. Communications in StatisticsSimulation and Computation **37**, 2, 241–257 (2008)
8. Grazian, C., Leisen, F., Liseo, B.: Modelling preference data with the Wallenius distribution. Journal of the Royal Statistical Society: Series A (Statistics in Society) **182**, 2, 541–558 (2019)
9. Harkness, W. L.: Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics **36**, 3, 938–945 (1965)
10. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., Rakow, T.: Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons (2006)
11. Wallenius, K. T.: Biased sampling; the noncentral hypergeometric probability distribution. Stanford University CA Applied Mathematics and Statistics Labs (1963)

Small area models for skew and kurtotic distributions

Modelli per piccole aree riferiti a distribuzioni con asimmetria e curtosi

Maria Rosaria Ferrante and Lorenzo Mori

Abstract Models for small area estimation, used in the economic field, often have the necessity to assume or to lead back through appropriate transformations to the normality of the dependent variable. This work tries to extend the well known methodology of small area estimation at unit level with a family of models (Generalized additive model for location, scale and shape), which do not need the assumption of normality but which, as special case, has in itself most of the uni-variate models presented in literature.

Abstract *I modelli di stima per piccole aree, utilizzati in ambito economico, hanno spesso l'esigenza di assumere, o di ricondursi attraverso opportune trasformazioni, alla normalità della variabile dipendente. In questo lavoro si cerca di estendere la metodologia di stima per piccole aree a livello unit ad una famiglia di modelli (Generalized additive model for location, scale and shape), la quale non necessita dell'assunzione di normalità ma che, come caso particolare, comprende la maggior parte dei modelli univariati utilizzati nell'ambito della stima per piccole aree.*

Key words: Small area estimation, unit level models, GAMLSS, exponential family

1 Introduction

Small area methods attempt to solve the low representativeness of surveys within areas or the lack of data for specific areas/sub-populations. This is accomplished by incorporating information from outside sources. Unit-level models, originally defined in the form of nested error linear regression model [10], were generalized with few more complex and flexible models. The fil-rouge that links all those models is

Maria Rosaria Ferrante and Lorenzo Mori
Department of statistical science P. Fortunati , Via Belle Arti, 41 - Bologna - Italy e-mail: maria.ferrante@unibo.it and e-mail: lorenzo.mori7@unibo.it

the attempt to release assumptions, in particular the normality one, present in the basic unit level model. Many economic continuous variables, are distributed on the positive real line and can be significantly skewed. The standard technique of transforming variables by means of log is not always helpful to achieve a normal distribution. In the field of the frequentist approaches, generalized linear (mixed) models (GLMM) with random area effects were proposed and they are well summarized in [5]. In [6] it was proposed a generalized additive (mixed) models (GAMM) approaches to SAE. In the filed of the Bayesian approaches a skew normal distribution was used in [2] and a multivariate beta regression in [13]. Following the time line of the models development, speaking out of small area contest, a generalization of GLM, GAM, GLMM and GAMM is the generalized additive models for location, scale and shape (GAMLSS). Rigby and Statiunopolus [11] introduced these models where the response distribution is not restricted to belong to the exponential family. Those models, at the best of our knowledge, were never been used for Small Area Estimation (SAE). In our opinion, they can be not only a generalization of the models presented in literature but, above all, the possibility to define each parameter of a distribution in terms of covariates and random effects could bring to a reduction of the mean square error.

2 Generalized additive model for location, scale and shape

GAMLSS models, assume independent observations y_i, $i = 1,\ldots,n$ from a random variable Y_i, with probability density function $f(y_i|\theta^i)$, conditional on $\theta^i = (\theta_{i1},\ldots,\theta_{ip})$ a vector of p distribution parameters[1], each of which can be a function to the explanatory variables. Rigby and Statiunopolus [11] define the original formulation of a GAMLSS model as follows. Let $y_i^T = y_1,\ldots,y_n$ be the n length vector of the response variable. Also for $k = 1,\ldots,p$, let $g_k(\cdot)$ be known monotonic link functions relating the distribution parameters to explanatory variables by:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{j_k}\gamma_{j_k} \tag{1}$$

where θ_k and η_k are vectors of length n, e.g. $\theta_k^T = (\theta_{1_k},\ldots,\theta_{n_k})$ and $\beta_k^T = (\beta_{1_k},\ldots,\beta_{J_k'k})$ is a parameter vector of length J_k', X_k is a known design matrix of order $n \times J_k'$, Z_{j_k} is a fixed known $n \times J_k$ design matrix and γ_{j_k} is a J_k-dimensional random variable. The population probability (density) function $f(y|\theta^i)$ in model 1 is deliberately left general with no explicit conditional distributional form for the response variable. In GAMLSS, in fact, the exponential family distribution assump-

[1] The first two population distribution parameters $(\theta_{i1},\theta_{i2}) = (\mu_i,\sigma_i)$ are usually characterized as location and scale parameters, while the remaining parameters, if any, are characterized as shape parameters, e.g., skewness and kurtosis parameters. For many families of population distributions a maximum of two shape parameters $\nu(=\theta_3)$ and $\tau(=\theta_4)$ suffice.

tion for the response variable, y, (essential in GLMs and GAMs), is relaxed and replaced by a general distribution family, including highly skew and kurtotic distributions[2].

GAMLSS allows a variety of smooth functions[3] ($\sum_{j=1}^{J_k} Z_{j_k}\gamma_{j_k}$) of explanatory variables but, given that our interest is on small area estimation techniques, we will refer to γ_{j_k} always as parametric random-effect.

2.1 GAMLSS for small area estimation

In SAE a finite population Ω with N units is distributed into J sub-population $\Omega_1,\ldots,\Omega_J$, called domains or areas, of size $N_1,\ldots,N_J$, where $N=\sum_{\omega=1}^{\Omega} N_J$. In this case, Y_{ji} is the measurement of the target variable in unit i form area j. The main aim is to estimate area parameters in the form $H_j = h(Y_j)$, $j=1,\ldots,J$, where $h(\cdot)$ is a real measurable function. These area parameters are estimated using the data coming from the sample s of size n drawn from the population Ω with sub-sample $s_j = s\cap\Omega_j$ from area s of size n_j with $n=\sum_{j=1}^{J} n_j$. We will denote by $r_j=\Omega_d - s_j$ the sample complement from the same area, $j=1,\ldots,J$. Following [12], we can define the following model using area random effects:

$$\begin{aligned} g_1(\mu) &= \eta_1 = X_1\beta_1 + Z_1\gamma_{j1} \\ g_2(\sigma) &= \eta_2 = X_2\beta_2 + Z_2\gamma_{j2} \\ g_3(\nu) &= \eta_3 = X_3\beta_3 + Z_3\gamma_{j3} \\ g_4(\tau) &= \eta_4 = X_4\beta_4 + Z_4\gamma_{j4} \end{aligned}$$

where the $Z's$ can now be complex design matrices depending on the number and type of random effects and contains area level covariates. With $\gamma_k \overset{iid}{\sim} N(\mathbf{0},\Psi_k)$ for $k=1,2,3,4$. The Ψ's are matrices associated with the variance-covariance structure of the random effects. Random effects are also supposed to be independent among themselves. Hence considering a distribution model, called $\mathscr{F}$, the conditional distribution of Y, by including a random effect term in each of the models for the parameters μ,σ,ν and τ is:

$$Y_{ij}|\gamma_{j1},\gamma_{j2},\gamma_{j3},\gamma_{j4} \sim \mathscr{F}(\mu_{ij},\sigma_{ij},\nu_{ij},\tau_{ij}) \tag{2}$$

Independently for $j=1,\ldots,J$ and $k=1,2,3,4$. With a series of arguments reported in [11] it is possible to conclude that the β_ks and the γ_{j_k}s are estimated within the GAMLSS framework by maximizing a penalized likelihood function l_p given by:

[2] For a list of the possible distributions: https://www.gamlss.com/

[3] For more details: Statiunopolus et *al.* [12], Ch 10.

$$l_p = \sum_{i=1}^{n} \log\{f(y_i|\theta^i)\} - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \gamma_{j_k}^T \Psi_{j_k}^{-1} \gamma_{j_k}, \quad \text{given } \theta^i \; \forall \; i = 1,\ldots,n \qquad (3)$$

The model fitting [4] of a GAMLSS is achieved by either of two different algorithmic procedures which maximizing the penalized likelihood of the data eq. 3. The first algorithm (RS) is based on the algorithm that was used for the fitting of the mean and dispersion additive models (based on Newton–Raphson algorithm), whereas the second (CG) is based on the Cole and Green algorithm which one is a "transformation" of the Fisher scoring algorithm.

2.2 Special cases and mean square error estimation

GAMLSS models using random effects, can be viewed as a generalization of a certain number of small area unit level models.[5] A linear mixed model is easily obtained form eq. 2 using a normal distribution with identity link function, identity matrix Z and with $g_2(\sigma) \sim 1$. If Z is a matrix of area level covariates this model is equal to the "two-level" small area model [10], par. 4.5.3, In the same way, choosing an appropriate distribution and a link function for $g_1(\mu)$ is possible to obtain a GLMM model. For example the model presented in [10], par. 4.6.1, can be rewrite in GAMLSS terms as a dichotomous dependent variable with a logit link function. With similar arguments is also possible to rewrite the models in [10], par. 4.5.4 and par. 4.6.4.

All those models need a specific formulation of Mean Square Error (MSE). Following [10] and [4] for the linear mixed models obtained through GAMLSS is possible to use an algorithm, based on a parametric bootstrap, that is almost equal to the ones used in [10] for the empirical best. The only difference is that for GAMLSS it's necessary to use a correction for the variance matrix of the error term $\hat{\Lambda}$ obtaining $\hat{\Lambda}_k^N$. This correction, proposed by Nelder in [9], is given due to the MAP estimation of random effects within GAMLSS. The other points of the algorithm are exactly the same with the models used in [10] replaced by the GAMLSS model described above. The estimated CV's obtained from those two models are at most equal.

3 Preliminary results and further developments

A preliminary result was obtained using the data set "Cornsoybean" [1]. To estimate area under corn and area under soybeans for each of $j = 12$ counties in north-central Iowa, using farm interview data in conjunction with LANDSAT satellite data were

[4] For an in-depth on the algorithms: Statiunopolus et *al.* [12].

[5] For all the model that we will present in this paragraph we will refer to Rao and Molina [10], in which is possible to find more accurate references to every single model.

used the EBLUP method and the GAMLSS model. Those last with a normal distribution with identity link function for μ and a log link function for σ. We use the same variables as [1] for μ but adding a random effect term in σ^2:

$$\mu = X_1\beta_1 + \gamma_{j1} \tag{4}$$
$$\log(\sigma) = \gamma_{j2} \tag{5}$$
$$\tag{6}$$

given eq. 4 the dependent variable has a normal distribution:

$$Y_{ij}|\gamma_{j1},\gamma_{j2} \sim \mathcal{N}(\mu_{ij},\sigma_{ij}) \tag{7}$$

Using random effects for σ^2 we could be able to capture the variation in terms of different dispersion of Y in the areas. Following [4] the MSE estimation for the GAMLSS models was obtained with a parametric bootstrap method for finite populations. As follows:

1. Fit the GAMLSS model eq. 4 to the sample data y_s, obtaining estimate for the model parameters;
2. For each $b = 1,\ldots,B$, with B large, generate are vectors $y_j^{(b)}$, $j = 1,\ldots,J$ form the normal population model Eq. 7 with the estimated parameters at point 1. Let $H_j^{(b)} = h(y_j^{(b)})$ be the true bootstrap parameter for area j;
3. Take the sample part of $y_j^{(b)}$, denoted $y_{js}^{(b)}$, $j = 1,\ldots,J$. With the whole sample data fit the GAMLSS model Eq. 4 and compute the small area estimation, $\hat{H}_j^{(GAb)}$ of $H_j^{(GAb)}$;
4. A bootstrap estimator of the MSE of $\hat{H}_j^{(GAb)}$ is obtained as

$$mse_B(\hat{H}_j^{(GAb)}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{H}_j^{(GAb)} - H_j^{(b)}\right)^2$$

Here, having to estimate the mean of Corn Hec. ($\hat{y_j}$) in 12 counties, following [10] we use

$$\hat{y}_j = \frac{1}{N_j}\left(\sum_{i\in s_j} y_{ij} + \sum_{i\in r_j}\tilde{y}_{ij}\right) \tag{8}$$

where $\tilde{y}_{ij}$ is is the predicted value for non-sample unit $i \in r_j$ given by replacing the mean area values of covariates into the μ equation of Eq. 4. Note that define also a specific function for σ in term of random effect only does not involve changes in the formula for the estimation of the mean. The mean values obtained from the two model are at most equal. The correlation between them is 0.98. The estimated CVs are less for the GAMLSS model with respect to EBLUP in 7 out of 12 counties, in 2 perform in the same way and in the remaining three it is the worst. From this first simple application we could conclude that defining a different σ_j^2 for each area

bring to an increase in precision of estimate. For sure more study are necessary to verify applicability and precision of those models. Authors are working on model and design simulations whit an higher number of both areas and sampled unit for each area.

What is presented is a first step to introduce the GAMLSS models into the SAE theory. Those models, releasing the assumption of the exponential family, could be very useful to implement functional-regression in a more accurate way than the ones that assume normality for the dependent variable. As shown, as special case, most common unit level small area models can be preform using GAMLSS models and these last seems to use very stable numerical-algorithms. Further directions of research involve expanding it, in particular for the MSE estimation, for those distributions whose supports are limited. Developing a simultaneous correlated system of estimation extending results in [8] and [7] to SAE.

References

1. Battese, G. E., Harter, R. M., and Fuller, W. A. An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 28-36. (1988)
2. Ferrante, M. R., and Pacei, S. Small domain estimation of business statistics by using multivariate skew normal models. Journal of the Royal Statistical Society: Series A (Statistics in Society), 1057-1088 (2017)
3. Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. Generalized linear models for small-area estimation. Journal of the American Statistical Association, 273–282 (1998)
4. Graf, M., Marín, J. M., and Molina, I. A generalized mixed model for skewed distributions applied to small area estimation. Test, 565-597 (2019)
5. Jiang, J., and Rao, J. S. Robust small area estimation: An overview. Annual Review of Statistics and Its Application, 337–360 (2020)
6. Khairil, A. N., Kurnia, A. and Kusman, S. Statistical models for small area estimation. The 3rd International Conference on Mathematics and Statistics (ICoMS-3) Institut Pertanian Bogor, Indonesia, (2008)
7. Mangiavacchi, L., Piccoli, L., and Rapallini, C. Personality traits and household consumption choices. The BE Journal of Economic Analysis & Policy, 433-468. (2021)
8. Marra, G., and Radice, R. Bivariate copula additive models for location, scale and shape. Computational Statistics & Data Analysis, 99-113, (2017)
9. Nelder, J. A., Contribution to the discussion of Rigby and Stasinopoulos, Generalized additive models for location, scale and shape. Applied Statistics (2006)
10. Rao, J. N. K., and Molina, I. Small area estimation. Wiley series in survey methodology. John Wiley & Sons, Inc (2015)
11. Rigby, R. A., and Stasinopoulos, D. M. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 507–554 (2005)
12. Stasinopoulos, D. M., Rigby, R., Heller, G., Voudouris, V., and De Bastiani, F. Flexible regression and smoothing: using GAMLSS in R. Chapman and Hall CRC the R Series. Chapman & Hall CRC (2017)
13. Souza, D. F., and Moura, F. A. Multivariate beta regression with application in small area estimation. Journal of Official Statistics, 747-768. (2016)

The use of remotely sensed data in sampling designs for forest monitoring

L'uso dei dati telerilevati nei disegni campionari per il monitoraggio forestale

C. Bocci, G. Chirici, G. D'Amico, S. Francini and E. Rocco

Abstract Technology development has led to a large availability of increasingly precise remotely sensed data ready-to-use, but several countries' forest monitoring programs are still based on the traditional systematic sampling design of National Forest Inventories (NFIs). It is well known that, in order to improve surveys estimates, auxiliary data can be used both in the design phase and in the estimation phase. Recent literature has presented some proposals of using remote sensing (RS) data to improve NFIs but all are limited to specific countries or areas. Our aim is to investigate how RS data can be exploited to produce global forest estimates in a more cost-efficiently way. We assess the use of a global Landsat-based cloud/noise free Best Available Pixel (BAP) composite image in the design phase in order to produce reliable estimates of the biomass and soil carbon density.

Abstract *Nonostante lo sviluppo tecnologico abbia portato a un'ampia disponibilità di dati telerilevati sempre più precisi e pronti per l'uso, i programmi di monitoraggio forestale di diversi paesi si basano ancora sul tradizionale disegno di campionamento sistematico degli inventari forestali nazionali. E' noto che l'utilizzo di informazioni ausiliarie, sia in fase di disegno che in fase di stima, produce stime campionarie più efficienti. Recenti proposte in letteratura hanno introdotto l'uso di variabili ausiliare telerivevate negli inventari forestali nazionali, ma tutte si concentrano su singole aree o paesi. Il nostro obiettivo è di investigare come i dati telerilevati possano essere sfruttati per ottenere stime di variabili forestali a livello globale in modo efficiente in termini di costi. In questo studio valutiamo l'uso in fase di disegno di immagini satellitari globali composite di fonte Landsat per produrre stime attendibili della densità di Carbonio da biomassa e suolo.*

Key words: National Forest Inventory, Spatial estimation, Auxiliary data

Chiara Bocci, Emilia Rocco
Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence,
e-mail: chiara.bocci@unifi.it, emilia.rocco@unifi.it

Gherardo Chirici, Giovanni D'Amico, Saverio Francini
Department of Agriculture, Food, Environment and Forestry, University of Florence,
e-mail: gherardo.chirici@unifi.it, giovanni.damico@unifi.it, saverio.francini@unifi.it

1 Introduction

Forest ecosystems are a critically important component of the global terrestrial carbon cycle absorbing almost 3 billion tons of anthropogenic carbon annually, or 30% of the total emissions associated with fossil fuel burning and net deforestation [1]. Unfortunately, the vast array of goods and services provided by forests are threatened by climate change, and monitoring forests and supporting decision-makers with reliable data is more important than ever. For this reason, monitoring the biomass and biomass variations due to human activities or climate changes is more essential than ever. On the other hand, while information related to tree structure is of great interest, such information is often not available wall-to-wall and over large areas due to large survey and processing costs.

Usually, forest variables estimates are provided by a design-based approach with data collected in the field in the frameworks of traditional National Forest Inventories (NFIs) [2, 5]. Estimates aggregated for large geographic areas are requested in the context of national and international reporting. However, NFI data are expensive, since they require long and costly field campaigns, so a major scientific challenge is to develop new methods to derive useful forest information in a more cost-efficiently way [8]. In the last decades, recent advancements of earth observation technologies opened the possibilities to use remote sensing (RS) data to support forest inventories, so that the strategies to collect information and derived estimates in the context of forest inventories changed consequentially ([7], [2], [4] among others). However, at best of our knowledge, most of these contributions limit their applications to specific countries or areas.

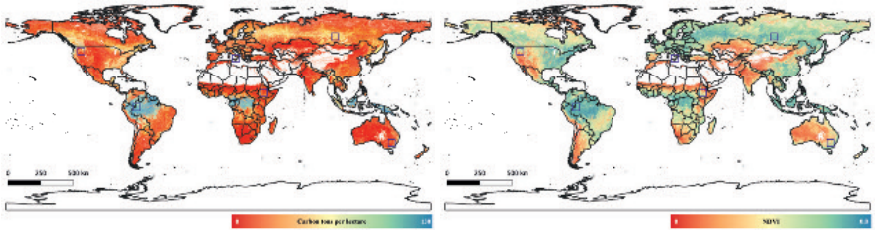
Our aim is to investigate how RS data can be exploited to produce global forest estimates in a more cost-efficiently way. RS data in combination with field data can be used to enhance the precision of large-scale forest inventories both in the design phase and in the estimation phase, but in this first contribution we limit our analysis to the design phase.

2 Data

The target population consists of all the pixels of a harmonized global map (with a resolution of 300 meters) of above-and-below-ground terrestrial carbon storage (tonnes (t) of Carbon per hectare (ha)) in biomass and soil for the reference year 2010 [6]. This map depicts the 'total' carbon density (CD) at global scale and is obtained by combining data from multiple publicly available datasets. A detailed description of the dataset can be found at the source website[1].

To produce estimates at global scale we require auxiliary variables with the same coverage, which can be obtained by composition. The Best Available Pixel (BAP) image compositing application in Google Earth Engine GEE-BAP enables the gen-

[1] http://developers.google.com/earth-engine/datasets/catalog/WCMC_biomass_carbon_density_v1_0



(a) (b)

Fig. 1 Above and Below Ground Biomass Carbon Density (a) and Normalized Difference Vegetation Index (b).

eration of annual global Landsat-based cloud/noise free BAP composite images. This kind of image represent a new paradigm in remote sensing that is no longer reliant on scene-based analysis and the Normalized Difference Vegetation Index (NDVI) calculable from it is used as auxiliary variable in our analysis. Landsat, and so the composite map, has a resolution of 30 meters; to be consistent with the CD we aggregate (using average) the BAP composite map at 300 meters resolution. The two maps of the target and auxiliary variables are showed in Figures 1(a) and 1(b).

3 Sampling strategy

The global study area is tessellated into squared cells, with a size of 6 degrees of latitude $\times$ 6 degrees of longitude (which roughly corresponds to the whole Italian surface). Each of these cells is a domain of study, that is we are interested in providing estimates globally as well as for each cell of the grid. For this, we can assume that each cell is a sort of stratum and from each of them we select a sample independently from the others. All the cells have the same size and include the same number of pixels L, but the actual size of the eligible population N_h varies between cells since there could be areas in which the target variable is not defined (over the oceans, lakes, deserts, etc...).

We assume that the sample size is constant for all the cells with the same eligible population. This value is then reduced proportionally for those cells with a lower number of eligible pixels. The sample size should be chosen considering the trade-off between the need to produce precise estimates and the organizational and cost constraints of the survey. To evaluate the influence of this choice, we consider three different settings: 10000, 5000 and 2000 sampled units.

Performance of the following different sampling designs is assessed through a simulation study. All designs exploit the geographical location of the study units and/or the values of the auxiliary variable NDVI.

1. **SPAT_LPM**: Spatially balanced sampling through Local Pivotal Method (LPM), suggested by [3], with equal inclusion probabilities $\pi_i s$. The LPM in its original

formulation produces samples that are well spread in the geographic space. It can be considered as a proxy of the systematic sampling commonly used for NFIs.
2. **AUX_LPM**: Sampling balanced through LMP in the space spanned by the auxiliary variable, with equal $\pi_i s$.
3. **BIV_LPM**: Sampling balanced through LMP in the space spanned by both the geographical coordinates and the auxiliary variable, with equal $\pi_i s$.
4. **UNEQ_LPM**: Spatially balanced sampling through LMP with unequal inclusion probabilities $\pi_i s$ proportional to the auxiliary variable.
5. **STRp_AUX_LPM** and **STRn_AUX_LPM**: Stratified sampling with AUX_LPM design in each stratum. The cell is partitioned in 100 sub-areas (strata) and then the AUX_LPM is applied in each stratum. Two allocation rules are considered: Proportional and Neyman's with respect to the variance of the auxiliary variable.
6. **STRp_BIV_LPM** and **STRn_BIV_LPM**: Stratified sampling with BIV_LPM design in each stratum. The same stratification designs described in the previous point, but with BIV_LPM applied in each stratum.

In addition, as a comparison, simple random sampling (SRS) is considered both over the entire area and within strata.

The basic idea of LPM introduced by [3] is to avoid that units close in distance appear together in the sample. First an inclusion probability $0 < \pi_i \leq 1$ is assigned to each unit so that their sum over the population is equal to the fixed sample size. The $\pi_i s$ can be either equal for all the units in the population or unequal (we considered unequal $\pi_i s$ in design 4). The sample is then obtained in at most N steps, where N is the population size. At each step one unit i is selected randomly from the available population and another unit j is chosen among the remaining units in the population by minimizing a distance function among them. This can be a univariate or a multivariate function that measures the distance with respect to one or more auxiliary variables, among which we can include the spatial coordinates. When all the variables are continuous the Euclidean distance is commonly used. Moreover, when multiple auxiliary variables are used, they are usually standardized in order to balance the contribution of each variable. After the selection of the unit i and j their inclusion probabilities are updated by using the following rule:

$$\begin{aligned}
&\text{if } \pi_i + \pi_j < 1 \text{ then } (\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \end{cases} \\
&\text{if } \pi_i + \pi_j \geq 1 \text{ then } (\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j} \end{cases}
\end{aligned} \tag{1}$$

As a result, in each step at least one unit is definitively excluded from the population, either because its probability becomes zero, and consequently it is definitely excluded from the sample, or because its probability becomes one and therefore is included in the sample. The procedure continues, updating at each step the probabilities of inclusion obtained in the previous step, until all units in the population

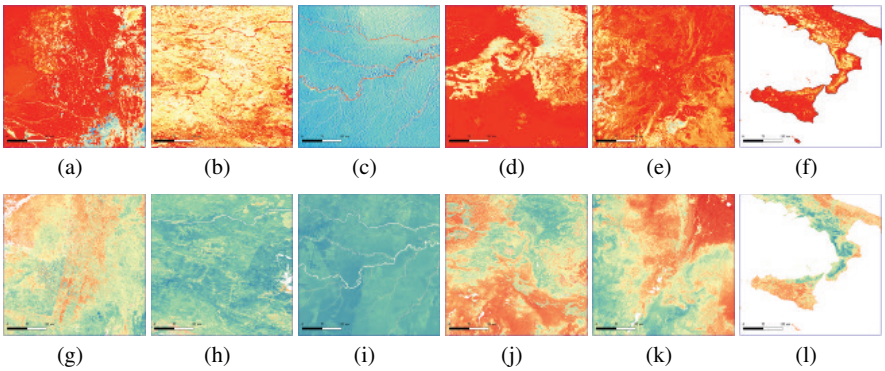


Fig. 2 Above and Below Ground Biomass Carbon Density: (a)-(f) Cells 1-6; Normalized Difference Vegetation Index: (g)-(l) Cells 1-6.

Table 1 Descriptive statistics of the study and the auxiliary variables.

Cell	μ_{CD}	μ_{NDVI}	CV_{CD}	CV_{NDVI}	$\rho_{CD,NDVI}$
1	28.49	0.424	0.856	0.516	0.501
2	63.14	0.735	0.468	0.145	0.317
3	27.69	0.487	1.643	0.342	0.390
4	32.21	0.414	1.141	0.514	0.615
5	177.42	0.798	0.177	0.108	0.348
6	19.86	0.469	1.307	0.445	0.739

are processed. LPM selects the units with the same probability $\pi_i s$ initially assigned to them, therefore the population mean can be estimated with the usual Horvitz-Thompson estimator.

4 Simulation results and discussion

We briefly present some first results of our simulation study. The performance of the designs described in the previous section is evaluated through Monte Carlo experiments in 6 cells, selected across the globe, which differ in the distributions of the auxiliary and study variables and in their relation (see Figures 2(a)-2(f) and 2(g)-2(l), and the descriptive statistics in Table 1).

Table 2 presents the root mean squared errors (RMSEs) for each sampling design with a sample size of 5000 units[2], obtained with 500 Monte Carlo replications. Results show that the use of any auxiliary variables (geographical location or NDVI) is always preferable to the basic SRS. Moreover, they confirm the usefulness of including the geographical information in the sampling design. About the use of

[2] Tables for the other sample sizes are not included for lack of space

the additional auxiliary variable, our preliminary results are not straightforward. It seems that its contribution to improving the estimates' efficiency depends on many factors, particularly on the correlation with the study variable and on the variability of both study and auxiliary variables. Moreover, we found that the usefulness of the additional auxiliary variable increases as the sample size decreases.

Table 2 Root mean square error, with 500 replications and sample size = 5000

						STRp			STRn		
Cell	SRS	SPAT LPM	AUX LPM	BIV LPM	UNEQ LPM	SRS	AUX LPM	BIV LPM	SRS	AUX LPM	BIV LPM
1	0.353	0.210	0.281	0.265	0.240	0.274	0.257	0.234	0.295	0.258	0.241
2	0.422	0.326	0.373	0.325	0.316	0.376	0.317	0.321	0.380	0.349	0.355
3	0.635	0.417	0.530	0.496	0.359	0.541	0.487	0.453	0.570	0.459	0.468
4	0.510	0.302	0.382	0.336	0.235	0.350	0.282	0.312	0.369	0.230	0.317
5	0.443	0.390	0.428	0.405	0.507	0.420	0.397	0.389	0.412	0.402	0.418
6	0.770	0.533	0.424	0.404	0.397	0.698	0.471	0.497	0.685	0.465	0.463

References

1. Canadell, J.G., Raupach, M.R.: Managing Forests for Climate Change Mitigation. Science **320**, 1456–1457 (2008) doi: 10.1126/science.1155458
2. Chirici, G., Giannetti, F., McRoberts, R.E., Travaglini, D., Pecchi, M., Maselli, F., Chiesi, M., Corona, P.: Wall-to-wall spatial prediction of growing stock volume based on Italian National Forest Inventory plots and remotely sensed data. Int. J. Appl. Earth Obs. Geoinf. **84**, 101959 (2020) doi: 10.1016/j.jag.2019.101959
3. Grafström, A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (2012) doi: 10.1111/j.1541-0420.2011.01699.x
4. Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K.T., Maltamo, M., Nilsson, M., Nord-Larsen, T., Næsset, E., Olsson, H.: Remote sensing and forest inventories in Nordic countries – roadmap for the future. Scand. J. For. Res. **33**, 397–412 (2018) doi: 10.1080/02827581.2017.1416666
5. McRoberts, R.E., Liknes, G.C., Domke, G.M.: Using a remote sensing-based, percent tree cover map to enhance forest inventory estimation. For. Ecol. Manage. **331**, 12–18 (2014) doi: 10.1016/j.foreco.2014.07.025
6. Soto-Navarro, C., Ravilious, C., Arnell ,A., de Lamo, X., Harfoot, M., Hill, S.L.L., Wearn, O.R., Santoro, M., Bouvet, A., Mermoz, S., Le Toan, T., Xia, J., Liu, S., Yuan, W., Spawn, S.A., Gibbs, H.K., Ferrier, S., Harwood, T., Alkemade, R., Schipper, A.M., Schmidt-Traub, G., Strassburg, B., Miles, L., Burgess, N.D., Kapos, V.: Mapping co-benefits for carbon storage and biodiversity to inform conservation policy and action. Philos. Trans. R. Soc. Lond. B **375**, 20190128 (2020) doi: 10.1098/rstb.2019.0128
7. Waser, L.T., Ginzler, C., Rehush, N.: Wall-to-Wall Tree Type Mapping from Countrywide Airborne Remote Sensing Surveys. Remote Sens. **9**, 766 (2017) doi: 10.3390/rs9080766
8. White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., Tompalski, P.: Remote Sensing Technologies for Enhancing Forest Inventories: A Review. Philos. Trans. R. Soc. Lond. B **42**, 619–641 (2016) doi: 10.1080/07038992.2016.1207484

Analyzing different causes of one–inflation in capture recapture models for criminal populations.

Modelli cattura–ricattura per la stima di popolazioni criminali: analisi delle diverse cause di one–inflation.

Davide Di Cecco, Andrea Tancredi and Tiziana Tuoto

Abstract Our goal is the estimation of the size of various criminal populations on the basis of administrative data on criminal proceedings by exploiting capture–recapture models for repeated count data. The data at our disposal exhibit an abundance of units that are captured exactly once, which suggests the necessity of explicitly modeling this deviation. We distinguish three possible causes for this phenomenon, namely, the erroneous inclusion of out–of–scope units, a particular behavioural effect preventing subsequent captures after the first one, and missed links in a Record Linkage step. We then propose three families of "one–inflated" models to estimate the number of uncaptured units which we compare on datasets of Italian criminal proceedings.

Abstract *L'utilizzo di modelli cattura ricattura per stimare il numero di criminali non è nuovo in letteratura. In questo lavoro intendiamo stimare il numero di individui connessi a specifiche attività criminali in Italia sulla base dei dati sul numero di provvedimenti giudiziari individuali. I dati mostrano chiaramente la necessità di modellizzare l'eccesso di individui catturati solo una volta. Individuiamo 3 cause possibili per questo fenomeno noto come "one–inflation": la presenza di unità che andrebbero escluse dalla popolazione; un behavioural effect già noto nela stima delle popolazioni animali; e la presenza di mancati abbinamenti in una fase preliminare di Record Linkage. Proponiamo di conseguenza 3 strategie per le 3 possibili cause.*

Key words: Criminal populations. Capture-recapture. Official statistics.

Davide Di Cecco
Università di Roma La Sapienza e-mail: davide.dicecco@uniroma1.it

Andrea Tancredi
Università di Roma La Sapienza e-mail: andrea.tancredi@uniroma1.it

Tiziana Tuoto
ISTAT, e-mail: tuoto@istat.it

1 Introduction

Estimating the number of people involved in certain illegal activities helps us in allocating the police forces to counter it, in evaluating the effectiveness of prevention and counteraction policies, and in evaluating the economic value associated to those activities. The European Parliament and Council identified the smuggling of goods, prostitution exploitation and drug trafficking as the main sources of illegal economic transactions to report in the national accounts aggregates. In this work we aim at estimating the number of people involved in these three illegal activities in Italy during 2014 by exploiting administrative data on criminal proceedings in a capture–recapture model.

Data from the Ministry of Justice report alleged crimes for which the judicial authority started a criminal proceeding. We count the number of times each individual appears in the Prosecutor's offices registers. Under a (reasonable) assumption of time–homogeneous capture probabilities, the data can then be summarized as counts of units captured j times, $\{n_j\}_{j=1,2,\ldots}$.

Figure 1 depicts our data. We observed a total of 3349 smugglers (top figure), 2740 prostitution exploiters (mid), and 34964 drugs traffickers (bottom).

The common parametric approach to estimate the number n_0 of individuals unreported to the justice system, is to define a counting distribution for the observed (truncated in zero) number of captures in the population. We considered various distributional alternatives: Poisson, Geometric, Negative Binomial, finite mixtures of Poisson and of Geometric. Even considering different parametric options, in all three observed distributions the number of units captured exactly once exceeds the expectation under the model hypothesis. This phenomenon is known in literature as "one–inflation" and is of particular importance, as it typically implies a substantial difference in the estimates of n_0.

2 Causes of one–inflation

One–inflation might be generated by different factors, and we have identified the following three possible mechanisms:

1. a specific behaviour of the units, which learn how to avoid subsequent captures after the first one;
2. the presence of spurious units which do not belong to the reference population;
3. errors in re-identifying the units (linkage errors), due to the lack of unique identifiers and error-prone soft identifiers.

In our data we cannot exclude any of the previous factors having an impact on the observed excess of "ones".

The first mechanism generating one–inflation can be viewed as an extreme form of what in animal abundance analysis is called "trap shy" behaviour. In our case, we suppose that, after the first capture, a portion of persons involved in illegal activities

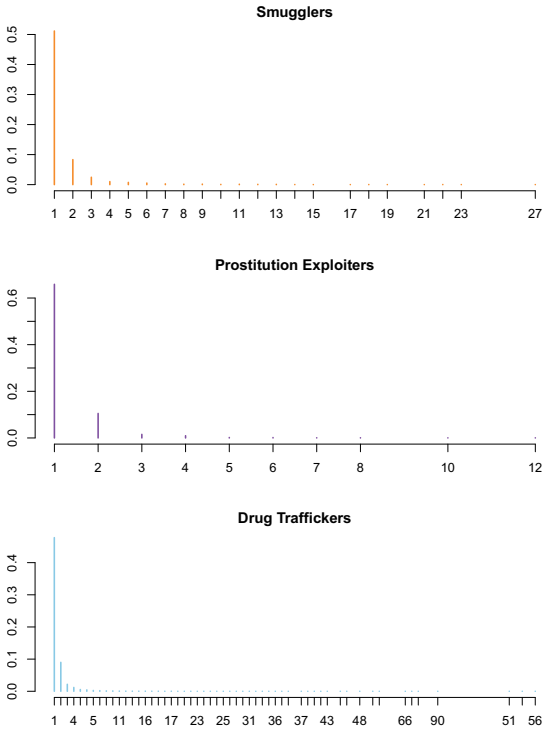


Fig. 1 Relative frequencies of observed number of captures for smugglers (top), prostitution exploiters (mid), and drug traffickers (bottom) in Italy in 2014

acquires the necessary knowledge/ability to avoid any subsequent capture. This can be formalized as a latent subpopulation presenting right–censored data, as we only know that their potential number of captures is larger than one. This hypothesis has been explored in depths in several recent papers (see, e.g., [6], [7], [8], [1]).

We cannot exclude the presence of spurious cases as source of one–inflation in our data. Indeed, we analyse records from the reported offences for which the judicial authority started a proceeding. However, we actually do not have information on whether the person reported was sentenced in court or acquitted. Then, a portion of the units captured once may be out–of–scope, and should be discarded from the analysis. The presence of out–of–scope or spurious captures has been considered, for example, in [2], [5]. Under this hypothesis, the observed value n_1 represents an upper bound for the real number of criminals captured once.

Finally, due to privacy issue, our data are not provided with a unique identifier, hence, we cannot exclude the possibility of one–inflation deriving from linkage errors. In fact, when matching information does not suffice to recognise multiple captures of the same individual, the resulting missing links erroneously increase the number of singletons. The effect of linkage errors on population size estimates in the

multiple system estimation framework has been investigated both with a Bayesian perspective, see [9] and with a frequentist approach [3], and [4].

To the best of our knowledge, the explicit modeling of linkage errors in repeated count data for population size estimation is a topic not previously considered. In this contribution, we will show some possible linkage errors modeling in this setting, to derive population size estimates that will take into account the linkage uncertainty. In addition, we compare and discuss the estimates produced by the different sources of one–inflation.

3 Linkage errors in repeated count data

The lack of unique identifiers is a common situation both in social science and in natural science applications. Linkage procedures allow recognising the same unit even in the absence of a unique identifier, hence they are a prerequisite for population size estimates. Usually, the linkage is far from being a simple operation and it might introduce some errors, namely false and missed links. The linkage procedure finds a trade–off between the two kinds of linkage errors. In addition, the two errors have opposite effects on repeated count data, and consequently on population size estimates. The false links erroneously increase counts greater than one, since different units are erroneously recognised as the same one. On the other side, missed links, i.e. failure in recognising the same unit, increase the number of singletons.

Since the linkage procedure is actually a statistical activity, the one-inflation generated by linkage errors can be considered from both a primary and a secondary analysis perspective. In the former, the linkage and the population size estimate are viewed as components of a unique process and, via a hierarchical structure, as in [9] and [10], we are able to propagate the uncertainty between the parameter estimation step and the matching procedure. A clear advantage of this approach is that we are able to take into account at the same time both the two types of linkage errors.

Due to its relevance in real life application, we also consider the secondary analysis perspective, that is we assume that the linkage has been performed by someone else, and only quality indicators like linkage errors are available to be included in the subsequent analysis. In this perspective we explicitly model the one–inflation caused by missed links, since usually the linker tries to minimise the false matches, at the price of accepting some highest level of missed link errors. A specific characteristic of this model is the additional uncertainty affecting the total number of distinct units, as a consequence of the linkage error itself. Actually, assuming that our data is affected by missed links reduces the sample size in terms of distinct units and the observed one can be considered an upper–bound of the true one. On the contrary, when modeling the one–inflation due to the behavioural effect and spurious cases, we can consider as fixed the observed number of distinct units.

4 Results

For the sake of brevity, we only show in Table 4 some results on prostitution exploiters data, considering both non inflated and one–inflated models and other standard estimates for the proposed setting. For the one-inflation model generated by linkage errors we assume a secondary analysis framework. In particular we show the results obtained by taking a Poisson and a Geometric distribution for the individual captures y^* and assuming that the observed captures y are given by a Binomial variate with size $y^* - 1$ and that $y^* - y$ extra singletons are added to the data set. The idea underlying the model is that when the individual records subsequent to the first capture are compared with the initial record we may fail to link the records reducing the number of captures for that individual from y^* to $y^* - m$ where m is the number of missing links. Moreover the missing links cannot be matched to any other records. In fact, we exclude both the possibility to have false matches and the chance to reconstruct, by the missing links, a sub-cluster of records referring to a given individual. This way the missing links may provide only single captures. Note also that while for the one-inflation caused by behavioural effects and spurious cases, Bayesian inference can be obtained via standard Markov Chain Monte Carlo simulation algorithms, for the one-inflation model caused by linkage error, due to the likelihood intractability, we opted for an approximate Bayesian computation (ABC) approach.

Estimator/Model	$\hat{N}$	95%CI.$\hat{N}$	
Ignoring one–inflation			
Poisson	7210	6780 - 7689	
Geometric	13332	12415 - 14394	
Chao	9851	8961 - 10868	
Zelterman	10030	9033 - 11027	
Poisson Mixture model	10073	9049 - 11110	
Modeling one–inflation		behavioural effect	$\hat{\omega}$
One-inflated Poisson	3895	3656 - 4156	0.645
One-inflated Geometric	8182	7406 - 9233	0.478
One-inflated Poisson mixture model	6613	5284 - 9870	0.261
Modeling one–inflation		spurious cases	$\hat{\psi}$
One-inflated Poisson	3154	3035-3304	0.560
One-inflated Poisson mixture model	4711	3744 - 8799	0.163
Modeling one–inflation		linkage error	$\hat{\mu}$
One-inflated Poisson	2362	2090 - 2681	0.645
One-inflated Geometric	9384	7896 - 10720	0.323

Table 1 Prostitution exploitation data: posterior mean and credible intervals for the population size N, posterior mean for the one-inflation parameters ω,ψ and μ accordingly to the source of one-inflation.

Specification on the different models and additional results will be presented during the conference and in an extended version of this contribution.

References

1. Böhning, D., Friedl, H. (2021). Population size estimation based upon zero–truncated, one–inflated and sparse count data. Statistical Methods & Applications, 1–21.
2. Böhning, D., van der Heijden, P. G. (2019). The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. The Annals of Applied Statistics,13(2), 1198–1211.
3. Di Consiglio, L., Tuoto, T. (2018). Population size estimation and linkage errors: the multiple lists case. Journal of official statistics,34(4), 889–908.
4. Di Consiglio, L., Tuoto, T., Zhang, L.C. (2019). Capture-recapture methods in the presence of linkage errors. In Analysis of integrated data(pp. 39–71). Chapman and Hall/CRC
5. Fegatelli, D., Farcomeni, A., Tardella, L. (2017). Bayesian population size estimation with censored counts. In *Capture-recapture methods for the social and medical sciences* (pp. 371–385). Chapman and Hall/CRC.
6. Godwin, R., Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. Journal of the Royal Statistical Society: Series C (Applied Statistics),66(2), 425–448.
7. Godwin, R. (2017). One-inflation and unobserved heterogeneity in population size estimation. Biometrical Journal, 59(1), 79–93.
8. Godwin, R. (2019). The one–inflated positive Poisson mixture model for use in population size estimation. Biometrical Journal, 61(6), 1541–1556.
9. Tancredi, A., Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. Annals of Applied Statistics,5(2B), 1553–1585
10. Tancredi, A., Steorts, R., Liseo, B. (2020). A unified framework for de-duplication and population size estimation (with discussion). Bayesian Analysis, 15(2), 633-682.

Administrative database and official statistics: an IT and statistical procedure

Database amministrativi e statistiche ufficiale: una procedura statistico-informatica

Caterina Marini, Vittorio Nicolardi

Abstract The availability of a complete harmonized database that includes all information, both administrative and statistical, on the phenomenon under investigation is a precious analytical instrument very often unavailable because of issues related to important obstacles that involve both statistical and computer problems, and administrative and legal impediments. In this work, we depict the IT and Statistical procedure used to provide the first Full Information Harmonised Real Estate Database in the analysis of the real estate economy.

Abstract *La disponibilità di un database completo e armonizzato che includa tutte le informazioni, sia amministrative che statistiche, riguardanti il fenomeno in analisi è uno strumento analitico prezioso molto spesso non disponibile a causa di problemi legati a importanti ostacoli che coinvolgono sia aspetti statistici e informatici, sia impedimenti amministrativi e legali. In questo lavoro, noi descriviamo la procedura informatica e statistica usata per produrre il primo Database Immobiliare Armonizzato a Informazione Completa da usare nell'analisi del fenomeno economico del settore immobiliare.*

Key words: Administrative Data, Big Data, Real Estate Economy, Harmonised Database.

1 Introduction

[1] Caterina Marini, Department of Economics and Finance University of Bari Aldo Moro, email: caterina.marini@uniba.it

Vittorio Nicolardi, Department of Economics and Finance University of Bari Aldo Moro, email: vittorio.nicolardi@uniba.it

The debate about the administrative databases and their use to provide a statistical information that is not differently affordable is internationally considered significant and relevant to face the nowadays scientific challenge. Many issues are involved, and all are basically fundamental to provide a complete depiction of the socio-economic scenario in a world that is becoming more complex to be analysed. It is believed that the administrative data are largely available and readily usable, and this is one the most common errors that is unconsciousness committed. When working with administrative data, the main problem is the integrated use of data that are yielded by independent administrative offices, sometimes belonging to the same organisation, public authority or company, and differently collected for specific administrative purpose. Furthermore, in most cases databases are structurally diverse because of well-established IT and administrative procedures that are not updated to align all information for users. As a consequence of their nature, administrative data can be affected by material and human errors, such as duplication, missing values and erroneous information, and therefore they are not controlled and statistically correct. Another aspect that is important to consider is that administrative data are a little part of the more complex phenomenon known as Big Data. In literature, the Big Data aspect of the problem is widely discussed (Harford, (2014); Kitchin, (2014 a, b); Marini and Nicolardi, (2020) (2021)). Finally, it is crucial the harmonisation of the administrative data, which are raw information, with the official statistical data as yielded by the National Statistical Institutes (NSIs, hereafter) and, therefore, consolidated statistically. In some of our recent works, we provided an original fully comprehensive statistical tool to describe the phenomenon of the real estate economy: the full information harmonised database is referred to the real estate phenomenon where administrative data and Italian NSI's data are perfectly aligned and the Big Data crucial aspects solved. In view of the current Italian government intention to reform the Italian National Registry of private and public properties (known as "catasto") as stated in the National Recovery and Resilience Plan document as part of the Next Generation EU programme, we consider our Full Information Harmonised Real Estate Database (FIHRE-DB, hereafter) and the economic experiment we run on a sole city to be an important starting point towards a fair reform. In this work, we describe the IT and statistical procedure we defined to create the FIHRE-DB. The full procedure needed SAS software and QGIS software to yield geo-maps and overlapping steps.

2 The Data

In our work, the experiment is restricted to the territory of the city of Bari, in southern Italy, because that was part of a national research project. The databases used to create the FIHRE-DB are essentially 5 key databases (DBs, hereafter) and 3 instrumental DBs, the latter useful for harmonising all data. The 5 key DBs are independent administrative DBs, generally managed by independent Public Administration (PA, hereafter) offices of the city of Bari, and provide autonomous real estate information. Four of 5 key DBs belong to the Real Estate Registry (RER,

hereafter) and contain all information related to the real estates. The last key DB belongs to the Real Estate Italian Observatory (REIO, hereafter) of the Italian Revenue Agency. The 3 instrumental DBs are mutually fully independent because they belong to independent PA: the first belongs to RER and allows the final connection between the administrative data and the official statistical data; the second belongs to REIO and the last belongs to the Italian NSI, and both are the geo-localisation datasets used for the final overlapping action.

3 The Method

The method we proposed is a *step by step* procedure, depicted in the subsections that follow, to obtain the final outcome, that is the FIHRE-DB. A first study of all databases highlights that the 5 RER DBs needed an inevitable prior action of cleaning because of the many duplicated, missing and erroneous data. Afterwards, that was possible to proceed with the harmonisation of all data based on key fields.

3.1 RER Databases

The administrative RER DBs are fundamental to create the final harmonised database and, at the same time, those that required more effort in terms of cleaning. Furthermore, the analysis of data highlighted also that a recoding action was necessary to correct some information that, in the past, was erroneously recorded because of some administrative territorial process that involved the inclusion of some little municipalities in the city. Therefore, the cleaning and recoding actions have been important to homogenise information provided by RER and align the 5 RER DBs. The first RER DB includes a list of records that contains all technical and economic information of each unit. Therefore, without affecting the statistical significance of the analysis we deleted the incomplete and duplicated records in this first RER DB. That is a double step procedure. The first step allowed to sort data by means of two key fields: *Id_Property*, which identifies each real estate in RER, and *Ordinal_Number*, which identifies the mutations corresponding to the units:

```
proc sort data=name_Database1;
by Id_Property DESCENDING Ordinal_Number;                    [1]
```

The second step of the procedure permitted to clean the DB and utilises as key field *Id_Property* to delete the oldest data related to each real estate mutation:

```
proc sort data= name_Database1 out= name_Database2 nodupkey ;
by Id_Property;                                              [2]
```

Furthermore, we proceeded to also delete all real estate units without income:

```
data mainDatabase1;
set name_Database2;                                          [3]
where (Income not in (.));
```

The second RER DB comprises other real estate information, mainly the Urban

Section and the Registry Sheet, Subordinate and Parcel. This second RER DB includes a much higher number of records than that of the first RER DB because of both duplications, caused by some administrative change, and the presence of some real estate unit whose record has not been deleted though the building was demolished. Therefore, we deleted the incomplete and duplicated records in this second RER DB and proceeded to recoding some field. That is a multiple step procedure. The first step allowed to sort data by means of two key fields: i.e. *Id_Property* and *Urban Section*:

```
proc sort data= name_Database3;
by Id_Property DESCENDING Urban Section;                    [4]
```

The second step of the procedure permitted to clean the DB and utilises as key field *Id_Property* to delete duplicated data for each real estate:

```
proc sort data= name_Database3 out= name_Database4  nodupkey ;
by Id_Property;                                             [5]
```

Furthermore, we proceeded to also delete all records where the real estate units have an erroneous code for the field *Registry Sheet*.

```
data name_Database5;
set name_Database4;                                         [6]
where (Registry Sheet in (report the right "Registry Sheet
numbers/code" set));
```

The third step of the procedure allowed to recode the sole field *Urban Section* in this final DB:

```
data name_Database6;
set name_Database5;                                         [7]
if (Urban Section in (report the old "Section code" set)) then
Section =  report the new "Section code";
…;
```

Finally, to homogenise information between the two achieved DBs the last step deleted all the records that were in the second DB and not in the first DB. That is a *right join* between the *name_Database6* and *mainDatabase1*. This final step yielded the *mainDatabase2*. The third RER DB comprises the toponyms of each real estate unit that are important to identify the exact localisation of each unit on the urban territory. This third RER DB shows many duplications because of the modifications of some toponym and/or building number. Therefore, a cleaning action is necessary and a double step procedure was defined. The first step allowed to sort data by means of two key fields: *Id_Property*, which identifies each real estate in RER, and *Ordinal_Number*, which identifies all the mutations in toponyms and/or building number corresponding to the real estate:

```
proc sort data= name_Database7;
by Id_Property DESCENDING Ordinal_Number;                   [8]
```

The second step of the procedure permitted to clean the database and utilises as key field *Id_Property* to delete the oldest data related to each real estate mutation:

```
proc sort data= name_Database7 out= name_Database8 nodupkey ;
by Id_Property;                                             [9]
```

Finally, to homogenise information between the two previous achieved databases a final step of the procedure deleted all the real estate records that are in this third database and not in the previous two. That is a *right join* between the

name_Database8 and *mainDatabase1*. This final step yielded the *mainDatabase3*.

The final step of the procedure that is referred to the first 3 key RER DBs is the creation of a sole omni comprehensive DB that aligned data by means of the key field *Id_Property*:

```
data mainDatabaseRER;
merge mainDatabase3 mainDatabase2 mainDatabase1;
by Id_Property;
```

The last RER DB contains all data referred to the Real Estate Income and Value, both administrative data, and the corresponding council tax rate. The cleaning and recoding problems we found in this last RER DB are the same we encountered in the second RER DB previously described. In this last RER DB the codification of the real estate is based on the taxpayer in a framework of Urban Section, Registry Sheet, Subordinate and Parcel. Duplication rises because more than one taxpayer can pay the council tax for the same building in a co-ownership status. Therefore, we followed exactly the same steps of the procedure as in [5] - [7] and worked out the needed databases: first the *name_DatabaseT1* starting from the raw database; second the *name_DatabaseT2* as the outcome of the procedure [6] and [7], and finally the *mainDatabaseCT* as the result of the procedure [5].

3.2 REIO Database

The REIO DB is the last key DB we use in our work and one of the most reliable sources of data to analyse the real estate monetary value dynamics in the Italian real estate market. The REIO real estate value data are calculated based on the trade price per square meter of the properties. They are open data on a biannual basis referred to the minimum and maximum price for all the different types of real estates at the level of the council territory. To use a univocal REIO value in the analysis, the midrange value for each record was calculated. In the REIO dataset, the council territory is split into homogeneous areas that experience the same economic and socioenvironmental characteristics, i.e. the REIO zones. The REIO DB comprises statistical valued data. Therefore, no cleaning and recoding actions were necessary.

3.3 Full Information Harmonised Real Estate Database

Once the RER DBs have been achieved in a homogenous format, the last section of the work is the creation of the FIHRE-DB. The most crucial step in this last part of the procedure is to find out a method to connect and merge all the RER data with the REIO data. In fact, both *mainDatabaseRER* and *mainDatabaseCT* cannot be connected with REIO DB by any field. Therefore, the instrumental databases as describe previously play an important role to obtain the FIHRE-DB. They are all geo-referred and, therefore, a geo-localised process is suitable to solve the problem. In fact, we defined a multiple procedure of merging all the key databases through

the geo-localised information. First, the creation of a "linking table" named BRIDGEDB allowed to connect the REIO zones with the Italian NSI census sections. In other words, the procedure allows to link the Italian NSI census sections, and indirectly all real estate units, with each REIO zone and, consequentially, the statistical information with all the administrative data. Second, the multiple procedure of merging all the key databases to yield a *Final Administrative DB* was composed of the following steps:

1. merging Italian NSI census sections and *mainDatabaseRER* through the key fields *Urban Section*, *Registry Sheet* and *Subordinate*
2. merging Italian NSI census sections and *mainDatabaseCT* through the key fields *Urban Section*, *Registry Sheet* and *Subordinate*
3. merging *mainDatabaseRER* and *mainDatabaseCT* through the key field *Id_Property*

Finally, the BRIDGEDB allowed to connect the *Final Administrative DB* to the REIO DB by means of Italian NSI census sections. This is the final outcome of our work, that is FIHRE-DB.

4 Concluding remarks

The importance of the analysis yielded in this work is unique and original in its attempt to provide a new omni comprehensive database that includes all information, both administrative and statistical, related to the analysis of the real estate economy.

The experiment we run to solve the dearth of a complete information because of the practical non-existence of a full information harmonized database can be an important input to replicate the procedure for any dimensional geographical area and for any other economic analysis. As seen in this work, the IT and statistical procedure we describe to create the Full Information Harmonised Real Estate Database is quite complex because of the many issues that need to be faced and the dimension of databases. However, in view of the current Italian government intention to reform the Italian National Registry of private and public properties, our outcome can be an important starting point towards a fair reform.

References

1. Hadford, T.: Big data: a big mistake? Significance, 11, 14–19 (2014).
2. Kitchin, R.: Data, new epistemologies and paradigm shift. Big Data & Society., 1, (2014a).
3. Kitchin, R.: The Data Revolution: Big Data, Open Data. Data Infrastructures and Their Consequences: Sage Publications, London (2014b).
4. Marini C., Nicolardi V.: Big Data and Economic Analysis: The Challenge of a Harmonized Database, Data Science and Social Research II. DSSR 2019. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Nature, (2020).
5. Marini C., Nicolardi V.: Administrative database and official statistics: The case of the real estate analysis. Statistica Applicata - Italian Journal of Applied Statistics - vol. 33 (1), pp.83-95 (2021).

Spatial modeling and Analyses

Spatial statistics analysis using microdata: an application at agricultural sector

Uso dei microdati per l'analisi statistica spaziale: una applicazione al settore agricolo.

Daniela Fusco, Maria Antonietta Liguori, Valerio Moretti, Francesco Giovanni Truglia

Abstract The increasing availability of geo-referenced microdata has led to important changes in the production of official statistics. These changes were been driven also by the need of policy makers to have information related to territorial characteristics. Agriculture is the sector that more than any other is linked to the territory and an analysis of the sector today cannot be separated from spatial analysis. The aim of the research is to analyse, through spatial statistical analysis, the territorial connections of farms, identified at the microdata level, which have survived at the changes of the last twenty years.

Abstract *La disponibilità crescente di microdati georeferiti ha condotto importanti cambiamenti nella produzione delle statistiche ufficiali. Tali cambiamenti sono stati condotti anche dalla necessità dei policy maker di disporre di informazioni legate alle caratteristiche territoriali. L'agricoltura è il settore che più di ogni altro è legato al territorio e una analisi del settore non può prescindere oggi dall'analisi spaziale. Obiettivo della ricerca è analizzare, attraverso l'analisi statistica spaziale, le connessioni territoriali delle aziende agricole, individuate a livello di microdati, sopravvissute ai cambiamenti dell'ultimo ventennio.*

Key words: spatial analysis, microdata, agriculture, autocorrelation

[1] Daniela Fusco, Istat; email: dafusco@istat.it
Maria Antonietta Liguori, Istat; email: liguori@istat.it
Valerio Moretti, Istat; email: vmoretti@istat.it
Francesco Giovanni Truglia, Istat; email: truglia@istat.it

1 Introduction

The new attention to specifying, estimating, and testing for the presence of spatial interaction in the mainstream of applied and theoretical econometrics can be attributed to two major factors [3]: the growing interest in the interaction of an economic agent with other heterogeneous agents in the system and the need to handle spatial data. This has been stimulated by the explosive diffusion of geographic information systems (GIS) and the associated availability of geocoded data.

It was be possible by the ever-increasing availability of statistical and geospatial data [5]. As the volume of geospatial data has been increasing exponentially, the European Statistical System (ESS) has undergone a comprehensive process of reform covering most aspects of its statistical production. This new statistical approach involved each study sector, particularly agriculture. Agriculture is essentially a spatial phenomenon, which is not independent of location. The integration between statistical and geospatial data is fundamental to improve the monitoring of development outcomes: land use and land cover changes are the result of the interplay between socio-economic, institutional and environmental factors [9]. Driving forces, which guide the changes, are generally subdivided into two groups: proximate causes and underlying causes. Proximate causes are the activities and actions that directly affect land use, e.g. wood extraction or road building. Underlying causes are the 'fundamental forces' that underpin the proximate causes, including demographic, economic, technological, institutional and cultural factors [8]. This allows differentiation into spatial and temporal expectations of change.

In Italy, the result of changes is an increase in the average farm size went from 5.5 hectares in 2000 to 8.4 in 2017. The livestock sector too is affected by a rationalization of resources and a conspicuous increase in intensive farming, leading to a progressive growth in the average size of the animals per farm [6]. In this period, some farms, unable to resist at the changes, are died and other were born. The linkage at micro data level identifies a panel of 531,536 farms surveyed for the first time in 2000 and listed in 2017. The aim of the analysis is to explore and describe the territorial dimension of "resilient" farms. The basic hypothesis is that the presence of territorial clusters formed by municipalities with a high density of farms favoured this feature.

2 Methodology: data used, global and local autocorrelation index

The reference panel was obtained through the combination of the Statistical Register of Italian farms (Farm Register 2017) with the results of the V and VI Agricultural Census (2000 and 2010). Only the farms that have not undergone changes of ownership during the reference period were considered, in accordance with the hypothesis that this would be equivalent to a substantial continuity of the

company's mission. In the first instance, the farms survived between the two censuses were identified applying a deterministic model of equality, using the Unique Code Farm as the coupling variable. Then the address and the name were linked applying a function of the distance of the strings via an indicator normalized between 0 and 1. It is defined as twice the number of terms common to compared entities divided by the total number of terms in both tested entities. Finally, the result was deterministically linked to the 2017 Farm Register through the Unique Code Farm only.

Measures of spatial autocorrelation have applied on the dataset. The concept of spatial autocorrelation is one of the most important issues of spatial statistics and it derives directly from the first law of geography by Tobler: "All Things Are Related, But Nearby Things Are More Related Than Distant Things" [12].

In this study, Moran Index (I) in global and local version (Local Indicator of Spatial Association - LISA) has been calculated: it provides an overall measure of Spatial Autocorrelation [10]. Spatial autocorrelation is a property of spatial data that exists whenever there is a systematic pattern in the value recorded at different locations in an area [7]. Moran scatter plot [2] allows achieving a graphic representation of spatial relationships and enables us to investigate possible local agglomerations, LISA allows us to take into account local effects of the phenomenon [1, 11]. In other words, the global index gives a measurement of similarity/dissimilarity of units in relationship with the variable studied. Generally, global index assumes values between 1 and -1. The value 1 means positive autocorrelation: contiguous municipalities have the same density values; the value -1 means negative autocorrelation: contiguous municipalities have different density values. In absent of autocorrelation, the index assumes value -1/N-1.

The local index (LISA) shows the statistical significant units that gives an important contribute ($p<0.05$) at global index construction. Based on the results, the units are dived in four groups:

- HH: municipalities with density higher than average, contiguous to municipalities with density higher than average, too;
- LL: municipalities with density lower than average, contiguous to municipalities with density lower than average, too;
- AL and LA: municipalities with density higher than average, contiguous to municipalities with density lower than average; vice versa for LA.

The first two groups are homogenous geo-statistical configuration; the last two are “encalves”.

3 Results

The figure 1 shows the density of farms survived at last 17 years, defined “resilient”. The density is expressed in quintiles and darken colours highlights a higher concentration: Sud-Tyrrhenian and Adriatic lines, part of Veneto, Friuli-Venezia Giulia, Lazio, Liguria and Piemonte.

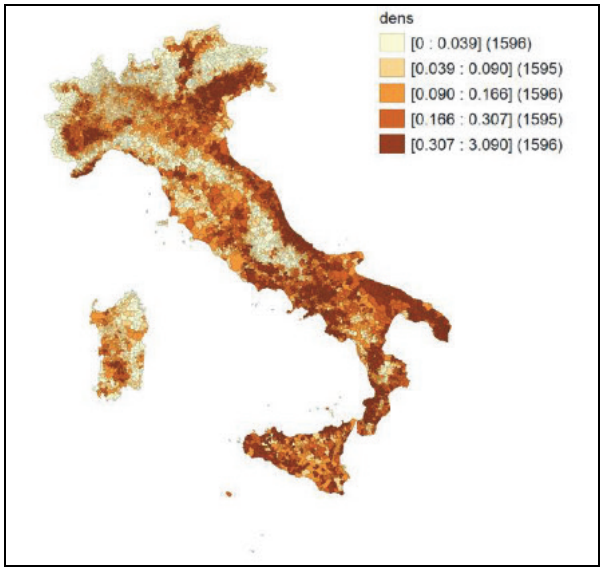


Figure 1: Density distribution of "resilient" farms for municipalities.

Spatial autocorrelation has been visualized adopting Moran scatter plot [4] and considering the density of resilient as abscissa and lagged density of residuals as ordinate. In the graph, Moran Index corresponds to direction coefficient of linear regression, which represents the scatter plot. Positive autocorrelation corresponds to spatial clusters in upper right and lower left quadrants. Lower right and upper left quadrants can classified as spatial outliers.

The Moran's index (figure 2) is 0.585. It's very significant ($p<0,001$): the positive sign and the intensity of this statistic reveal that the municipalities, with a similar density value, are located territorially close to each other.

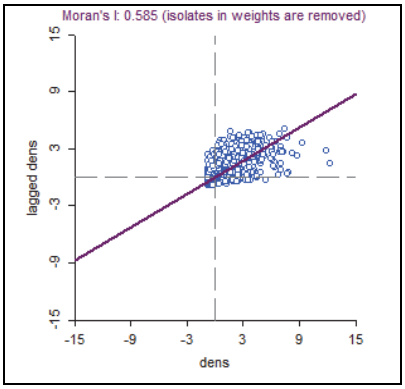


Figure 2: Moran scatter plot.

The scatter plot in figure 2 shows that most of the municipalities are in the first quadrant: municipalities with high density are close to other municipalities with similar density. In any case, the other municipalities are near the intersection of the Cartesian axes.

Figure 3 identifies several clusters of HH municipalities, located in almost all of Puglia; on the coast of Abruzzo and part of Molise; in Veneto and Trentino Alto Adige; in Campania, Calabria and Sicily. Small clusters are in Liguria, Piemonte, Emilia Romagna and Lazio. They are a kind of "hard core" of resilience. LL cluster are located mainly along Alps, Apennines and Sardegna. These areas can be considered as places where this type of structure has not established.

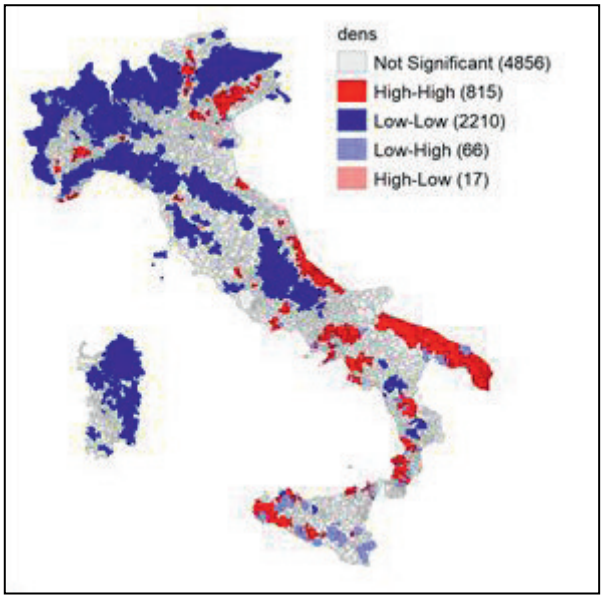


Figure 3: Local Indicator of Spatial Association - LISA

4 Conclusion

The micro data availability allows to explore the mechanisms that influence the farms resistance at the economic changes. A cohort of farms survived at the last 17 years was identified. The results show the importance of territory as a physical support where socio-economic information can be spread out. The results highlight the potential of spatial autocorrelation to prove the surviving spreads via some mechanisms of ‘contamination’ or spatial spill-overs.

There are some differences in the territorial distribution of this phenomenon, not only between north and south, but also between the Tyrrhenian and Adriatic lines where the number of farms survived at the last twenty years is grown.

We choose the Moran index, global and local variant, as a statistical measure of this event [13]. It registers units' autocorrelation with respect to the intensity of a statistical variable (farms surviving) in contiguous territorial units (municipalities). The value of the Moran index is very significant. This situation indicates the presence of a process of contagion that leads to the formation of territorial aggregations with higher than average levels of farms surviving. The LISA index highlights four main types of territorial aggregations, the two most common of which are displayed in blue (Low-Low aggregation) and red (High-High aggregation) where the creation of municipalities' cluster (territorial spill-over) is strongest. In particular, we can note the presence of a great spill-over effect in south of Italy, principally in Puglia region and Abruzzo coast, where the agricultural sector is linked to deep socio-cultural roots. The smaller presence of the spill-over effect in northern Italy probably depends on a greater farm liveliness in these areas, in fact, we can observe little clusters of HH just in North-Est.

References

1. Anselin, L.: "Local Indicators of Spatial Association-LISA". Geographical Analysis 27, (1995)
2. Anselin, L.: GeoDa 0.9 "User's Guide, Spatial Analysis Laboratory." Department of Agricultural and Consumer Economics and CSISS. University of Illinois, p. 125 (2003)
3. Anselin, Luc. "Spatial econometrics". In A Companion to Theoretical Econometrics. Edited by Badi H. Baltagi,pp. 310-330. Blackwell Publishing (2003).
4. Anselin, L.: "The Moran scatterplot as an ESDA tool to assess local instability in spatial association." In Spatial Analytical Perspectives on GIS, pp. 111-126. Edited by Manfred Fischer, Henk J Scholten, and David Unwin. London: Routledge (2019)
5. Eurostat "Merging statistics and geospatial information. Experiences and observations from National Statistical authorities". Luxembourg: Publications Office of the European Union (2019).
6. Fusco, D., Liguori, M. A., Moretti V. "Multisource approach for trends evaluation. An application at the agricultural sector." Paper presented at IES2022 Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, January 27th-28th 2022 at the University of Campania "L. Vanvitelli".
7. Mazziotta, C., Mazziotta, M., De Castris, M. – Spatial Autocorrelation of Alternative Infrastructural Indicator - Atti della XLIII riunione scientifica (2006)
8. Geist, H.J., Lambin E.F,. "Proximate causes and underlying driving forces of tropical deforestation." BioScience 52(2):143–150 (2002).
9. Lesschen, J. P., Verburg, P. H., Staal, S. J. "Statistical methods for analysing the spatial dimension of changes in land use and farming systems" LUCC Report Series 7. Published by:The International Livestock Research Institute, Nairobi, Kenya & LUCC Focus 3 Office, Wageningen University, the Netherlands (2005).
10. Moran, P.A.P.: "The interpretation of statistical map". Journal the Royal Statistical society B, 243–251 (1948)
11. Scardaccione, G., Scorza, F., Las Casas, G., Murgante, B., - Spatial autocorrelation analysis for the evaluation of migration flows: the Italian case - Lecture Notes in Computer Science, vol 6016. Springer, Berlin, Heidelberg. ICCSA (2010)
12. Tobler, W.R.: "A computer movie simulating urban growth in the Detroit region." Economic Geography 46(2), 234–240 (1970)

13. Truglia, F. G, Zeli, A.: “Spatial analysis of economic and social determinants of vote: the case of the European Parliament and constitutional referendum votes in Italy” Italian Political Science Review/Rivista Italiana di Scienza Politica page 1 of 18 doi:10.1017/ipo.2019.29 (2019).

Bayesian spatial modeling of extreme precipitation

Modellazione spaziale bayesiana per precipitazioni estreme

Federica Stolf

Abstract Maps of return levels provide information about the spatial variations of the risk of extreme precipitation and are expected to be useful for infrastructure planning. In this paper we analyze a collection of spatially distributed time series of precipitation in Georgia (USA): exploiting a spatial hierarchical Bayesian model we can produce maps of precipitation return levels with uncertainty measures. Inference about the parameters and spatio-temporal predictions are obtained via Markov Chain Monte Carlo (MCMC) simulation.

Abstract *Le mappe dei valori di ritorno forniscono informazioni importanti sulla variazione spaziale del rischio di precipitazioni estreme e sono utili per la pianificazione delle infrastrutture. In questo contributo analizziamo un insieme di serie storiche di precipitazioni con collocazione spaziale in Georgia (USA): tramite un modello spaziale gerarchico bayesiano possiamo produrre mappe dei valori di ritorno delle precipitazioni con le relative misure di incertezza. L'inferenza sui parametri e le previsioni spazio-temporali sono ottenute tramite simulazioni Markov Chain Monte Carlo (MCMC).*

Key words: Extreme value, Bayesian Hierarchical model, rainfall, Georgia.

1 Introduction

Extreme value theory finds wide application in environmental sciences. Extreme meteorological events such as high rainfall and windstorms arise due to physical processes and are spatial in extent. These events are usually characterized by limited predictability and can cause significant economical and social damages. We mention for example the catastrophic flood that impacted North Georgia, in partic-

Federica Stolf
Department of Statistical Sciences, University of Padova, Padova, Italy.
e-mail: federica.stolf@phd.unipd.it

ular the Atlanta metropolitan area, on September 2009 as a result of multiple days of prolonged rainfall. The flood is blamed for at least 10 deaths and $500 million in damage (National Weather Service). Although these extreme precipitation events are rare, understanding their frequency and intensity is important for public safety and long-term planning.

A rich statistical literature concerned with modeling extreme events is available and [2] provides a comprehensive introduction. Standard approaches utilize generalized extreme value (GEV) distributions [5] and Generalized Pareto Distribution (GPD) [6]. In [4] are identified three main classes of statistical models for spatial extremes: Bayesian hierarchical models, copula based models, and max-stable process models. Latent variable models arise naturally in the Bayesian framework and have been widely used in the context of extremes (see e.g. [3, 1]). A particular issue when dealing with extremes is that although vast amounts of data may be available rare events are necessarily unusual and so the quantity of directly relevant data is limited. One of the advantages of the Bayesian approach is the possibility to incorporate reliable information supplementary to the data in the form of prior distributions. In this paper we focus on the Bayesian method recently proposed in [9], a specification that does not make any asymptotic assumption and explicitly takes into account the spatial dependence of the data.

2 A hierarchical Bayesian Extreme Value Model

Let $x_{ij}(s)$ denote the magnitude of the i-th event within the j-th block for the site s, where $j = 1,\dots,J$ with J the number of blocks in the observed sample, $i = 1,\dots,n_j(s)$ with $n_j(s)$ the number of events observed within the j-th block for the site s and $s = 1,\dots,S$ with S the total number of stations. Traditional approaches, exploiting the asymptotic results stated in [6, 5], focus on the distribution of block maxima for each station $Y_j(s) = \max_i\{x_{ij}(s)\}$ only, discarding the 'ordinary values'. This approach has several limitations. First, the number of events per block may be often not large enough for the asymptotic argument to hold [7] and second the assumption of a constant parent distribution is unrealistic in many contexts. Based on these considerations a hierarchical Bayesian extreme value model that avoids the asymptotic argument and accounts for possible inter annual variability in the magnitude of the events was introduced in [10], building upon the results discussed in [8].

By considering some physical process such as rainfall, one can expect that nearby locations will exhibit similar behavior, and in the Bayesian framework the reduction in uncertainty gained from pooling over space is particularly useful. Thus, spatial modelling of extremes is expected to reduce the overall uncertainty in extreme values estimates, by borrowing strength across spatial locations. For these reasons starting from the approach proposed in [10], [9] include the spatial dependence of the data in the model, incorporating in the layers of the hierarchical model geographical features, to make predictions at unobserved sites. We adopt the latter model speci-

fication, where the events within a block, $x_{ij}(s)$, conditionally on unobserved latent processes are assumed to be conditionally independent with common parametric cdf $F(\cdot;\theta_j(s))$, with $\theta_j(s) \in \Theta$ unknown parameter vector. For technical details see [9].

3 Georgia rainfall data analysis

This study uses daily precipitation observations from the United States Historical Climatological Network (USHCN), a high-quality source of data sets freely available. We use all available stations in Georgia which contain more than 80 years of data from 1892 to 2021. The records characterized by non-blank quality flag were removed, as well as the years characterized by more than 30 daily missing observations. The number of stations included is 20. These data are particularly appropriate for studying heavy precipitation events because of their long-term records, indeed the records for most of the stations are longer than 100 years.

Data are analyzed applying the spatial hierarchical Bayesian model (sHMEV) of [9]. In particular, we consider as geographic covariates to be included in the model, in addition to latitude and longitude, also the elevation since there are some mountainous ranges in the northern part of the region. We fit the model only on the first 20 years of observations for each station and we use the remaining records to validate it.

Adopting a Bayesian methodology allows to make inference on any functional of the posterior distribution (like our target, the cumulative probability of block maxima) and uncertainty measures result naturally from the sampling procedure. A common measure of extreme events is the return level: the r-year return level is the quantile that has probability $1/r$ of being exceeded in a particular year. Figure 1 shows maps of the predictive pointwise posterior mean for the 25 and 100 year return levels, with pointwise 90% credible intervals width. To create these maps the study region was divided into a grid of points, and considering the posterior draws and the values of the covariates for each point is straightforward to obtain draws for the posterior distribution at any grid point. We observe higher return levels for the north mountain area and for the south area, in particular on the southeast of the region near the Atlantic Ocean. There are a few points in the mountainous area with rather high return values and great variability. This is due to the high values of elevation for those points, much greater than for the sites in the data, that lead to a more uncertain extrapolation.

In order to evaluate the performance of the model we compare it with standard alternative methods: Bayesian implementations of the classical generalized extreme valued distribution (GEV) and peak over threshold (POT) Poisson point process models. Since the focus is the right tail of the distribution of the block maxima, we evaluate the predictive accuracy in estimating the true distribution of block maxima on the test set. Figure 2 shows two representative examples. Specifically, the quantile versus return time plots obtained for the different methods for two sites, Waycross and Gainesville, are reported. The fist station is in south Georgia, while

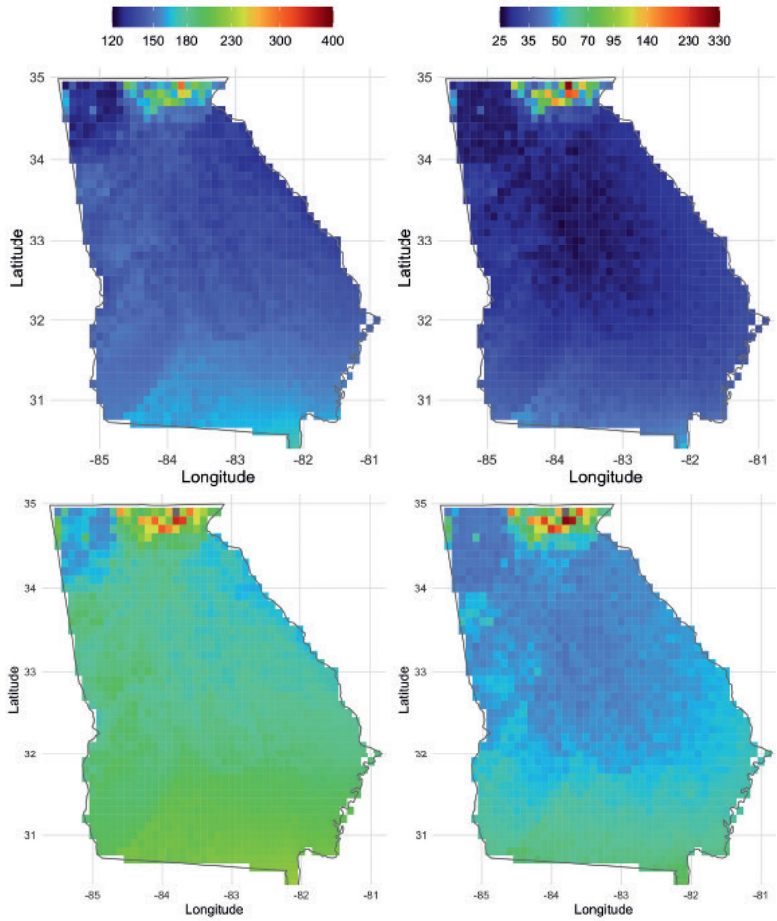


Fig. 1 Maps of the predictive pointwise 25 (top panel) and 100 (bottom panel) year return level estimates for rainfall (mm). Predictive pointwise posterior mean (left) and the width of the 90% pointwise credible intervals (right).

the second one is located north of Atlanta. For both stations the spatial hierarchical model presents an overall good agreement with the empirical frequencies associated to the annual maxima extracted from the entire record, and yields quantile estimates with narrower credibility intervals. POT and GEV appear to be more sensitive to the smallest observations in the training set and tend to underestimate the quantiles. This behavior is expected given the limited length of the records, as observed in [9, 10]. In this case the spatial hierarchical model, exploiting also information from the other sites (*borrowing strength*), manages to obtain more accurate and less variable estimations than the competitors.

This analysis shows the potentiality of the spatial hierarchical models framework for extreme precipitation, although a more detailed comparison via simulated and

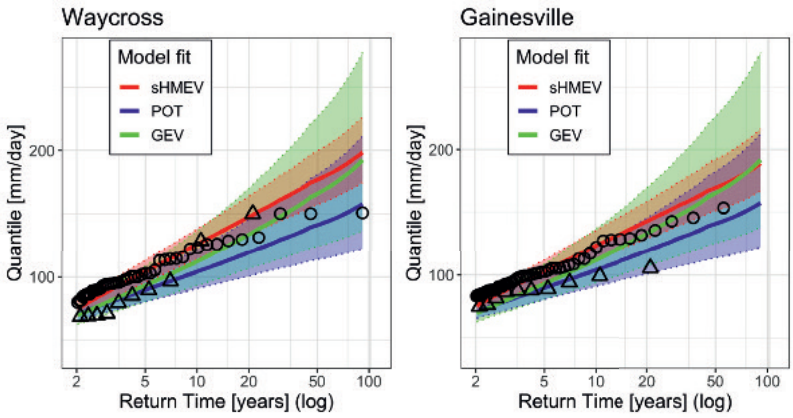


Fig. 2 Quantiles predicted for the stations of Waycross and Gainesville by the GEV (green), POT (blue), and spatial hierarchical Bayesian (red) models. Solid lines show the expected value of the quantile for a given return time, while dashed lines represent the bounds of 90% credibility intervals. Triangles represent the maxima on the training set, while the circles represent the maxima on the test set.

real data sets is needed. A major asset of latent variable models is flexibility: the approach discussed above can be generalized or extended to study different natural phenomena.

References

1. Bracken, C., Holman, K. D., Rajagopalan, B., Moradkhani, H.: A bayesian hierarchical approach to multivariate nonstationary hydrologic frequency analysis. Water Resources Research, **54**(1), 243-255 (2018)
2. Coles, S.: An introduction to statistical modeling of extreme values. Springer-Verlag (2001)
3. Cooley, D., Nychka, D., Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels. Journal of the American Statistical Association, **102**, 824-840 (2007)
4. Davison, A. C., Padoan, S. A., Ribatet, M.: Statistical modeling of spatial extremes. Statistical Science, **27** (2), 161 –186 (2012)
5. Fisher, R. A., Tippett, L. H. C.: Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of the Cambridge Philosophical Society, **24** (2), 180–190 (1928)
6. Gnedenko, B.: Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of Mathematics, **44** (3), 423-453 (1943)
7. Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. Hydrological Sciences Journal, **49**(4) (2004)
8. Marani, M., Ignaccolo, M.: A metastatistical approach to rainfall extremes. Advances in Water Resources, **79**, 121-126 (2015)
9. Stolf, F., Canale, A.: A hierarchical Bayesian non-asymptotic extreme value model for spatial data. Technical report (2022)
10. Zorzetto, E., Canale, A., Marani, M.: Bayesian non-asymptotic extreme value models for environmental data. arXiv:2005.12101 (2020)

A proposal to adjust local Moran's I for measuring residential segregation

Una proposta di modifica dell'indice di Moran locale per misurare la segregazione residenziale

Antonio De Falco, Antonio Irpino

Abstract The study of the spatial distribution of social groups within urban areas has been gaining increasing attention in recent years due to the recent growth of both social inequalities and levels of urban segregation in European countries and cities. Over time, various measures have been proposed to analyze the phenomenon according to its main dimensions of analysis. Among these, detecting the concentration of social groups in specific areas is an essential indicator for revealing the presence of spatial clusters and the social geography of urban areas. The work aims to propose an implementation of the Local Moran's I to integrate some topological aspects of the territories to define a different structure of interaction between spatial units to calculate the index.

Abstract *Lo studio della distribuzione spaziale dei gruppi sociali all'interno delle aree urbane ha guadagnato una crescente attenzione negli ultimi anni a causa della recente crescita dei livelli di disuguaglianze sociali e di segregazione urbana nei paesi e nelle città europee. Nel corso del tempo sono state proposte diverse misure per analizzare il fenomeno secondo le sue principali dimensioni di analisi. Tra queste, il rilevamento della concentrazione di gruppi sociali in specifiche zone è un importante indicatore per rilevare la presenza di cluster spaziali e la geografia sociale delle aree urbane. Il lavoro propone un'implementazione dell'I di Moran locale capace di integrare alcuni aspetti topologici dei territori per definire una diversa struttura di interazione tra unità spaziali per il calcolo dell'indice*.

Key words: residential segregation, LISA, spatial analysis, Google Api

[1] Antonio De Falco, Università di Napoli Federico II; email: antonio.defalco3@unina.it
Antonio Irpino, Università della Campania Luigi Vanvitelli; email: antonio.irpino@unicampania.it

1 Introduction

Residential segregation is generally used to indicate the physical separation between two or more groups defined on relevant social characteristics such as socio-economic status or ethnicity that involve unequal distribution across the urban space. Although not problematic in an absolute sense, residential segregation can represent a severe threat to the social inclusion of disadvantaged groups: this is because the physical and social characteristics of the residential areas, along with an unequal spatial allocation of services and resources, contribute to defining the opportunity/constraint structure of social groups, feeding the r(e)production and transmission of social inequalities (Tammaru *et al.,* 2021).

Since the 1950s, the phenomenon has interested scholars with different disciplinary backgrounds, leading to the development of theoretical models and indices aiming to describe and identify patterns and processes underlying urban segregation. Several indices have been proposed and systematized according to the multidimensional conceptualization of the phenomenon suggested by Massey and Denton (1988). However, in recent times different re-conceptualizations and measures deemed more effective have delineated new analysis approaches (Reardon and O'Sullivan, 2004). Within this framework, the definition of indices that can incorporate the spatial dimension has a particular interest. In fact, they appear to be more effective in capturing different spatial patterns and the variability of the phenomenon across the space.

A proposed way to analyze segregation consists of using spatial clustering techniques based on spatial autocorrelation, such as Geary's C and Moran's I statistics. In particular, Anselin (1995) proposed a local approach to detect spatial clustering by introducing the local indicators of spatial association (LISA) derived by decomposing Moran and Geary indices into several local values for detecting local clusters and spatial outliers. However, critical steps arise when spatial correlation statistics for segregation analysis are used, such as the choice of neighborhoods and the type of spatial relationship among areal units, which significantly affects the results. Further, when summary statistics of the individuals describe spatial units, spatial autocorrelation indices may usually rely on compositional data, vectors whose components are the proportion or percentage of a constant-sum set (respectively 1 or 100). Because compositional data are constrained to be into a *simplex* inducing specific properties, standard statistics designed for unconstrained data – such as correlation coefficient –could lead to biases in the results (Aitchison, 1986).

Starting from this evidence, we propose a modified version of the local Moran's I calculation to investigate socio-economic phenomena and their spatial configuration, such as residential segregation, by introducing two major changes to deal with the issues mentioned above. Firstly, we suggest a different way for defining the structure of the spatial relationship between territorial units by introducing into the spatial weight matrix temporal-distance criteria which consider the spatial connectivity between areal units as a result of different travel time mode distributions (by walking, by car, by public transport, by bicycle). This approach allows us to consider topological features of the study area, trying to give a more realistic representation of

the existing degree of spatial interaction among territorial units than methods based on adjacency or geographical distance criteria. Secondly, since distributions express the variable used for Moran's I calculation, we propose to apply the compositional data analysis approach if the support is categorical (Aitchinson, 1986) or the distributional analysis of variables if the support is quantitative (Irpino and Verde, 2015).
We apply the proposed procedure to Population Census Data collected by the National Institute of Statistics (Istat), the primary statistical source of demographic and socio-economic information, available at high territorial detail.

2 Data and methods

The starting point is the Moran's index (Moran, 1948) of spatial autocorrelation, which is widely used to detect and quantify how a phenomenon of interest is related to itself in space. The existence of an association between the spatial distribution of values of the same variable is known as "spatial dependence." Moran's index can be considered an extension of Pearson's coefficient characterized by a matrix of spatial weights that defines the neighborhood relationships between spatial units, which is necessary to quantify spatial dependence.

The global Moran index measures the spatial autocorrelation in a study area; most often, in reality, spatial autocorrelation reports some level of spatial heterogeneity captured using a local Moran statistic derived from a decomposition of the global index. It can be defined as:

$$I_i = z_i \sum_{j,i \neq j}^{n} w_{ij} z_j \qquad (1)$$

where:

- z_i and z_j are the standardized values of the observed variable in the spatial units *i and j;*
- w_{ij} is the spatial weight matrix that defines the spatial relation between *i* and *j*.

It is worth noting that the Local Moran's index is sensitive to the definition of the relationships between the spatial units partitioning the study area. Since it was conceived for identifying local clusters and local spatial outliers, if the spatial weight matrix does not reflect the actual structure of spatial interactions, the index results may underestimate or overestimate the phenomenon of interest.

Usually, the spatial proximity between units is accounted for through their adjacency or a geographical distance. At the same time, a second step involves the assignment of weights to their proximity links (i.e., inverse distance or negative length of shared boundary), allowing the definition of a spatial weights matrix that is the basis for local Moran's I computation. However, in the study of socio-economic phenomena, relationship systems based on such criteria may not be adequate to catch

the complexity of factors that could influence the connection between territorial units and how and to what extent they interact.

If we focus on urban segregation, it might be helpful to integrate some topological aspects of urban areas, such as road networks and public transportation infrastructures affecting human mobility. Actually, human mobility patterns represent better the degree of connectivity of urban areas (Tóth *et al.*, 2021). In this perspective, a matrix of spatial weights is specified by adopting a criterion based on a temporal dissimilarity that is defined as a function of travel time distributions of people moving from one spatial unit to another. Other authors attempted to model spatial interaction according to road networks using time accessibility (Netrdová and Nosekor, 2020) and travel distances (Qin *et al.*, 2020), but, to our knowledge, there is no similar procedure for analysing socio-economic phenomena such as residential segregation.

In this work, we propose a time-spatial weight matrix, namely, an $n \times n$ positive symmetric matrix W with a generic element w_{ij}, where the weights for each pair of units, respectively i and j, are assigned by pre-set criteria describing the existing spatial connectivity structure among locations.

We consider a temporal similarity as follows:

$$w_{ij} = \frac{1}{t_{ij} + 1}; \qquad (2)$$

where:
t_{ij} is the needed time (in minutes) to get from point i to point j
+1 (minute) is the assumed minimum travel time

To specify the connectivity structure between areal units based on this formulation and incorporate it in local Moran's I calculation, we apply the following procedure using R on both census variables and geographical data released by Istat.

Firstly, we proceed by getting the centroids of spatial polygons and converting them into spatial points; after that, travel time distances can be obtained querying the Distance Matrix API, a Google service that provides travel distance and time between two points or two vectors of points allowing to calculate the travel time for each pair of centroids according to four route modes (bicycle, walking, car, and public transportation); next, we select the median time among the generated time distributions.

Once the time-spatial dissimilarity/distance matrix is created, we use an inverse distance function to define weights in the spatial weight matrix modeling neighborhood relationships: the greater the distance between territorial units, the lesser the connection between them, and vice versa. The result is stored as an instance of an "nb" object and then used as an argument in the global and local Moran's I calculation implemented in the *Spdep* package.

3 Preliminary results and further developments

The method described in the previous section has been applied to a sub-group of Istat Census Area (ACE) 2011, considering the incidence of high-medium skilled professions as a proxy of the middle-upper class. The preliminary results confirm the proposed procedure's validity and encourage applying the modified local Moran's I index to explore the spatial configuration of socio-economic phenomena such as residential segregation to a larger study area (i.e., municipalities or metropolitan areas).

Further analysis considering different ways to establish the spatial relationships between areal units is required to evaluate to what extent the Moran statistic based on dissimilarity distance is more effective in capturing spatial interaction and detecting the presence of autocorrelation in the study area. Additional developments concern the adoption of a strategy based on spatial interpolation techniques to define differently areal units on which the index is calculated to address Modifiable Areal Unit Problem (Openshaw, 1984) and give a more reliable representation of the variable of interest.

For the sake of brevity, we omit the detailed results here.

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman and Hall, London (1986)
2. Irpino A., Verde, R.: Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification **9**, 143–175 (2015)
3. Massey, D.S., Denton, N.A.: The dimension of residential segregation. Social Forces **67**, (2), 281–315 (1988)
4. Moran, P.A.P.: The Interpretation of Statistical Maps. Journal of the Royal Statistical Society. Series B **10**, (2), 243–251 (1948)
5. Netrdová, P., Nosek, V.: Spatial Dimension of Unemployment: Space-Time Analysis Using Real-Time Accessibility in Czechia. International Journal of Geo-Information **9**, (6), 401 (2020)
6. Openshaw, S.: The Modifiable Areal Unit Problem. Concepts and techniques in modern geography. Geobooks, Norwich (1984)
7. Qin, J., Liu, Y., Yi, D. et al.: Spatial Accessibility Analysis of Parks with Multiple Entrances Based on Real-Time Travel: The Case Study in Beijing. Sustainability, 12, (18), 7618 (2020)
8. Reardon, S.F., O'Sullivan, D.: Measures of Spatial Segregation. Sociological Methodology **34**, 121 – 162 (2004)
9. Tóth, G., Wachs, J., Di Clemente, R. et al.: Inequality is rising where social network segregation interacts with urban topology. Nature communications **12**, 1–9 (2021)
10. Tammaru, T., Knapp, D., Silm, S., van Ham, M., Witlox, F.: Spatial Underpinnings of Social Inequalities: A Vicious Circles of Segregation Approach. Social Inclusion **9**, (2), 65–76 (2021)

Accurate Directional Inference for Gaussian Graphical Models

Inferenza Direzionale Accurata per Modelli Grafici Gaussiani

Claudia Di Caterina, Nancy Reid and Nicola Sartori

Abstract Directional tests to compare incomplete undirected graphs are developed in the context of covariance selection for Gaussian graphical models. The exactness of the underlying saddlepoint approximation leads to exceptional accuracy of the proposed approach. This is verified by simulations in settings with high-dimensional parameter of interest, when the accuracy of standard asymptotic approximations to the likelihood ratio test and some of its higher-order modifications fails.

Abstract *Test direzionali per confrontare grafi non orientati incompleti sono sviluppati nell'ambito di selezione della covarianza per modelli grafici Gaussiani. L'esattezza dell'approssimazione saddlepoint coinvolta rende l'approccio proposto estremamente accurato. Ciò è illustrato da simulazioni in cui il parametro di interesse ha dimensione elevata, situazione in cui le approssimazioni asintotiche standard alla distribuzione del test di rapporto di verosimiglianza e di alcune sue modificazioni di ordine superiore falliscono.*

Key words: covariance selection, likelihood-ratio test, undirected graph.

1 Introduction

Undirected graphical models [3, 15] are probabilistic graphs describing complex multivariate distributions of variables (nodes) via the product of simpler sub-models for low-dimensional subsets of the graph (cliques). An important question is inferring the structure of large graphs, i.e. the underlying connections (edges) between

Claudia Di Caterina
Department of Economics, University of Verona, Italy; e-mail: claudia.dicaterina@univr.it

Nancy Reid
Department of Statistical Sciences, University of Toronto, Canada; e-mail: reid@ustat.toronto.edu

Nicola Sartori
Department of Statistical Sciences, University of Padova, Italy; e-mail: sartori@stat.unipd.it

the variables under study. This task is known in the literature by the name of covariance selection, especially if the assumed distribution is normal [9, Sect. 5.2].

Directional inference on a vector-valued parameter of interest was introduced by [6] in nonnormal linear regression models and then generalized in [13]. Substantial progress from both a methodological and computational perspective was made by [5], where the computation of the directional p-value by one-dimensional numerical integration proved especially accurate in several settings. The procedure was extended from linear exponential families to nonlinear parameters of interest in general continuous models by [7]. Besides its accuracy, the directional approach was found to coincide with exact results in several classical situations [10].

Here we show how to employ directional tests for covariance selection in Gaussian graphical models. The directional p-value derived by [5, Sect. 5.3] for testing the complete graph under the saturated model is computationally much simpler. We generalize that to situations where some previous information on the relationships among variables is available. Simulation studies confirm the exceptional performance of the directional approach, even in high-dimensional scenarios where the number of nodes is of the same order of magnitude as the sample size; the likelihood ratio statistic and its higher-order modifications [14] break down in these settings.

2 Likelihood quantities for Gaussian graphical models

Gaussian graphical models are useful for describing normal multivariate distributions. The nodes represent the variables and the lack of an edge between two nodes models conditional independence of the two variables, given the remaining ones. This corresponds to a zero entry in the concentration (inverse covariance) matrix, and covariance selection involves identifying these conditional independencies.

Let $y_1,\dots,y_n$ be a random sample from the normal distribution $N_q(\mu,\Omega^{-1})$, where the mean is $\mu\in\mathbb{R}^q$ and the $q\times q$ concentration matrix Ω is positive definite. For covariance selection the interest is not on the mean, so we focus on the restricted likelihood based on the marginal Wishart distribution of the ML estimator

$$\hat{\Omega}^{-1}=y^\top y/n-y^\top 1_n 1_n^\top y/n^2\,,$$

where y denotes the $n\times q$ matrix with ith row vector $y_i^\top$ and 1_n is a n-vector of ones. Exploiting the symmetry of Ω, we can express the corresponding log-likelihood in the standard exponential family form as

$$\ell(\omega;u)=\frac{n-1}{2}\log|\Omega|-\frac{n-1}{2}\omega^\top Ju\,, \tag{1}$$

where $\omega=\operatorname{vech}\Omega$, $u=n/(n-1)\operatorname{vech}\hat{\Omega}^{-1}$ and the matrix $J=G^\top G$ is diagonal with elements equal to either 1 or 2. If A is a $q\times q$ symmetric matrix, $\operatorname{vec}A$ is the $q^2\times 1$ vector which stacks the columns of A on top of one another, while $\operatorname{vech}A$ retains only the $q^*=q(q+1)/2$ entries in the lower triangle of A. The two vectors are

linked by the relationship $\mathrm{vec}A = G\,\mathrm{vech}A$, which also gives the $q^2 \times q^*$ duplication matrix G [1, Sect. 11.3].

Assume some off-diagonal elements in the concentration matrix are zero, and define the edge sets $k = \{(i,j) : \Omega_{ij} \neq 0, i \leq j\}$ and $h = \bar{k}$. As in [12], we can rearrange the elements of ω, u and the leading diagonal of J such that

$$\omega = \left(\omega_k, \omega_h\right)^{\mathrm{T}}, \quad u = \left(u_k, u_h\right)^{\mathrm{T}}, \quad J = \mathrm{diag}\{J_{kk}, J_{hh}\}. \tag{2}$$

Since $\omega_h = 0$, in the unsaturated model $\Omega = \Omega_k = \Omega(\omega_k)$ so that (1) becomes

$$\ell(\omega_k; u_k) = \frac{n-1}{2}\log|\Omega_k| - \frac{n-1}{2}\omega_k^{\top} J_{kk} u_k, \tag{3}$$

which is a function of the p-dimensional canonical parameter ω_k, with $p \geq q$. Differentiation of (3) with respect to ω_k leads to the score function

$$\ell_{\omega_k}(\omega_k) = \frac{n-1}{2} J_{kk}(\sigma_k - u_k),$$

where σ_k is the partition of $\sigma = \mathrm{vech}\,\Omega_k^{-1}$ obtained as in (2). Solving the score equation leads to $\hat{\sigma}_k = u_k$ and to the corresponding ML estimate $\hat{\omega}_k$, usually derived numerically [5, Sect. 5.3]. As the observed and expected information matrices are equal in canonical exponential families, from [12, Sect. 3] follows that

$$j_{\omega_k\omega_k}(\omega_k) = \frac{n-1}{4} J_{kk}\mathrm{Iss}(\Omega_k^{-1})_{kk} J_{kk}, \tag{4}$$

where $\mathrm{Iss}(\Omega_k^{-1})_{kk}$ is a $p \times p$ partition of the Isserlis matrix [8] of Ω_k^{-1}.

Note that we allow the sample size to be smaller than the number of nodes q, but large enough for the ML estimate $\hat{\omega}_k$ to exist. Particularly, n must be larger than the maximal clique size of the hypothesized graph or of its decomposable version [4, 9, Sect. 5.3.2].

3 Directional and likelihood-based tests for covariance selection

Consider the partition $\omega_k = (\psi, \lambda)^{\mathrm{T}}$ of the canonical parameter, where ψ is the component of interest having dimension $d \leq p - q$. The null hypothesis $H_0 : \psi = 0$ tests whether d additional off-diagonal entries $\Omega_{ij}, i < j$, are zero. Hence, the reduced model is nested in the unsaturated full model of Section 2.

The log-likelihood ratio statistic for testing H_0 results [9, Sect. 5.2.2]

$$w(\psi_0) = -(n-1)\log|\hat{\Omega}_k^{-1}\hat{\Omega}_0|, \tag{5}$$

where $\hat{\Omega}_k = \Omega(\hat{\omega}_k)$ is the ML estimate of Ω obtained from (3), and $\hat{\Omega}_0 = \Omega(\hat{\omega}_{k0})$ is its constrained ML estimate under H_0, given $\hat{\omega}_{k0} = (0, \hat{\lambda}_0)^{\mathrm{T}}$. The null asymptotic

distribution of (5) is χ^2_d, as is that of both higher-order modifications

$$w^*(\psi_0) = w(\psi_0)\left\{1 - \frac{\log\gamma(\psi_0)}{w(\psi_0)}\right\}^2 \quad \text{and} \quad w^{**}(\psi_0) = w(\psi_0) - 2\log\gamma(\psi_0) \quad (6)$$

introduced by [14], being the term $\gamma(\psi_0)$ in this case equal to

$$\gamma(\psi_0) = \frac{2\{(\hat\sigma_{k0}-\hat\sigma_k)^\top \mathrm{Iss}(\hat\Omega_0^{-1})^{-1}_{kk}(\hat\sigma_{k0}-\hat\sigma_k)\}^{d/2}}{\{-\log|\hat\Omega_k^{-1}\hat\Omega_0|\}^{d/2-1}(\hat\omega_k-\hat\omega_{k0})^\top J_{kk}(\hat\sigma_{k0}-\hat\sigma_k)}\left\{\frac{|\mathrm{Iss}(\hat\Omega_0^{-1})_{kk}|}{|\mathrm{Iss}(\hat\Omega_k^{-1})_{kk}|}\right\}^{1/2}.$$

Due to limited space, we refer to [5] for a thorough derivation of the general formula for the directional p-value when testing hypotheses on a parameter of interest which is linear in the canonical parameterization of an exponential family. We shall thus obtain the relevant quantities in the special context of this paper. First, we find

$$s_\psi = -\ell_{\omega_k}(\hat\omega_{k0}) = \frac{n-1}{2}J_{kk}(u_k - \hat\sigma_{k0}),$$

where $\hat\sigma_{k0}$ is the component of $\hat\sigma_0 = \mathrm{vech}\,\hat\Omega_0^{-1}$. Then, the so-called tilted log-likelihood function along the line $s(t) = (1-t)s_\psi$ follows from (3):

$$\ell\{\omega_k; s(t)\} = \frac{n-1}{2}\log|\Omega_k| - \frac{n-1}{2}\omega_k^\top J_{kk}\{\hat\sigma_{k0} + t(u_k - \hat\sigma_{k0})\}. \quad (7)$$

Maximization of (7) entails $\hat\Omega_k^{-1}(t) = t\hat\Omega_k^{-1} + (1-t)\hat\Omega_0^{-1}$, and by taking the inverse $\hat\omega_k(t)$ is obtained accordingly. The replacement of ω_k in (7) with $\hat\omega_k(t)$ and $\hat\omega_{k0}$, respectively, delivers after some manipulation the result

$$\exp[\ell\{\hat\omega_{k0}; s(t)\} - \ell\{\hat\omega_k(t); s(t)\}] \propto |\hat\Omega_k(t)|^{-\frac{n-1}{2}}.$$

By (4), we get $|j_{\omega_k\omega_k}\{\hat\omega_k(t); s(t)\}|^{-1/2} \propto |\mathrm{Iss}\{\hat\Omega_k^{-1}(t)\}_{kk}|^{-1/2}$ and consequently

$$\begin{aligned} h\{s(t);\psi_0\} &\propto \exp[\ell(\hat\varphi^0_\psi; s) - \ell\{\hat\varphi(s); s\}]\,|j_{\varphi\varphi}\{\hat\varphi(s); s\}|^{-1/2} \\ &\propto |\hat\Omega_k^{-1}(t)|^{\frac{n-1}{2}}|\mathrm{Iss}\{\hat\Omega_k^{-1}(t)\}_{kk}|^{-1/2}. \end{aligned} \quad (8)$$

The directional measure of deviation from H_0 in the direction indicated by the data [5] is computed as

$$p(\psi_0) = \frac{\int_1^{t_{\max}} t^{d-1}h\{s(t);\psi_0\}\,\mathrm{d}t}{\int_0^{t_{\max}} t^{d-1}h\{s(t);\psi_0\}\,\mathrm{d}t}. \quad (9)$$

The value $t_{\max}$ is the largest t such that $\hat\Omega_k(t)$ is positive definite, and is determined numerically. The saddlepoint approximation (8), when normalized, is the exact conditional distribution of $s(t)$ in the multivariate normal setting [2, p. 189]. Normalization occurs implicitly in the ratio (9), so the only approximation error involved in the calculation of the directional p-value $p(\psi_0)$ stems from the numerical integration.

Table 1 Empirical null p-value distributions (%) based on 100 000 replications for testing Markov dependence of order 1 versus different orders $m \in \{9, 18, 28\}$, with $n = 60$ and $q = 30$.

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
vs MD(9), $d = 196$											
Likelihood ratio, (5)	11.1	19.1	28.4	41.5	64.6	84.6	95.3	98.7	99.5	99.8	99.9
Skovgaard's w^*, (3)	0.3	0.9	2.0	4.4	13.3	32.3	57.9	78.5	87.1	92.5	96.4
Skovgaard's w^{**}, (3)	0.3	0.8	1.7	3.9	12.1	30.2	55.4	76.5	85.7	91.4	95.7
Directional, (9)	0.9	2.3	4.8	9.7	24.7	50.3	75.8	90.5	95.4	97.7	99.1
vs MD(18), $d = 340$											
Likelihood ratio, (5)	53.8	66.9	76.9	86.0	95.0	98.8	99.8	100.0	100.0	100.0	100.0
Skovgaard's w^*, (3)	0.0	0.1	0.3	0.7	3.0	10.7	27.4	48.8	62.1	72.8	82.9
Skovgaard's w^{**}, (3)	0.0	0.0	0.1	0.4	1.7	6.9	19.5	38.2	51.2	62.5	74.3
Directional, (9)	0.8	2.2	4.6	9.5	24.7	50.2	76.0	90.8	95.6	97.8	99.2
vs MD(28), $d = 405$											
Likelihood ratio, (5)	86.2	92.3	95.6	97.9	99.5	99.9	100.0	100.0	100.0	100.0	100.0
Skovgaard's w^*, (3)	0.0	0.0	0.0	0.2	0.9	4.3	13.8	30.0	42.5	53.9	67.0
Skovgaard's w^{**}, (3)	0.0	0.0	0.0	0.0	0.2	1.4	5.9	15.5	24.5	33.9	46.4
Directional, (9)	1.0	2.4	5.1	10.1	25.2	50.1	75.1	90.1	95.1	97.5	99.0
Standard error	0.0	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.0

4 Simulation studies

The performance of the directional approach is presented via Monte Carlo studies based on 100 000 replications, along with that of the test statistics $w(\psi_0)$, $w^*(\psi_0)$ and $w^{**}(\psi_0)$ expressed in Section 3.

To enable comparison with [5, Sect. 5.3], samples of size $n = 60$ are drawn from a q-variate Gaussian distribution under the hypothesis of first-order Markovian dependence (MD). For each $q \in \{11, 30, 50\}$, this null model is tested against higher MD orders for the concentration matrix. These unsaturated alternative models are chosen to check the sensitivity of the various methods to the number d of parameters of interest. Table 1 reports results obtained for $q = 30$. The empirical null distribution of the directional p-value is extremely accurate in all settings, almost unaffected by the size of d; on the contrary, its competitors break down as d increases.

A second simulation study considers q-variate Gaussian samples of size $n \in \{60, 90, 120\}$, with $q = 50$ and block diagonal concentration matrix with two main-diagonal blocks. The larger unsaturated model allows instead for some nonzero entries in the off-diagonal blocks, so that $d = 250$. Figure 1 shows the exceptional agreement with the uniform benchmark of the empirical directional p-values, as opposed to the classical likelihood-based alternatives that are found to be strongly influenced by the value of n.

Despite being intrinsically a conditional method, the directional p-value proves to enjoy extremely good marginal properties in this high-dimensional framework, particularly compared to the omnibus tests which quantify the observed departure from H_0 taking all directions of the parameter space into account. A future line of research could extend the same approach to graphical models for discrete data [11].

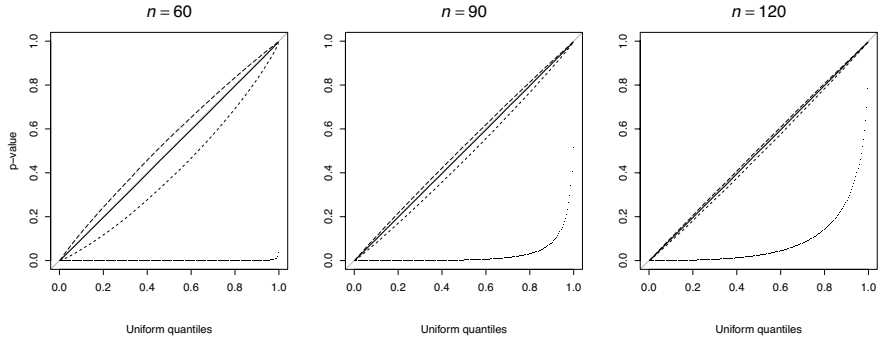


Fig. 1 Results from 100 000 Gaussian samples simulated with block diagonal concentration matrix and $q = 50$. The empirical p-values obtained via w (dot-dashed), w^* (dashed), w^{**} (long-dashed) and the directional test (solid) are compared with the uniform distribution given by the gray diagonal for $d = 250$ and several sample sizes: $n = 120$ (left), $n = 90$ (middle) and $n = 60$ (right).

References

1. ABADIR, K. M., AND MAGNUS, J. R. *Matrix Algebra*. Cambridge University Press, 2005.
2. BARNDORFF-NIELSEN, O. E., AND COX, D. R. Edgeworth and saddle-point approximations with statistical applications (with Discussion). *Journal of the Royal Statistical Society: Series B 41* (1979), 279–312.
3. BORGELT, C., AND KRUSE, R. *Graphical Models: Methods for Data Analysis and Mining*. John Wiley & Sons, 2002.
4. BUHL, S. L. On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics 20* (1993), 263–270.
5. DAVISON, A. C., FRASER, D. A. S., REID, N., AND SARTORI, N. Accurate directional inference for vector parameters in linear exponential families. *Journal of the American Statistical Association 109* (2014), 302–314.
6. FRASER, D. A. S., AND MASSAM, H. Conical tests: Observed levels of significance and confidence regions. *Statistische Hefte 26* (1985), 1–17.
7. FRASER, D. A. S., REID, N., AND SARTORI, N. Accurate directional inference for vector parameters. *Biometrika 103* (2016), 625–639.
8. ISSERLIS, L. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika 12* (1918), 134–139.
9. LAURITZEN, S. L. *Graphical Models*. Oxford University Press, Oxford, 1996.
10. MCCORMACK, A., REID, N., SARTORI, N., AND THEIVENDRAN, S. A. A directional look at f-tests. *The Canadian Journal of Statistics 47* (2019), 619–627.
11. ROVERATO, A. *Graphical Models for Categorical Data*. Cambridge University Press, 2017.
12. ROVERATO, A., AND WHITTAKER, J. Standard errors for the parameters of graphical Gaussian models. *Statistics and Computing 6* (1996), 297–302.
13. SKOVGAARD, I. M. Saddlepoint expansions for directional test probabilities. *Journal of the Royal Statistical Society: Series B 50* (1988), 269–280.
14. SKOVGAARD, I. M. Likelihood asymptotics. *Scandinavian Journal of Statistics 28* (2001), 3–32.
15. WHITTAKER, J. *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons, 2009.

Advances in Classification

Measures of interrater agreement based on the standard deviation

Misure di accordo tra valutatori basate sulla deviazione standard

Giuseppe Bove

Abstract Interrater agreement for measurements on quantitative (discrete or continuous) scales is usually evaluated by overall measures across subjects like the intraclass correlation coefficient. In this paper, a new index that allows to evaluate the agreement between two or more raters for each single case (subject or object) is presented, and a global measure for the whole group of cases is obtained as the arithmetic average of the single case values. Some advantages of the proposed indices are highlighted by an application related to the measurement of tumor sizes.

Abstract *L'accordo assoluto tra misurazioni su scala quantitativa (discreta o continua) è valutato generalmente con misure globali di accordo come il coefficiente di correlazione intraclasse. In questo lavoro viene presentato un nuovo indice che consente di valutare l'accordo tra due o più valutatori per ogni singolo caso (soggetto o oggetto), e di ottenere anche una misura globale dell'accordo tra i valutatori per l'intero gruppo di casi valutati. Alcuni vantaggi degli indici proposti sono evidenziati con un'applicazione relativa alla misurazione della dimensione dei tumori.*

Key words: interrater agreement, quantitative measurement, tumor measurement

1 Introduction

Agreement among measurements provided by two or more raters (humans or devices) is considered in applications regarding education, psychometrics, biomedical sciences, and other disciplines. For instance, the agreement among clinical diagnoses

[1] Giuseppe Bove, Dipartimento di Scienze della Formazione, Università Roma Tre; giuseppe.bove@uniroma3.it

provided by more physicians on a nominal scale is analysed for identifying the best treatment for the patient, or the agreement among ratings of educators who assess on a new ordinal rating scale the language proficiency of a corpus of argumentative (written or oral) texts is considered to test reliability of the new scale. Besides, agreement among quantitative measurements can be analysed, for instance, when radiologists measure the tumor size of lung cancer patients who could be considered in a clinical trial. In all these applications, the main interest is to analyse interrater absolute agreement, that is the extent raters assign the same (or very similar) values on the rating scale.

Many indices have been proposed to measure the agreement between ratings or measurements of two or more raters on a whole group of subjects (objects) (for a review see, for example, Shoukri (2011), Broemeling (2009), von Eye and Mun (2005)). Moreover, methods to detect subsets of raters who demonstrate a high level of interobserver agreement were considered, for instance, by Landis & Koch (1977). Less frequently agreement on a single subject has been considered (O'Connell & Dobson, 1984), in spite of the fact that having evaluations of the agreement on the single case is particularly useful, for example, in situations where the rating scale is being tested, and it is necessary to identify any changes to improve it, or to request the raters for a specific comparison on the single case in which agreement is poor. Recently, Bove *et al.* (2021) and Bove (2021) proposed subject-specific measures of absolute agreement for ratings provided on an ordinal and nominal scale, respectively. In the next sections, indices measuring the interrater agreement for quantitative measurements on a single subject based on the standard deviation are proposed. Furthermore, a global measure of agreement on the whole group of subjects obtained as the arithmetic average of the subject values of the indices will be also considered and applied to a data set concerning the agreement of radiologists measuring tumor size.

2 Method

There are several well-known methods that are employed to reveal the patterns of agreement between scores provided for a group of subjects (objects) by raters using a quantitative scale X. For quantitative discrete scales with a limited number of levels, weighted versions of kappa-type indices are available (e.g., Gwet (2014)). For quantitative continuous scales, the intraclass correlation coefficient is the traditional approach (e.g., McGraw and Wong (1996)), by which the agreement on the whole group of subjects can be evaluated using mean squares from an analysis of variance (ANOVA). Another very common and simple approach is based on the graphical method proposed by Bland and Altman (1986), according to which, for each pair of raters, differences in the paired scores versus the mean of the paired scores are plotted to determine if the fitted line is flat and if the observations are within two standard deviations of the mean.

A different approach based on a regression framework is considered in this section to propose subject-specific measures of absolute agreement between scores on a

quantitative scale. A global measure of agreement is also obtained as the average of the values of the subject-specific measure. We consider first the case of two raters and then the case of several raters.

Two raters

The scores of two raters (X_1, X_2) can be depicted in a scatter plot, the raters are in perfect agreement $(X_1 = X_2)$ if the regression goes through the origin with a slope of 1. Normally, the points in the scatter plot deviate from perfect agreement, so the level of agreement can be explored as the dispersion around the straight line $x_1 = x_2$. For subject *i* having the scores x_{i1}, x_{i2} $(i = 1,2, \dots, n)$, the distance l_i of the point (x_{i1}, x_{i2}) from the straight line $x_1 = x_2$ is

$$l_i = \sqrt{(\sum_{j=1}^{2} x_{ij}^2) - \frac{(\sum_{j=1}^{2} x_{ij})^2}{2}} = \sqrt{2s_i^2}$$

with s_i^2 the variance of the scores x_{i1}, x_{i2}. When min(X) and max(X) are known for the quantitative scale X, a subject-specific measure of absolute agreement g_i can be defined as

$$g_i = 1 - \frac{l_i}{\max(l_i)} = 1 - \frac{s_i}{(\max(X) - \min(X))/2}$$

with $g_i = 0$ for maximum disagreement and $g_i = 1$ for perfect agreement. When the minimum and maximum of X are unknown, a relative measure of agreement can be obtained (disregarding the constant $\sqrt{2}$) by the coefficient of variation $CV_i = \frac{s_i}{\bar{x}_i}$, with $\bar{x}_i$ the arithmetic average of the scores x_{i1}, x_{i2}, (or the absolute value if $\bar{x}_i < 0$).

Global measures of agreement are obtained by averaging the subject-specific values

$$g_{ave} = \frac{1}{n} \sum_{i=1}^{n} g_i \quad , \quad CV_{ave} = \frac{1}{n} \sum_{i=1}^{n} CV_i$$

for the two cases, respectively.

More raters

The scores of *k* raters $(X_1, X_2, \dots, X_k)$ are now represented in *k* dimensions, so the level of agreement can be explored as the dispersion around the straight line $x_1 = x_2 = \cdots \dots = x_k$. For subject *i* having the scores $x_{i1}, x_{i2} \dots, x_{ik}$ $(i = 1,2, \dots, n)$, the distance l_i of point $(x_{i1}, x_{i2} \dots, x_{ik})$ from the straight line $x_1 = x_2 = \cdots \dots = x_k$ is

$$l_i = \sqrt{(\sum_{j=1}^{k} x_{ij}^2) - \frac{(\sum_{j=1}^{k} x_{ij})^2}{k}} = \sqrt{k s_i^2}$$

with s_i^2 the variance of the scores $x_{i1}, x_{i2} \ldots, x_{ik}$. Formulas for g_i and g_{ave} remain the same.

Generalizations of the coefficient of variation CV_i and the average CV_{ave} are immediate for any k.

Some experiences with real applications suggest the following indications of thresholds for the interpretation of CV_i and CV_{ave} expressed in percentage: values less than or equal to 10% excellent agreement, values between 10% and 20% good agreement, values greater than 20% fair to poor agreement.

3 Application

A discussion of disagreement between radiologists is considered in Erasmus et al. (2003). The purpose of the study was to assess the consistency of measurements performed by readers evaluating lung tumors, this is a critical component of many cancer trials because measurements can be used, for instance, to warrant additional testing of an agent or to determine whether therapy should be continued.

Patients were selected with non-small-cell lung cancer and with 40 lung lesions whose size exceeded at least 1.5 cm. Measurements were performed independently by five thoracic radiologists using printed film by computed tomography. Each radiologist read each of 40 images measuring a) the longest diameter (unidimensional measurement) and b) the longest diameter and the perpendicular longest diameter (bidimensional measurement) of each lesion. Measurement was repeated on the same images 5–7 days later, that means each radiologist was looking at the same image (and the same tumor size) twice (Table 6.18 in Broemeling (2009) contains the data of the two replications of the unidimensional measurements). Variation between and within (the two replications of) the five radiologists measuring tumor size is examined in Erasmus et al. (2003) with different methods, in order to ascertain how to improve measurement consistency in terms of reduction of incorrect interpretation of tumor response.

In the following, we are only interested in analysing agreement applying the proposed indices CV_i and CV_{ave} to the unidimensional measurements of the five radiologists in the first and second replication. Some descriptive statistics regarding the first replication of the unidimensional measurement are provided in Table 1. The

similarity of the means in Table 1 reflects a pretty good level of agreement, with radiologist 2 reporting the smallest mean tumor size and the smallest standard deviation. The level of agreement is confirmed by the values of the intraclass correlation coefficient (random factorial design assumed) and the CV_{ave} provided in Table 2, where it can also be noted that at least half of the forty CV_i values are less or equal to 10% in both the repetitions.

Table 1: Descriptive statistics of tumor measurements (centimeters) of five radiologists on *n*=40 lung lesions (first replication)

Radiologist	*Mean*	*Median*	*Range*	*SD*
1	3.92	3.80	1.5 to 8	1.59
2	3.71	3.80	1.2 to 7.8	1.50
3	4.42	4.20	1.5 to 9	1.54
4	4.37	4.10	1.5 to 9	1.60
5	4.14	3.95	1.7 to 9	1.54
Total	**4.11**	**4.00**	**1.2 to 9**	**1.58**

On the other hand, some high CV_i values (not reported) are present in each repetition (e.g., images 14, 16, 30, 34), these images could be selected for a comparison between radiologists and to detect particular types of lesions (irregular edge and/or irregular contour) difficult to measure.

Table 2: Intraclass correlation values and descriptive statistics of the distributions of the coefficient of variation (percentage) in repetitions 1 and 2

Repetition	*Intraclass correlation*	CV_{ave}	CV_{min}	CV_{max}	$CVi \leq 10\%$	$10\% < CVi \leq 20\%$	$CVi > 20\%$
1	0.837	12.89	1.27	39.24	20	13	7
2	0.842	12.61	0	36.23	23	8	9

In Erasmus et al. (2003) the measurements referring to the same radiologist in the two replications (intraobserver) and the measurements referring to two different radiologists in the same repetition (interobsever) were compared for each tumor to detect the presence of misclassifications (differences in the measured size of tumors higher than a predefined cutoff). The level of intraobserver misclassifications was lower than the level of interobserver misclassifications, so it was adviced that the same reader should perform measurements for any one patient for the duration of the clinical cancer trial.

The indices CV_i and CV_{ave} were computed for each radiologist in the two replications and for each pair of radiologists in each replication. CV_{ave} values are provided in Table 3. It can be noted that off-diagonal values are higher than diagonal values in Table 3, and this confirms that intraobsever measurements have less dispersion than interobserver measurements.

Table 3: Pairwise intraobserver and interobserver CV_{ave} values (percentage) of five radiologists

Radiologist	1	2	3	4	5
1	4.26				
2	8.80	4.54			
3	9.30	13.03	3.47		
4	6.84	10.55	7.39	4.36	
5	6.06	9.69	7.84	5.26	4.20

4 Conclusions

This contribution provides new subject-specific measures of interrater agreement, extending to scores on quantitative scales the proposals for categorical data presented in Bove et al. (2021) and Bove (2021). The approach is mainly descriptive and it is based on a geometrical formalization. Preliminary results obtained in the application to the measurement of tumor sizes seem encouraging. Future developments may concern a study of the sampling properties of the proposed indices.

References

1. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurements. The Lancet, 327 (8476), 307-310 (1986)
2. Bove, G.: A subject-specific measure of interrater agreement based on the homogeneity index. In: Porzio G.C. et al. (eds.), CLADAG 2021 Book of short papers, pp. 272-275. Firenze University Press, Firenze (2021)
3. Bove, G., Conti, P.L., Marella, D.: A measure of interrater absolute agreement for ordinal categorical data. Statistical Methods & Applications, 30 (3), 927-945 (2021)
4. Broemeling, L.D.: Bayesian methods for measures of agreement. Chapman & Hall/CRC, London (2009)
5. Erasmus, J. J., Gladish, G. W., Broemeling, L., Sabloff, B. S., Truong, M. T., Herbst, R. S. , Munden, R. F.: Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. J. Clin. Oncol. 21(13), 2574-2582 (2003)
6. Gwet, K.L.: Handbook of inter-rater reliability (4-th ed.), Advanced Analytics, LLC, Gaithersburg MD, USA (2014)
7. Landis, J.R., Koch, G.G.: An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. Biometrics, 33, 363-374 (1977)
8. McGraw, K. O., Wong, S. P.: Forming inferences about some intraclass correlation coefficients. Psychological Methods*, 1*(1), 30-46 (1996)
9. O'Connell, D.L., Dobson, A.J.: General Observer-Agreement Measures on Individual Subjects and Groups of Subjects. Biometrics, 40 (4), 973-983 (1984)
10. Shoukri, M. M.: Measures of interobserver agreement and reliability. Taylor and Francis Group, Boca Raton, Florida (2011)
11. von Eye, A., Mun, E. Y.: Analyzing rater agreement. Manifest variable methods. Lawrence Erlbaum Associates, Mahwah, New Jersey (2005)

A comparison of accuracy measures for classification tasks

Confronto tra misure di accuratezza delle classificazioni

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract Evaluating the ability of a classifier for making predictions on unseen data and increasing it by tweaking the learning algorithm are two of the main reasons motivating the evaluation of classifier predictive performance. While the commonly adopted performance measures are suitable for balanced data sets, in the case of data imbalances it is suggested to adopt a balanced performance measure. In this study the properties of the Balanced AC_1 — a recently introduced classifier accuracy measure — are investigated under different class imbalance conditions via a Monte Carlo simulation. The study results reveal the suitability of Balanced AC_1 with both balanced and imbalanced data sets.

Abstract *Valutare la capacità di un classificatore di processare nuovi dati e cercare di incrementare tale capacità migliorando l'algoritmo di apprendimento sono due delle principali ragioni per cui è necessario misurare le prestazioni di un classificatore. Mentre le misure di prestazione comunemente utilizzate sono adatte a dati ugualmente distribuiti tra le classi, in caso di dati sbilanciati è suggerito l'utilizzo di misure bilanciate. Attraverso uno studio in simulazione Monte Carlo, questo lavoro vuole investigare le proprietà del Balanced AC_1 — una misura di accuratezza del classificatore introdotta recentemente — in diverse condizioni di sbilanciamento dei dati. I risultati ottenuti rivelano l'adeguatezza del Balanced AC_1 sia con dati bilanciati che sbilanciati.*

Key words: Classifier predictive performance, Accuracy, Balanced AC_1, Monte Carlo Simulation

[1]Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; email: amalia.vanacore@unina.it

[2]Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; e-mail: mariasole.pellegrino@unina.it

1 Introduction

Several measures have been proposed in the specialized literature for evaluating classifier predictive performance (see [5, 6] for an overview). The most commonly used are the measures determined on the basis of confusion matrix, a cross table that records how cases are distributed over predicted (on columns) and actual (on rows) classes in such a way that cells on the main diagonal count the correctly classified cases whereas off-diagonal cells the cases incorrectly assigned to the class. Each measure deals with a different performance aspect so that the appropriate classifier performance measure for the problem at hand is generally chosen according to the performance aspect to be investigated.

Sensitivity (or recall), precision, specificity and accuracy are four of the commonly used classifier performance measures based on confusion matrix. In the most popular classification task with 2 non-overlapping classes, sensitivity and precision focus on the performance on positive class, specificity focuses on the performance on negative class and accuracy focuses on overall effectiveness. Specifically, sensitivity counts the proportion of cases predicted as positive that are truly positive related to the total count of truly positive cases whereas precision counts the proportion of cases predicted as positive that are truly positive related to the total count of cases predicted as positive. Specificity, instead, counts the proportion of cases predicted as negative that are truly negative related to the total count of truly negative cases. Whereas accuracy, by far the most widespread performance measure, is defined as the number of both positive and negative successful predictions relative to the total number of classifications.

Despite its popularity, a main criticism raised against accuracy is that it fails to compensate the non-zero probability that some classifications match only by chance. To fill this gap, the Cohen's K coefficient [1], a κ-type coefficient introduced in social and behavioral sciences for measuring the degree of agreement between two raters, has been adopted in the last decades as an alternative measure of classifier accuracy. However, its adoption is still debated since the probability of classifications matching by chance is estimated through marginal frequencies in such a way that a reasonably high proportion of observed agreement is converted into a much smaller coefficient value when the marginal frequencies are unequal [2]. A variant of Cohen's K is the AC_1 proposed by Gwet [3] who formulates the chance agreement term as independent from trait prevalence; this allows the coefficient to be recognized as a reliable measure of classifier predictive performance [4].

Moreover, in the case of imbalanced data sets, common in many real world classification problems, it is suggested to adopt a balanced performance measure which averages the performance values estimated for each class and thus treats classes equally avoiding the dependency over the majority class. The recently introduced Balanced AC_1 [7] is an accuracy measure able to treat both classes equally while compensating the non-zero probability that some classifications match only by chance.

This research work aims at investigating, via a Monte Carlo simulation study, the statistical behavior of Balanced AC_1 for binary classification tasks with different

class imbalance conditions. The performance of Balanced AC_1 is compared against that of the commonly adopted accuracy measures. It is worth to note that our simulation study has been conceived to generate confusion matrices with known randomness rather than generating synthetic data from specific classification algorithms.
The remainder of the paper is organized as follows: in Section 2 the classifier accuracy measures under study are introduced; in Section 3 the Monte Carlo simulation algorithm and the main results are fully discussed; finally, conclusions are summarized in Section 4.

2 Classifier accuracy measures

Let n be the number of cases classified on $k = 2$ non-overlapping classes during a binary classification task. The provided classifications are arranged in a 2×2 confusion matrix reported in Table 1.

Table 1 2×2 confusion matrix

		Predicted		
		Positive (+)	Negative (−)	TOTAL
Actual	Positive (+)	*tp*	*fn*	AP
	Negative (−)	*fp*	*tn*	AN
	TOTAL	PP	PN	*n*

The widespread accuracy is formulated as follows:

$$\text{Accuracy} = \frac{tp + tn}{n} \tag{1}$$

Cohen's K coefficient [1] compensates the effect of classifications matching by chance by correcting the observed agreement (i.e. accuracy in Eq. 1) as reported in the following:

$$K = \frac{(tp + tn)/n - (\text{AP} \cdot \text{PP} + \text{AN} \cdot \text{PN})/n^2}{1 - (\text{AP} \cdot \text{PP} + \text{AN} \cdot \text{PN})/n^2} \tag{2}$$

Gwet's AC_1 coefficient [3] quantifies the probability of agreement between two series of evaluations *exclusively* on the cases not susceptible to agreement by chance. According to Gwet, the cases whose classification is not certain are difficult to classify and these are the only cases that could lead to chance agreement. This notion of chance agreement let the Gwet's AC_1 coefficient be formulated as follows:

$$AC_1 = \frac{(tp + tn)/n - 2\left[\frac{\text{AP}+\text{PP}}{2n}\left(1 - \frac{\text{AP}+\text{PP}}{2n}\right)\right]}{1 - 2\left[\frac{\text{AP}+\text{PP}}{2n}\left(1 - \frac{\text{AP}+\text{PP}}{2n}\right)\right]} \tag{3}$$

The balanced version of Accuracy suitable to deal with imbalanced data sets is given by:

$$\text{Balanced Accuracy} = \left(\frac{tp}{\text{AP}} + \frac{tn}{\text{AN}}\right) \cdot \frac{1}{2} \tag{4}$$

The Balanced AC_1 [7] is formulated as a relative measure that corrects the Balanced Accuracy with the balanced proportion of classifications matching by chance obtained by averaging the values estimated for each class:

$$\text{Balanced } AC_1 = \frac{\text{Balanced Accuracy} - p_{a|c}^{AC_1}}{1 - p_{a|c}^{AC_1}} \tag{5}$$

where $p_{a|c}^{AC_1}$ is estimated as follows:

$$p_{a|c}^{AC_1} = \frac{\left[\frac{tp}{\text{AP}} \cdot \left(\frac{tp}{\text{PP}}\left(1 - \frac{tp}{\text{PP}}\right) + \frac{fn}{\text{PN}}\left(1 - \frac{fn}{\text{PN}}\right)\right) + \frac{tn}{\text{AN}} \cdot \left(\frac{fp}{\text{PP}}\left(1 - \frac{fp}{\text{PP}}\right) + \frac{tn}{\text{PN}}\left(1 - \frac{tn}{\text{PN}}\right)\right)\right]}{2} \tag{6}$$

3 Monte Carlo simulation

To compare the biases of the coefficients under investigation, a Monte Carlo simulation study has been conducted in order to generate R confusion matrices of dimension 2×2. Each confusion matrix contains the cross-classifications of $n = 100$ cases into '+' and '−' classes under the assumption of a prevalence rate (Pr) of class '+'. Specifically, at the beginning of each experiment, classifier sensitivity $\alpha = 0.9$, classifier specificity $\beta = 0.6$ and classifier propensity of random classifications $\theta = 0.05$ have been set. Three different scenarios have been investigated, corresponding to balanced (i.e. Pr = 0.5), moderately imbalanced (i.e. Pr = 0.7) and highly imbalanced (i.e. Pr = 0.9) data sets. In order to simulate the data sets, the n cases are firstly randomly assigned to the actual class in such a way that the probability of belonging to class '+' is Pr; the cases with actual class '+' are then randomly assigned to the predicted '+' class with probability α, whereas the cases with actual class '−' are randomly assigned to the predicted class '+' with probability $1 - \beta$. A proportion of θ cases are then randomly classified into one of the two classes with the same probability 1/2 assuming that a non-random rating leads to a correct classification.
For each scenario, $R = 1000$ Monte Carlo data sets have been generated and cross-classified in 2×2 confusion matrices. For each Monte Carlo data set, Accuracy, K, AC_1, Balanced Accuracy and Balanced AC_1 have been assessed and the relative bias of each coefficient i ($i = 1, \ldots, 5$) assessed as follows:

$$\text{RelBias}_i = \frac{\frac{1}{R}\sum_{r=1}^{R} \hat{p}_{r|i} - \text{P}^*}{\text{P}^*} \tag{7}$$

being $\hat{p}_{r|i}$ the classifier performance estimated with coefficient i for r^{th} data set and P^* the 'true' value of classifier performance given by:

$$P^* = \frac{(\alpha+\beta)/2-\theta/2}{1-\theta/2} \tag{8}$$

The adopted α, β and θ parameters give a 'true' value of classifier performance equal to 0.7436.

Table 2 Mean value and relative bias of performance measures based on $R = 1000$ Monte Carlo replications

		Accuracy	K	AC_1	Balanced Accuracy	Balanced AC_1
Pr = 0.5	Mean	0.7382	0.4735	0.4903	0.7377	0.6969
	RelBias	- 0.73	- 36.31	- 34.07	- 0.79	- 6.28
Pr = 0.7	Mean	0.7960	0.4908	0.6579	0.7385	0.7022
	RelBias	7.05	- 33.99	- 11.53	- 0.68	- 5.56
Pr = 0.9	Mean	0.8529	0.3587	0.8075	0.7377	0.7146
	RelBias	14.73	- 51.76	8.60	- 0.79	- 3.91

Simulation results, reported in Table 2, reveal that the mean value of classifier accuracy changes with class imbalance for imbalanced measures: the accuracy estimates vary from 0.4735 to 0.7382 with balanced data set (i.e. Pr = 0.5) and from 0.3587 to 0.8529 with highly imbalanced data set (i.e. Pr = 0.9).
The lowest accuracy estimate is always given by Cohen's K whereas the highest estimate is provided by Accuracy, since this latter does not compensate the non-zero probability of chance classifications. Cohen's K shows low sensitivity to moderate changes of class prevalence from 0.5 to 0.7 but K value decreases in the presence of highly imbalanced data set; indeed, the probability of classifications matching by chance is estimated through marginal frequencies and it is not clear how the classifier performance over majority and minority classes is balanced. Due to the strong dependency on the performance over the majority class, Accuracy increases with class prevalence and the same behavior is observed for AC_1, although its chance agreement has been formulated with the aim of compensating class imbalance.
Vice-versa, the balanced measures (i.e. Balanced Accuracy and Balanced AC_1) are not sensitive to class prevalence, as confirmed by the relative bias always at the most equal to 6%.

4 Conclusions

The blindly adoption of performance measures could result in dangerously misleading conclusions if they are not suited for the problem at hand. Although κ-type coefficients are able to account for chance agreement, they do not treat classes equally being thus not recommended for imbalanced data sets; on the other hand, Balanced Accuracy treats classes equally but it does not account that some classifications could match only by chance. The Balanced AC_1 coefficient is a novel κ-type coefficient able to treat both classes equally while accounting for predicted classifications matching the actual class by chance alone.
This research study aims at investigating, via a Monte Carlo simulation, the statistical behavior of Balanced AC_1 under different class imbalance conditions. Simulation results reveal the coefficient stability with class imbalance and its similarity with true classifier performance, being thus recommended as a good classifier performance measure able to deal with imbalanced data sets, which are a rule in many real-word classification problems.

References

1. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
2. Delgado, R., Tibau, X.A.: Why cohen's kappa should be avoided as performance measure in classification. PloS one **14**(9), e0222,916 (2019)
3. Gwet, K.: Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods for Inter-rater Reliability Assessment **1**(6), 1–6 (2002)
4. Labatut, V., Cherifi, H.: Evaluation of performance measures for classifiers comparison. arXiv preprint arXiv:1112.4133 (2011)
5. Sammut, C., Webb, G.I.: Encyclopedia of machine learning. Springer Science & Business Media (2011)
6. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information processing & management **45**(4), 427–437 (2009)
7. Vanacore, A., Pellegrino, M.S., Ciardiello, A.: Testing the predictive performance of multiclass classifiers. In: Book of short papers of IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment, pp. 141–147. PKE Publisher (2022)

Iterative Threshold-based Naïve Bayes Classifier: an efficient Tb-NB improvement

Il classificatore Iterative Threshold-based Naïve Bayes: un efficace miglioramento di Tb-NB

Maurizio Romano, Gianpaolo Zammarchi, and Giulia Contu

Abstract While analyzing online reviews on Booking.com, we proposed an ad-hoc classification model (Threshold-based Naïve Bayes Classifier, Tb-NB) to evaluate Customer Satisfaction, starting from the reviews' content, and predicting them as positive/negative. The log-likelihood ratios attributed to each word included in a review are then used to estimate a numeric sentiment score. In this paper we propose an improved version of Tb-NB called "iterative" Tb-NB. It results in a second step of Tb-NB: starting from the output of Tb-NB and reclassifying reviews with a probabilistic approach, it refines iteratively the threshold value used to classify a given subset of reviews.

Abstract *Analizzando le recensioni online presenti su Booking.com, è stato proposto un classificatore ad hoc (Threshold-based Naïve Bayes Classifier, Tb-NB) per valutare la Customer Satisfaction che, partendo dal contenuto delle recensioni stesse, le classifica in positive/negative. La log-verosimiglianza attribuita ad ogni parola contenuta in una recensione è successivamente utilizzata per stimare un sentiment score numerico. In questo paper si propone una versione migliorata chiamata "iterative". La nuova versione si comporta come un secondo step: a partire dall'output di Tb-NB si riclassificano le recensioni con un approccio probabilistico, raffinando la regola-soglia di classificazione per un sottoinsieme di recensioni iterazione dopo iterazione.*

Key words: Threshold-based Naïve Bayes Classifier, Iterative Threshold-based Naïve Bayes Classifier, Customer Satisfaction, Sentiment Analysis, General Sentiment Decomposition

---

Maurizio Romano
University of Cagliari, Cagliari, Italy, e-mail: romano.maurizio@unica.it

Gianpaolo Zammarchi
University of Cagliari, Cagliari, Italy, e-mail: gp.zammarchi@unica.it

Giulia Contu
University of Cagliari, Cagliari, Italy, e-mail: giulia.contu@unica.it

1 Introduction

Analyzing textual data has proved to be extremely useful in many field [7, 1, 4] but, likewise the language is complex so is the data [3]. In [2, 6] we have proposed Threshold-based Naïve Bayes Classifier (Tb-NB), a framework able to produce a good sentiment classifier with a particular focus on the model interpretability. In this paper we introduce an "iterative" step able to improve the performance of Tb-NB while reclassifying subsets of observations. Furthermore, we compare the performance of the new proposal with the previous one using the same Booking.com reviews data.

2 The data

For this study, with an ad hoc web scraping Python program, we have collected two datasets from Booking.com with a total of 127 features:

- Hotels dataset (86 features): Hotel (3), Review (8), Reviewer (2), Booking's score (11), Accommodation (32), Guest (8), Length of stay (6), Other info (4), Score components (12);
- Comments dataset (41 features): Hotel (2), Comment (6), Reviewer (2), Accommodation (16), Guest (4), Length of stay (3), Other info (2), Score components (6).

More in detail, the web-scraped data, is related to:

- 619 hotels located in Sardinia;
- 66,237 reviews, divided in 106,800 comments (in Italian or English): 44,509 negative + 62,291 positive;
- Period: Jan 3, 2015 – May 27, 2018.

3 Iterative Threshold-based Naïve Bayes Classifier

Considering a Natural Language text corpora as a set of reviews $\boldsymbol{r}$ s.t.:

$$r_i = comment_{pos_i} \cup comment_{neg_i}$$

where $comment_{pos}$ ($comment_{neg}$) are set of words (a.k.a. comments) composed by only positive (negative) sentences, and one of them can be equal to $\emptyset$, the basic features of the Threshold-based Naïve Bayes classifier applied to reviews' content are as follows. For a specific review r and for each word w ($w \in$ *Bag-of-Words*), we consider the log-odds ratio of w,

$$\begin{aligned} LOR(w) &= \log\left[\frac{P(c_{neg}|w)}{P(c_{pos}|w)}\right] \approx \\ &\approx \log\left[\frac{P(w|c_{neg})}{P(\bar{w}|c_{neg})}\cdot\frac{P(w|c_{pos})}{P(\bar{w}|c_{pos})}\cdot\frac{P(c_{neg})}{P(c_{pos})}\right] = \ldots = \\ &\approx pres_w + abs_w \end{aligned}$$

where $c_{pos}(c_{neg})$ are the proportions of observed positive (negative) comments whilst $pres_w$ and abs_w are the log-likelihood ratios of the events $(w \in r)$ and $(w \notin r)$, respectively.

While computing those values for all the w ($w \in$ *Bag-of-Words*) words, it is possible to obtain an output such that reported in Table 1, where we report c_{pos}, c_{neg}, $pres_w$ and abs_w for each word in the considered *Bag-of-Words*.

Table 1 Threshold-based Naïve Bayes output

	w_1	w_2	w_3	w_4	w_5	...
$P(w_i\|c_{neg})$	0.011	0.026	0.002	0.003	0.003	...
$P(w_i\|c_{pos})$	0.007	0.075	0.005	0.012	0.001	...
$pres_{w_i}$	0.411	-1.077	-1.006	-1.272	1.423	...
abs_{w_i}	-0.004	0.052	0.003	0.008	-0.002	...

We have then used cross-validation to estimate a parameter τ such that: c is classified as "negative" if $LOR(c) > \tau$ or as "positive" if $LOR(c) \leq \tau$.

We knew now that the Threshold-based Naïve Bayes Classifier has engaging performances with respect to other algorithms [6, 2]. However, still, there are some problems during the classification in Pos/Neg of text data that have a LOR-value close to the selected value of τ. In particular, we can observe this in Fig. 1: if the LOR-value is far from the estimated threshold, there are few chances to obtain an incorrect classification.

Considering that the LOR-value is a numeric sentiment value, it is obvious than if this value is close to the threshold even a word with an irrelevant sentiment value could be classified incorrectly.

To improve the performance of the Threshold-based Naïve Bayes Classifier for LOR-values located close to the threshold, we add a new step that refines the threshold while classifying only those comments that are close to the selected value of τ. Then, we classify again those comments with a new estimated value of τ.

The "augmented" Iterative Threshold-based Naïve Bayes works as follows:

1. a proportion of observations close to τ, and located in the uncertainness area, are marked for being reclassified.
2. A new decision rule is created for those observations only, estimating a new value of τ.
3. Observations in the uncertainness area are re-classified and Steps 1-3 are repeated until the proportion of cases in the uncertainness area reduces consistently.

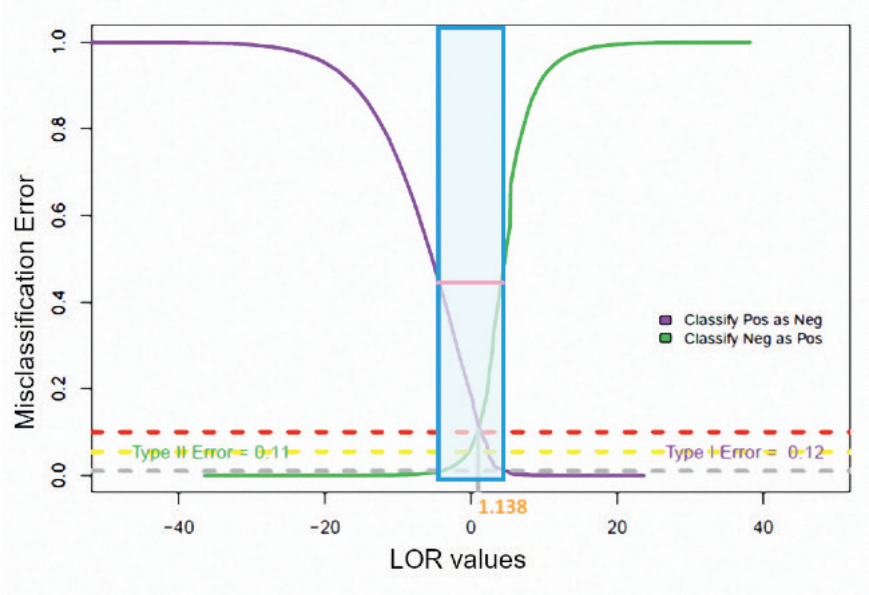


Fig. 1 Threshold-based Naïve Bayes uncertainness area

We can then transform that logic in an algorithm that better describes such a process:

---

1. Set $i = 0$
2. Map the distribution of LOR-values into two functions:

$$x = LOR(\pi_c)$$

$$f_{pos} : x \mapsto [0;1], f_{neg} : x \mapsto [0;1] \qquad \forall x \in (-\infty;+\infty)$$

   s.t. the input (x) is the LOR-value of a comment c and the output of f_{pos} (f_{neg}) is the probability that c is positive (negative).
3. Define max_{pos}, max_{neg} s.t.:

$$max_{pos} = \underset{x\in(-\infty;+\infty)}{\text{argmax}} \; f_{pos}(x)$$

$$max_{neg} = \underset{x\in(-\infty;+\infty)}{\text{argmax}} \; f_{neg}(x)$$

4. Compute the intersection point τ_i s.t.

$$\tau_i = \begin{cases} \underset{x \in (max_{pos}; max_{neg})}{\text{argmin}} \; (f_{neg}(x) - f_{pos}(x)) & \text{if } max_{pos} < max_{neg} \\ \underset{x \in (max_{neg}; max_{pos})}{\text{argmin}} \; (f_{neg}(x) - f_{pos}(x)) & \text{if } max_{pos} > max_{neg} \end{cases}$$

5. if $max_{pos} < max_{neg}$ update the decision rule:
   $LOR(\pi_c) > \tau_i \rightarrow$ c is classified as "Negative"
   $LOR(\pi_c) \leq \tau_i \rightarrow$ c is classified as "Positive"
6. otherwise, if $max_{pos} > max_{neg}$ update the decision rule:
   $LOR(\pi_c) > \tau_i \rightarrow$ c is classified as "Positive"
   $LOR(\pi_c) \leq \tau_i \rightarrow$ c is classified as "Negative"
7. Set $i = i + 1$
8. Repeat steps 1-7 with a proportion ω of comments located in the interval $\tau_i \pm \partial\omega$ until a stopping criterion is satisfied.

The proportion of observations marked for being re-classified is estimated while considering the uncertainness area around τ. Using a large value of ω leads to results that are similar to those obtained from Threshold-based Naïve Bayes classifier, whilst using a low value for ω might lead to consider few observations only, thus producing a decision rule which overfits the data. Our experiments suggest that, empirically, a good trade-off is reached by setting $\omega = 0.20$.

Whereas, the **stopping criterion** of the iterative Tb-NB depends on at least one of the following conditions:

1. the number of cases (positive and negative comments) used to produce a reliable distribution of LOR-values, and the associated function that map them, is too small;
2. there is no intersection point (other than zero) between $f_{pos}(x)$, *and* $f_{neg}(x)$ and $max_{pos} \neq max_{neg}$. The two distributions are well separated, thus the τ estimated by the standard Threshold-based Naïve Bayes Classifier at the previous step is the final value of τ;
3. $max_{pos} = max_{neg}$. We cannot distinguish the positive distribution from the negative one. Thus, there is not a geometric solution to this problem and the classification of those comments has to be done with the τ estimated by the Threshold-based Naïve Bayes Classifier.

Moreover, it is interesting to highlight how empirical results suggest that the stopping criterion usually is reached after no more than two iterations ($0 \leq i \leq 2$).

Furthermore, in Table 2 we have compared the performance of the proposed classifier (iTb-NB) with the previous version (Tb-NB) and that of other well known competitors, in particular: Logistic Regression (LOG), Random Forest (RF), standard Naïve Bayes (NB E1071 [5]), Naïve Bayes using kernel estimated densities (NB KlaR [8]), Decision Trees (CART), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

It is possible to notice that iTb-NB is the most accurate classifier with respect to four out of the five measures reported in the table.

Table 2 Performance metrics on raw data using 5-fold cross validation

Classifier	ACC	Sensitivity	Fall-out	F1	MCC
Tb-NB	0.911	**0.929**	0.117	0.926	0.813
iTb-NB	**0.925**	0.923	**0.072**	**0.939**	**0.829**
LOG	0.850	0.884	0.532	0.877	0.361
RF	0.811	0.873	0.591	0.849	0.303
NB(E1071)	0.806	0.804	0.389	0.834	0.390
NB(KLAR)	0.806	0.804	0.389	0.834	0.390
CART	0.768	0.842	0.587	0.815	0.272
LDA	0.764	0.860	0.641	0.816	0.246
SVM	0.793	**0.929**	0.290	0.771	0.621
average	0.826	0.872	0.401	0.851	0.469

Notes: ACC = Accuracy; F1 = F1-score; MCC = Matthews Correlation Coefficient

4 Conclusions

We have proposed an improved, and "iterative", Threshold-based Naïve Bayes Classifier. Furthermore, we have compared the performance of the new proposal with the previous one using the same Booking.com reviews data. Reclassifying subsets of observations has then been proved as a way to reduce the misclassification rate and hence boosting the performance of Tb-NB whilst keeping the original interpretability of the results.

References

1. Boyd, D., Crawford, K.: Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society **15**(5), 662–679 (2012)
2. Conversano, C., Romano, M., Mola, F.: Hotel search engine architecture based on online reviews' content. In: Arbia, G., Peluso, S., Pini, A., Rivellini, G. (eds.) Smart Statistics for Smart Applications. Book of Short Papers SIS2019, pp. 213-218. Pearson, Milan (2019)
3. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Synthesis Lectures on Human Language Technologies **10**(1), 1–309 (2017)
4. Halevy, A., Norvig, P., Pereira, F.: The Unreasonable Effectiveness of Data. IEEE Intelligent Systems **24**(2), 8–12 (2009)
5. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: Misc. Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2019).
6. Romano, M., Frigau, L., Contu, G., Mola, F., Conversano, C.: Customer Satisfaction from Booking. In: Mieli, M., Volpe, C. (eds.) Selected papers Conferenza GARR_18 Data (R)evolution, pp. 111-118. Associazione Consortium GARR, Cagliari (2018)
7. Schmunk, S., Höpken, W., Fuchs, M., Lexhagen, M.: Mobile Social Travel Recommender System. In: Xiang, Z., Tussyadiah, I. (eds.) Information and Communication Technologies in Tourism 2014, pp. 3-16. Springer International Publishing, Cham (2013)
8. Weihs, C., Ligges, U., Luebke, K., Raabe, N.: klaR Analyzing German Business Cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (eds.) Data Analysis and Decision Support, pp. 335-343. Springer-Verlag, Berlin/Heidelberg (2005)

Reprogramming FairGANs with Variational Auto-Encoders: A New Transfer Learning Model

Auto-Encoders variazionali per la riprogrammazione di FairGANs: un approccio basato su Transfer Learning

Beatrice Nobile, Gabriele Santin, Bruno Lepri and Pierpaolo Brutti

Abstract Fairness-aware GANs (FairGANs) exploit the mechanisms of Generative Adversarial Networks (GANs) to impose fairness on the generated data, freeing them from both disparate impact and disparate treatment. Given the model's advantages and performance, we introduce a novel learning framework to transfer a pre-trained FairGAN to other tasks. This reprogramming process has the goal of maintaining the FairGAN's main targets of data utility, classification utility, and data fairness, while widening its applicability and ease of use. In this paper we present the technical extensions required to adapt the original architecture to this new framework (and in particular the use of Variational Auto-Encoders), and discuss the benefits, trade-offs, and limitations of the new model.

Abstract *Le FairGANs rappresentano una classe di modelli di recente introduzione che si sono dimostrati particolarmente efficaci nell'indurre imparzialità nei dati generati sfruttando i meccanismi propri delle Generative Adversarial Networks (GANs). Sulla base di queste premesse, nel presente lavoro introduciamo un nuovo schema di transfer learning per estendere l'applicabilità di una FairGAN pre-addestrata ad altri data-set e task, anche profondamente diversi da quelli visti in training. Questo processo di riprogrammazione ha lo scopo di mantenere i principali obiettivi di FairGAN ampliandone al tempo stesso il campo e la facilità d'uso. Qui presentiamo le estensioni tecniche necessarie per adattare l'architettura originale al nuovo framework (in particolare l'uso di Variational Auto-Encoder) e discutiamo i vantaggi, i compromessi e le limitazioni del nuovo modello.*

Key words: FairGAN, Reprogramming, Variational Auto Encoder, Fairness

Beatrice Nobile · Pierpaolo Brutti
Sapienza University of Rome, e-mail: nobile.1908315@studenti.uniroma1.it, pierpaolo.brutti@uniroma1.it

Gabriele Santin · Bruno Lepri
DIGIS, Bruno Kessler Foundation, e-mail: gsantin@fbk.eu, lepri@fbk.eu

1 Introduction

The research on Fairness in automated systems has recently advanced considerably, even if recent developments have made the field increasingly complex. The paper [6] lists as many as 23 possible biases in the data, and 6 different types of discrimination that may arise from it, but most importantly they also list 10 different definitions of fairness with no method that can address them all at the same time [8].

Within this complex landscape, FairGANs [10] offer a simple and actionable mechanism to address biases in the data. Rather than debiasing the data at hand, FairGANs use a GAN-like mechanism to generate new data that must be as similar as possible to the original, but with the additional feature that no evident correlation with the protected attribute is present. As such FairGANs can ensure that the system would be free of both *disparate treatment* and *disparate impact*, i.e., implicit and explicit discrimination in its decision making.

Since the need for retraining at each usage may be cumbersome at best and unfeasible in many cases where there is not enough data or computing power, this paper aims at creating a transfer learning model for FairGANs. This could lead to a wider adoption of such a method given an increased ease of usage, minimized resource demands, and the ability to re-use knowledge from high-quality data, potentially without loss of accuracy.

To achieve these goals, we make use of Adversarial Reprogramming techniques [3, 5]), which can retrain the model without the need for the original dataset, and only modifiying a reduced part of the model. The biggest issue in applying transfer learning for FairGANs with Adversarial Reprogramming is that the two architectures have been designed to work on different types of data. In fact, while FairGANs have to do with tabular data, Adversarial Reprogramming so far has only been applied to images and sequences [5, 7]. Therefore, the first contribution of this paper is designing a model that is able to perform Reprogramming on tabular data. To do so, we have used Variational Auto-Encoders [2] which allow for the necessary dimensional flexibility. The second contribution is understanding whether transfer learning is possible on FairGANs, or if instead direct training is absolutely necessary at each new task. Our findings suggest that balancing all three objectives of FairGANs, data utility and fairness and classification utility, is not trivial when performing transfer learning. Yet, even extreme cases have yielded positive results in most experiments, suggesting that it is indeed possible.

2 Materials and Methods

We recall the tools used in our model in Section 2.1, and introduce the complete solution in Section 2.2.

2.1 Background on GANs, FairGANs, and reprogramming

GANs were first introduced in 2014 [4]. Given an initial dataset $\mathscr{D} := \{x_i\}$ of certain objects, the idea is to train two neural networks, a generator G and a discriminator D, so that G is able to generate, from a random input seed $\bar{z}$, an object $\bar{x}$ that is as similar as possible to the true ones, while D is able to classify real and fake objects. The two networks are trained to minimize a minimax loss [4].

FairGANs [10] follow the same mechanism but applied now to an extended dataset $\mathscr{D} := \{(x_i, y_i, s_i)\}$, where x_i is a certain object, y_i is an associated binary label, and s_i is a binary value representing the membership of x_i in a protected class. In this case G is seeded with a random input $\bar{z}$ and a value $\bar{s} \in \{0,1\}$, and outputs a generated pair $(\bar{x}, \bar{y})$. A first discriminator D_1 tries to tell apart the triples $(\bar{x}, \bar{y}, \bar{s})$ from those found in $\mathscr{D}$, while a second discriminator D_2 tries to classify the pairs $(\bar{x}, \bar{y})$ generated with an input $\bar{s} = 0$ from those generated with an input $\bar{s} = 1$, of course without access to the actual value of $\bar{s}$. A modification of the minimax optimization problem forces G to fool both discriminators, thus learning to generate data that are realistic (*data utility*), and for which the protected attribute is hard to identify if not disclosed (*data fairness*). Once the training is completed, G can be used to generate a bias-free dataset, that can finally be used to train a classifier that predicts the binary label y for a previously unseen object x, without relying on s.

In order to perform transfer learning with a pre-trained FairGAN, we additionally use Adversarial Reprogramming techniques [3]. In general terms, given a model $f : X \to Y$ between sets X, Y and a target model $g : X' \to Y'$ between sets X', Y', the goal is to modify the input and output of f in order to mimic the response of g. A typical example is the modification of a pre-trained classifier that maps a set of images X to suitable labels Y, in order to obtain a new classifier g that works on a different set of images X' of possible different resolution, with other labels Y'. This goal is achieved by training two parametric functions $h_f : X' \to X$, $h_g : Y \to Y'$ so that the input-output modified model $h_g \circ f \circ h_f : X' \to Y'$ approximates g. This general technique has been applied to GANs in [5], where an existing generator-discriminator pair (G, D) is updated to generate objects of the same original dimension, but following a different distribution. An input modification function h_f is applied to obtain a new generator $G' := G \circ h_f : Z' \to X$, where Z' is a second set of input noise seeds. Since the GAN training step requires the join optimization of the generator and discriminator, the new discriminator D' is instead trained from scratch, therefore moving away from an adversarial attack paradigm towards one of transfer learning.

2.2 Reprogramming FairGANs

We assume to have a dataset $\mathscr{D} := \{(x, y, s)\} \subset X \times Y \times \{0,1\}$, where X are tabular data, and initially train a classifer $\overline{C} : X \to Y$, that will remain fixed. Additionally, we train a standard VAE $\overline{\text{Enc}}, \overline{\text{Dec}}$ on $X \times S$, i.e., $\overline{\text{Dec}}(\overline{\text{Enc}}((x,s))) := (G_x(x,s), G_s(x,s)) \approx (x,s)$ for all $(x,s) \in X \times S$. This will serve as the base for the

construction of the generator and allow us to tackle reprogramming with dimension expansion, which is a notoriously hard to solve problem [5].

We then consider a second target dataset $\mathscr{D}' := \{(x', y', s')\} \subset X' \times Y \times \{0,1\}$ with tabular input X'. We assume that the columns of X' are either a subset or a superset of those of X. Observe that a special setting of interest, which we will explore in the following, is the case when $X = X'$ or $Y = Y'$. Our goal is to adapt the classifier $\overline{C}$ to $\mathscr{D}'$ under two constraints: creating realistic data, and creating bias-free data.

To reprogram with the realism constraint only, we define a generator $G : (X', S') \to (X, S)$ defined as $G := \overline{\text{Dec}} \circ \text{Enc}$ which maps the new columns to the old ones. We also define a discriminator $D : X \to \{0,1\}$. Observe that the decoder part of the generator is kept fixed. Instead, we train Enc with two goals: we want to have $\overline{C}(G_x(x')) \approx y'$, i.e., the classification accuracy is preserved, and we want that D_1 can't distinguish $G(X' \times S')$ from $X' \times S'$. Observe that D_1 acts only on the columns of $G(X')$ that are in common with those of X', so that it can be trained by having access to X' only, and not to the original set X. That is, the discriminator should discriminate only those columns that represent the current data. This goal is realized by the optimization of Enc and D_1 by minimization of a loss composed by the usual GAN loss with ℓ_2 norm penalty, plus the cross entropy loss for classification. The balance between the two losses is controlled by a parameter $\gamma \geq 0$ (large γ means large importance to classification). In this way we realize a transfer of the classifier from $\mathscr{D}$ to $\mathscr{D}'$, in a way that preserves a realism condition on the generated data. Observe that even without the realism constraint, this *tabular reprogramming* is an unexplored field for adversarial reprogramming.

When adding a fairness constraint, we obtain the full *FairGAN reprogramming* model, which differs solely for the presence of additional discriminators. In this case, a second discriminator $D_2 : G(X' \times S') \to S'$ tries to distinguish between the generated columns that are generated with an original input with $s' = 0$ or $s' = 1$. This second discriminator is trained together with Enc and D_1, by adding to the loss a second GAN-discriminator loss term controlled by a second parameter $\delta \geq 0$, where $\delta = 0$ means that D_2 is deactivated. With this design, the reprogrammed FairGAN guarantees data fairness because the "fairness" discriminator D_2 forces the generator to produce data that is independent of the sensitive information.

3 Results

For our experiments we use two datasets, COMPAS Recidivism Racial Bias [1] comprising after cleaning 60843 observations and 15 columns, and a Candidate Selection dataset [9] comprising 31349 observations and 10 columns. We consider as target labels y either *gender* (both datasets), or *ethnic code* (Caucasian or not, in COMPAS) and *selected* (hired or not - Candidate).

Table 1 reports the accuracies of these four classifiers $\overline{C}$, that we consider as baselines for our experiments. For the reprogramming tasks we consider four scenarios: i) We don't change neither dataset nor task (Same Dataset Same Task, SDST); ii) We change the task but not the dataset (Same Dataset Different Task, SDDT); iii)

We change dataset while maintaining the same task (Different Dataset Same Task, DDST); iv) We change both dataset and task (Different Dataset Different Task, DDDT). We remark that passing from COMPAS to Candidate requires a dimensionality reduction, while the vice versa requries a dimensionality expansion.

In all experiments, we use the architecture of Section 2.2. The latent dimension is set to 10, which is the smallest value that does not affect the classification accuracy in the results of Section 3.1.

	COMPAS-Gender	COMPAS-Ethnic code	Candidate-Gender	Candidate-Selected
Accuracy	82.8	67.2	78.8	71.7

Table 1: Accuracy of the classifiers trained on different dataset and target labels.

3.1 Tabular Reprogramming

We first test simply tabular reprogramming, i.e., without fairness constraint. In other terms, the only objective we are trying to achieve is to efficiently fool the classifier in order to maintain high accuracy, regardless of how we generate the data. We consider the classifier trained on the Candidate dataset to predict the label *gender*, and report the results of the four transfer learning scenarios SDST, SDDT, DDST, DDDT in the left part of Table 2. The model works very well in three out of four cases, even outperforming the baseline. Observe that this is possible since the architecture includes now also the optimizable encoder, and thus the overall model has more degrees of freedom. Only in the DDDT setting, which is ideed the more challenging, we observe a limited reduction of 4% in the accuracy. Thus, we can affirm that we have successfully reprogrammed across tabular datasets.

	SDST	SDDT	DDST	DDDT	γ	SDST	SDDT	DDST	DDDT
Reprogramming	78.9	78.7	79.1	74.5	0.5	74.2 ✓	57.6 ✓	61.1 ✓	52.7 ✓
					100	76.3 ×	64.8 ×	69.1 ✓	64.8 ✓
Baseline	78.8	78.8	78.8	78.8		82.8	67.2	78.8	71.7

Table 2: Accuracy results of tabular reprogramming (left, see Section 3.1) and GAN reprogramming (right, see Section 3.2), compared to the corresponding baseline models obtained by direct training.

3.2 GAN and FairGAN Reprogramming

We now add the data realism constraint, so that we both reprogram the classifier and require the generated data to be similar to the input one. As described in Section 2.2, a parameter γ balances the importance of classification and of data realism, where larger values of γ increase the relative importance of classification. The setting of γ is itself a challenging task. Cross validation over the values of γ in combination

with the learning rate η have identified a values $\gamma = 0.5$, $\eta = 10^{-3}$. However, preliminary investigations have shown that reasonable results may be achieved with $\gamma \in [10^{-1}, 10^{2}]$. We report the results in Table 2, where we also report the case of $\gamma = 0.5$ for comparison. Each experiment reports also a green check or a red cross to indicate if the data realism constraint is met. In these preliminary experiments, this is checked by comparing the histograms of the distributions of the single true and generated columns. Unsurprisingly, moving across datasets seems to be more challenging for our model, but we manage to keep a relatively small decay in the accuracy, especially considering the semantic distance between the two datasets and the respective tasks.

Adding the fairness constraint we can finally test the entire FairGAN model, where we tested the model on several randomly selected columns used as the protected attribute. In this case the loss comprises the three terms of classification accuracy, realism, and fairness, and their relative importance is controlled now by the two hyperparameters δ and γ. To measure the success of the fairness constraint, we check if the validation accuracy of the discriminator is close to 50% on a validation set, meaning that the protected attribute is not predictable from the cleaned data. In this case, no value of the two parameter give a sufficiently accurate result in SDST and SDDT, while for DDST and DDDT we managed to obtain a validation accuracy close to 50%. This apparently counter-intuitive result may be due to preliminary nature of our experiments, and will be addressed in future research.

4 Conclusions

In this work we introduced for the first time a framework for the reprogramming FairGANs on tabular data, i.e., the possibility of using a pre-trained FairGAN and apply it to other datasets and/or tasks, maintaining acceptable levels of *data utility*, *classification utility* and *data fairness*. Our novel architecture leverage VAEs as a basic building block, and combines it with two dicriminators which are trained with a GAN loss to forse the generation of data that meets the desired constraints. Preliminary experiments on two benchmark datasets have shown that the model can be effectively trained to reprogram FairGANs on tabular dataset. Nevertheless, the incremental addition of multiple constraints makes the training increasingly challenging. Even if the experiments lean towards positive results, some issues remain unanswered, especially in the design of an effective balancing mechanism between the various components of the model.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica **23**, 77–91 (2016)
2. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)

3. Elsayed, G.F., Goodfellow, I., Sohl-Dickstein, J.: Adversarial reprogramming of neural networks. arXiv preprint arXiv:1806.11146 (2018)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
5. Lee, K., Suh, C., Ramchandran, K.: Reprogramming GANs via input noise design. In: ECML PKDD 2020: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML-PKDD (2020)
6. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv e-prints pp. arXiv–1908 (2019)
7. Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F.: Adversarial reprogramming of text classification neural networks. arXiv preprint arXiv:1809.01829 (2018)
8. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. arXiv preprint arXiv:1709.02012 (2017)
9. Tarun Chilkur: Candidate selection. predicting the selection of a candidate (2020). Data retrieved from https://www.kaggle.com/tarunchilkur/client
10. Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 570–575. IEEE (2018)

Robust statistics

Combinatorial Analysis of Factorial Designs with Ordered Factors

Analisi Combinatoria di Piani Fattoriali con Fattori Ordinali

Roberto Fontana and Fabio Rapallo

Abstract In recent literature a new combinatorial algorithm for the selection of robust fractional factorial designs has been introduced. In this work we analyze the application of this algorithm in the case of ordered factors.
Abstract *È stato sviluppato recentemente un nuovo algoritmo combinatorio per la selezione di piani fattoriali frazionari robusti. In questo lavoro analizziamo la sua applicazione nel caso di fattori ordinali.*

Key words: Algebraic statistics, Design of experiments, Optimality, Robust fractions

1 D-optimality, robustness, and combinatorial objects

The choice of a design from a set of candidate runs is one of the most relevant problems in Design of Experiments. When working in the framework of factorial designs, the candidate set is usally the full-factorial design containing all the possible level combinations of the factors. There are several criteria for choosing a design. Here we restrict our attention to model based techniques. Thus the linear model on the candidate set $\mathscr{D}$ is written in the form

$$\mathbf{y} = X_{\mathscr{D}}\beta + \varepsilon\,, \tag{1}$$

Roberto Fontana
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, e-mail: roberto.fontana@polito.it

Fabio Rapallo
Dipartimento di Economia, Università di Genova, via Vivaldi 5, 16126 Genova, e-mail: fabio.rapallo@unige.it

where $X_{\mathscr{D}}$ is the full-design model matrix with dimensions $K \times p$, β is the p-dimensional vector of the parameters, ε is the error, and $\mathbb{E}(\mathbf{y}) = X_{\mathscr{D}}\beta$.

The classical theory leading to the class of alphabetical optimality criteria (D-optimality, A-optimality, etc.) is based on the maximization of some quantities computed using the model matrix $X_{\mathscr{F}}$ of the selected fraction $\mathscr{F} \subset \mathscr{D}$. In their basic form, the selection algorithms work with a pre-defined and fixed dimension $n = \#\mathscr{F}$ of the fraction. As general reference for optimal designs, refer to [6].

When the design may be incomplete, e.g. for time limitations, there are methods to choose the order of the runs in order to achieve first the most informative runs, so that a possibly incomplete design is as much effective as possible for parameter estimation. Fractional Factorial Designs with removed runs are studied in, e.g., [1], [7]. In such a case the set $\mathscr{D}$ is usually a candidate set different from the full-factorial design.

Both optimality with a fixed run size and with possibly incomplete designs has been recently analyzed under a geometric and combinatorial point of view using a special representation of the basis of the kernel $\ker(X_{\mathscr{D}}^t)$ of the model matrix for the candidate set, namely the circuit basis. In particular, the property that naturally reflects the geometry of the design points is the robustness, first introduced in [4].

Definition 1. The robustness of a fraction $\mathscr{F}$ with design matrix $X_{\mathscr{F}}$ is defined as

$$r(X_{\mathscr{F}}) = \frac{\#\left\{ \text{ saturated } \mathscr{F}_p \right\}}{\#\left\{\mathscr{F}_p\right\}} = \frac{\#\left\{ \text{ saturated } \mathscr{F}_p \right\}}{\binom{n}{p}}$$

where $\mathscr{F}_p$ denotes a fraction with p runs and $\#\{\cdot\}$ denotes the cardinality of the set $\{\cdot\}$. $\mathscr{F}$ is a saturated fraction if $\#\mathscr{F} = p$ and the parameters β are estimable.

On the other hand, the circuit basis of the matrix $A_{\mathscr{D}} = X_{\mathscr{D}}^t$ is defined as follows.

Definition 2. 1. A vector $\mathbf{u} = (u(1), \ldots, u(K))$ in $\ker(A_{\mathscr{D}})$ is a circuit if it has relatively prime entries and minimal support. The support of a vector is the set of indices for which the entries are non-zero.
2. The (finite) set of all the circuits is the circuit basis of $\ker(A_{\mathscr{D}})$, and it is denoted by $\mathscr{C}(A_{\mathscr{D}})$.

For a concise reference on the circuits, their combinatorial properties, and their applications to optimization problems, the reader can refer to [8]. The circuit basis of an integer matrix $A_{\mathscr{D}}$ can be computed through several packages for symbolic computation. The computations presented in the present paper are carried out with `4ti2`, see [9].

In [2] and [3] it is shown that robust fractions correspond to the fractions which minimize the intersections between the fraction and the support of the circuits. For space limits, we do not introduce here the full details of the theory, but we summarize the algorithm for finding nested robust fractions using the circuit basis.

1. Start with an arbitrary fraction $\mathscr{F}$ of a specified size n;
2. Repeat:

a. Consider the circuits of $\mathscr{C}(X_{\mathscr{D}})$ which are contained in $\mathscr{F}$;
b. For each run R in $\mathscr{F}$, compute the number of circuits in which R is contained. This is the loss function associated to R;
c. Remove from the fraction the run with the highest loss function. In case of ties, randomize.

2 Circuits in case of ordered factors

In the previous section we have not mentioned the problem of the choice of the coding of the factor levels. Indeed, the combinatorial analysis introduced above is usually applied in the framework of qualitative nominal factors. In such a case, the linear model in Eq. (1) can be written in the standard ANOVA form. For instance:

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{2}$$

with the constraints $\sum_i \alpha_i = 0$, $\sum_i \beta_i = 0$, $\sum_i (\alpha\beta)_{ij} = 0$,$\sum_j (\alpha\beta)_{ij} = 0$.

Using the model written in the form of Eq. (2) with qualitative factors, there is a large class of codings which are equivalent in terms of the kernel of the matrix $A_{\mathscr{D}} = X_{\mathscr{D}}^t$. Among these parametrizations, one can use a polynomial model in the form $\mathbb{E}(Y) = \sum_{\alpha \in L} c_\alpha X^\alpha$, where c_α are real coefficients, X^α are monomials, and L is a suitable list of exponents. The proof of the equivalence is based on the Identity Theorem for Polynomials, see [5] for a detailed analysis and examples.

To encode ordered factors we use here two codings:

1. For a linear ordered factor (e.g., the discretization of a quantitative factor) with s levels, we use the set $\{0,\dots,s-1\}$ for a linear effect, and its powers for higher-order effects;
2. For a cyclic factor with s levels, we use the coding based on the roots of the unity:

$$\left\{ \omega_k = \frac{2\pi i k}{s} \ : \ k = 0,\dots,s-1 \right\} \tag{3}$$

With this choice, the monomials $X, X^2, X^3, \dots$ encode the cyclical nature of the factor. For computational reasons, the roots of the unity in Eq. (3) can be replaced with suitable Fourier-type functions, such as linear combinations of sin and cos functions.

Notice that there is a major difference between nominal and ordered factors. While for nominal factors all parametrizations are equivalent, when ordered factors are considered one can add to the model only a linear effect, or the linear effect plus some powers. This implies that when using the algorithm described in the previous section with ordered factors, special attention must be given on the circuit basis, taking the correct effects in the model matrix. In the next section we illustrate two examples involving ordered factors.

3 Examples

The first example considers two 2-level factors (X_1, X_2) and one 5-level factor (X_3) with two different models. The first model is linear in X_1, X_2, and X_3 and contains a constant term: $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. The number of degrees of freedom of this model is $p = 1+3 = 4$. For this model, the circuit basis is formed by 44 circuits with cardinality of the supports ranging from 2 to 5. The robustness of a D-optimal design with $n = 10$ runs is analyzed. The 10-run D-optimal design has been obtained using the full factorial design $\mathscr{D} = \{0,1\}^2 \times \{0,\ldots,4\}$ as candidate set. The exact distributions of the values of the robustness of the fractions which are obtained removing $k = 1,\ldots,n-p = 6$ points are computed and compared with the values of the robustness corresponding to the fractions found by the algorithm. Table 1 compares the values of the robustness of the fractions found by the algorithm (r_*) with the 75th, 90th and 95th percentile of the distributions of the robustness (p_{75}, p_{90}, p_{95} respectively) for different number of points (k) removed by the initial design. The value corresponding to the robustness of the initial design (r_0) is given at $k = 0$. It is worth noting that for each number k of points removed the algorithm provides values of robustness equal to the 95th percentile.

k	p_{75}	p_{90}	p_{95}	r_*
0		r_0=0.457		
1	0.476	0.476	0.476	0.476
2	0.529	0.529	0.529	0.529
3	0.629	0.629	0.629	0.629
4	0.6	0.6	0.867	0.867
5	0.8	1	1	1
6	1	1	1	1

Table 1: Example 1, $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Unlike the first model, the second one contains also the powers of X_3 of order 2,3,4: $\mathbb{E}(Y_{(x_1,x_2,x_3)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{k=1}^{4} \beta_{3k} x_3^k$. The number of degrees of freedom of this model is $p = 1+2+4 = 7$. From the combinatorial point of view this model is simpler than the linear one: there are only 4 circuits in the circuit basis. Also in this case the robustness of a D-optimal design with $n = 10$ runs and obtained using the full factorial $\mathscr{D}$ as candidate set is analyzed. The exact distributions of the values of the robustness of the fractions which are obtained removing $k = 1,\ldots,n-p = 3$ points are computed and compared with the values of the robustness corresponding to the fractions found by the algorithm. In this case we do not display all the results, but the performance of the algorithm is similar to the previous example.

The second example is taken from [11] and is based on [10]. An animal scientist wants to compare wildlife densities in four different habitats over a year. However, due to the cost of experimentation, only $n = 16$ observations can be made (in the original example the requested size was $n = 12$, but $n = 16$ allows us to describe the

method better than $n = 12$). The following model is postulated for the density $Y_j(t)$ in habitat j during the month m:

$$\mathbb{E}(Y_j(t)) = \mu_j + \gamma m + \sum_{k=1}^{4} \alpha_k \cos\left(k\frac{\pi m}{4}\right) + \sum_{k=1}^{3} \beta_k \sin\left(k\frac{\pi m}{4}\right) \tag{4}$$

The model includes the habitat as a classification variable ($\mu_j, j = 1,\ldots,4$), the effect of time with an overall linear drift term γm, $m = 1,\ldots,12$ ($m = 1$ corresponds to January, $\ldots$, $m = 12$ corresponds to December), and cyclic behavior in the form of a Fourier series. There is no intercept term in the model and the number of parameters is $p = 4+1+4+3 = 12$.

The Optex procedure [11] is used to generate a D-optimal design $\mathscr{F}$ with $n = 16$ runs using the full factorial arrangement of four habitats by 12 months (48 runs) as candidate set. The model matrix $X_{\mathscr{F}}$ corresponding to the 16-run D-optimal design that has been generated by the Optex procedure is reported in Table 2. The month $m \in \{1,\ldots,12\}$ is expressed as a number $t \in [-1,+1]$ using the linear transformation $t = -1 + (2/11)(m-1)$.

μ_1	μ_2	μ_3	μ_4	γ	α_1	α_2	α_3	α_4	β_1	β_2	β_3
0	0	0	1	-0.636	-0.707	0	0.707	-1	0.707	-1	0.707
0	0	0	1	-0.091	0	-1	0	1	-1	0	1
0	0	0	1	0.273	1	1	1	1	0	0	0
0	0	0	1	0.636	0	-1	0	1	1	0	-1
0	0	1	0	-0.818	0	-1	0	1	1	0	-1
0	0	1	0	-0.273	-0.707	0	0.707	-1	-0.707	1	-0.707
0	0	1	0	0.091	0.707	0	-0.707	-1	-0.707	-1	-0.707
0	0	1	0	0.818	-0.707	0	0.707	-1	0.707	-1	0.707
0	1	0	0	-0.455	-1	1	-1	1	0	0	0
0	1	0	0	-0.273	-0.707	0	0.707	-1	-0.707	1	-0.707
0	1	0	0	0.273	1	1	1	1	0	0	0
0	1	0	0	0.455	0.707	0	-0.707	-1	0.707	1	0.707
1	0	0	0	-1	0.707	0	-0.707	-1	0.707	1	0.707
1	0	0	0	-0.091	0	-1	0	1	-1	0	1
1	0	0	0	0.091	0.707	0	-0.707	-1	-0.707	-1	-0.707
1	0	0	0	1	-1	1	-1	1	0	0	0

Table 2: Model matrix $X_{\mathscr{F}}$ corresponding to the 16-run D-Optimal design

The circuits of a matrix A can be computed for an integer matrix A. Then we have to build an approximate version $\tilde{X}_{\mathscr{F}}$ of $X_{\mathscr{F}}$. Let us denote with x_{ij} and $\tilde{x}_{ij}$ the elements of the matrices $X_{\mathscr{F}}$ and $\tilde{X}_{\mathscr{F}}$ respectively, $i = 1,\ldots,16$, $j = 1,\ldots,12$. We do not modify the coding of the habitats, $\tilde{x}_{ij} = x_{ij}, j = 1,\ldots,4$, we code the month using again the integer m as above ($\tilde{x}_{ij} = m$) and we round the remaining values defining $\tilde{x}_{ij} = 10r(x_{ij})$ where $r(x)$ returns the rounding of x to one decimal place. For example for the first row of $\tilde{X}_{\mathscr{F}}$ we have $\tilde{x}_{14} = 1$, $\tilde{x}_{15} = 3$ and $\tilde{x}_{16} = -7$ which correspond to the parameters μ_4, γ, and α_1 respectively. However the circuits

encodes the complexity of the model: there are 26 circuits in the circuit basis but the supports now range from 8 to 10 points.

From Table 3 it is worth noting that for each number k of points removed the algorithm provides values of robustness equal to the 95th percentile.

k	p_{75}	p_{90}	p_{95}	r_*
0		r_0=0.527		
1	0.527	0.527	0.527	0.527
2	0.571	0.571	0.571	0.571
3	0.615	0.769	0.769	0.769
4	1	1	1	1

Table 3: Example 2, $\mathbb{E}(Y_j(t)) = \mu_j + \gamma m + \sum_{k=1}^{4} \alpha_k \cos(k\frac{\pi m}{4}) + \sum_{k=1}^{3} \beta_k \sin(k\frac{\pi m}{4})$

Acknowledgements

Roberto Fontana gratefully acknowledges financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. Roberto Fontana and Fabio Rapallo are members of the GNAMPA-INdAM group.

References

1. Butler, N.A., Ramos, V.M.: Optimal additions to and deletions from two-level orthogonal arrays. J. R. Stat. Soc. Ser. B **69**, 51–61 (2007)
2. Fontana, R., Rapallo, F., Wynn, H.P.: Circuits for robust designs. Stat. Pap. (Berl.), online first, 1–22 (2022)
3. Fontana, R., Rapallo, F.: Robustness of Fractional Factorial Designs through Circuits. In *SIS 2021 - Book of short papers*, C. Perna, N. Salvati, F. Schirripa Spagnolo Eds., Pearson (2021)
4. Ghosh, S.: On robustness of designs against incomplete data. Sankhya Ser. B **40**, 204–208 (1979)
5. Pistone, G., Riccomagno, E., Wynn, H.P.: Algebraic Statistics. Computational Commutative Algebra in Statitics Chapman & Hall/CRC, Boca Raton (2001)
6. Pukelsheim, F.: Optimal Design of Experiments, *Classics in Applies Mathematics*. SIAM, Philadelphia, PA (2006)
7. Street, D.J., Bird, E.M.: *D*-optimal orthogonal array minus *t* run designs. J. Stat. Theory Pract. **12**, 575–594 (2018)
8. Sturmfels, B.: Gröbner bases and convex polytopes, *University Lecture Series*, vol. 8. American Mathematical Society, Providence, RI (1996)
9. 4ti2 team: 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces (2018). URL https://4ti2.github.io
10. Mitchell, T. An algorithm for the construction of "*D*-optimal" experimental designs. Technometrics, **16**, 203–210 (1974)
11. SAS-Institute. SAS/QC 9.1: User's Guide. SAS Institute, Cary, NC (2004)

Robustifying the Rasch model with the Forward Search

Modello di Rasch robusto, mediante la Forward Search

Anna Comotti and Francesca Greselin

Abstract We introduces a forward search method for identifying atypical observations in Item Response Theory models for binary data (Rasch models). Our proposal introduces diagnostic tools, based on robust high-breakdown methodologies, to avoid distortion in the estimation of the model, and to single out outlying response patterns. Atypical response patterns usually deserve further investigation. Methods to initialize, progress, and monitor the Forward Search are explored. The simulated dataset showcases the effectiveness of the method in the presence of outliers.
Abstract *Introduciamo un metodo di Forward Search per identificare osservazioni atipiche nei modelli di Item Response Theory per dati binari (modelli di Rasch). Presentiamo strumenti diagnostici, basati su metodologie robuste, per evitare distorsione nella stima del modello e per individuare risposte anomale. Risposte atipiche solitamente necessitano ulteriori indagini. Vengono introdotti metodi per inizializzare, far progredire e monitorare la Forward Search. Risultati ottenuti mediante simulazioni mostrano l'efficacia del metodo in presenza di valori anomali.*

Key words: Item Response theory, Latent variable models, Outliers, Masking, Swamping, Monitoring, Robust estimations

1 Introduction

Dichotomous data frequently arise in the social sciences, and Factor Analysis, Item Response Theory, and its multivariate extension MIRT, are the most frequently em-

Anna Comotti
University of Milano-Bicocca, Department of Statistics and Quantitative Methods, e-mail: anna.comotti@unimib.it.

Francesca Greselin
University of Milano-Bicocca, Department of Statistics and Quantitative Methods, e-mail: francesca.greselin@unimib.it.

ployed models for assessing unobserved constructs, such as abilities, attitudes, exposure to an illness, and socioeconomic status.

Data are usually collected through surveys and measured over a scale. Atypical response patterns could arise due to cheating or guessing - when measuring abilities - and careless or systematic responding, exaggerated answers and also erroneous interpretation of an item, in all cases. Such cases should be identified and dealt with specific care. Moreover, aberrant response patterns could severely bias parameter estimation, and robust methodologies are needed to overcome this issue.

One of the first studies on the binary IRT model to remove the effects on random guessing in latent ability estimates is due to Waller Waller (1974). Based on the consideration that an item which is very difficult for a particular individual is an item which invites guessing, a conditional ML Estimation method is proposed, omitting from estimation any interaction for which this estimated probability is lower than some cutoff point, P_c.

A few years later, Wainer & Wright (1980) shows that gains in recovering abilities in the presence of guessing and untoward responses of other kinds can be obtained through the use of a robust jackknife. The jackknife procedures were supported only on empirical bases, therefore results cannot be generalized to two- or three-parameter logistic model.

A robust estimator of location is introduced in Mislevy & Bock (1982) based on the principle of Tukey's biweight. The larger the weighted differences between difficulty and ability, the heavier the down-weighting of the observation. The latter method is prone to produce infinite ability estimates for unexpected response patterns, whenever correct answers are sparse, so that Schuster & Yuan (2011) propose to adopt the Huber-type estimator.

A more recent paper Patton et al. (2019) proposes an iterated approach. First, the careless respondents are detected using the person-fit statistics in its standardized log-likelihood version l_z^* (Drasgow et al. 1985), then the corresponding responses are removed from the dataset, and this process continue by iteratively updating the item parameter estimates. An analogous method was used by Hong & Cheng (2019) and applied to the graded response models (polytomous scored items). First, the item parameters are estimated from the full data. Next, normalized l_z^*'s p-values are used as the weights to re-estimate the model.

Among some very recent contributions, Hong et al. (2020) investigated the efficacy of Insufficient Effort Responding (IER) detection methods as a data cleansing method. They evaluated six different detection methods and demonstrated which were best for flagging different types of IER.

A further iterative method was introduced by Liu & Liu (2021). The authors proposed to apply an iterative purification process based on a response time residual method with fixed item parameter estimates to detect non-effortful responses.

An IRT approach for detecting person-level outliers which rely on likelihood-based statistics has been proposed in Felt et al. (2017) to determine the most typical response patterns. The model is firstly estimated over the whole data and then new weights (derived from the limited information test statistic M_2, defined in Section 2), are employed to correct it.

Most methods cited till now start upon a first inference performed on the entire dataset. Therefore, they will not be preserved from the issue of *masking* and *swamping*. The masking effect occurs when an outlier is undetected because of the presence of a cluster of outliers and the swamping effect occurs when a *good* observation is incorrectly identified as an outlier.

The remaining part of the paper is organized as follows: Section 2 outlines the methodological extension of the Forward Search (FS) algorithm to the IRT model; Section 3 presents an application of the proposed methodology in published IRTs and simulated datasets; Section 4 discusses the main findings and provides directions for using the proposed diagnostic methodology for IRT.

2 The Forward Search for Rasch models

We propose a procedure for the detection of multiple outliers in IRT data based on the Forward Search (FS). The FS algorithm was initially introduced as an outlier detection tool for the estimation of covariance matrices Hadi (1992), and regression models Atkinson et al. (2000). It was subsequently extended to standard multivariate methods Atkinson et al. (2004), and factor analysis Mavridis & Moustaki (2009), and was recently applied in meta-regression Petropoulou et al. (2021).
We will use graphs to understand the relationship between a model and the data to which the model is fitted. Along model estimation, in a classical setting, we can lose information about the effect of individual observations on inferences about the form and parameters of the model. FS methods, instead, reveals how the fitted model depends on individual observations and on groups of observations. Robust procedures can sometimes reveal this structure, by down-weighting or discarding some observations. The novelty in FS is to combine robustness and a *forward search* through the data with model diagnostics and computer graphics. We introduce here easily understood plots for IRT models that use information from the whole sample to display the effect of each observation on a wide variety of aspects of the fitted model. The examples on simulated data show the amount of information the plots generate and the insights they provide.

Let Y be an $n \times k$ data matrix representing n observations on k items. We first introduces a plausible method to select a *basic* subset which contains p *good* observations and a *non-basic* subset which contains the remaining $n-p$ observations. Starting from the n initial observations, we extract a sub-sample of size p. The best basic set is chosen according to a test statistic which indicates how well the data are fitted by the estimated model. We have chosen to adopt the *limited information test statistic* M_2^* to evaluate such fit, due to its well known asymptotic properties

$$M_2^* = n(\pi - \hat{\pi})' C (\pi - \hat{\pi}). \tag{1}$$

It is called *limited information test statistic* because it is calculated for bivariate margins with positive responses to pair of items and it is widely used to assess

goodness-of-fit in the IRT field. C is a weight matrix of response patterns, π and $\hat{\pi}$ are the vectors of observed and expected (under the model) proportions for all the possible distinct 2^k response patterns. The M_2^* statistic is asymptotically χ^2 distributed and indicates an adequately fitting model when $p > 0.05$.

Second, for progressing in the forward search we increase the *basic* set adding the response pattern with the highest likelihood contribution. The likelihood contribution of the response pattern i, denoted by $\mathbf{Y_i} = (Y_{i1}, ..., Y_{ik})$, is given by

$$l_i = \prod_{j=1}^{k} p_{ij}^{Y_{ik}} (1 - p_{ij})^{1-Y_{ik}} \tag{2}$$

where, according to 1-PL model with $\lambda_j = 1 \; \forall j$, $p_{ik} = \dfrac{e^{\theta_i - \beta_k}}{1 + e^{\theta_i - \beta_k}}$. Likelihood contribution l_i is then weighted by the frequency of the $i-th$ response pattern already present in the basic set, to avoid that those patterns observed only once are most likely to enter in the last steps of the search.

Third, we rearrange such $n-p$ observations in the *non-basic* subset in ascending order accordingly, then increase the *basic* subset by adding to it the most plausible observation and decrease the *non-basic* subset to $n-p-1$ observations. This process is repeated until an appropriately chosen stopping criterion is met (usually derived from the FS plot itself).

3 Applications to simulated data

We generate 500 responses pattern from the 1-PL model with $k = 4$ items, with difficulty parameter estimation equal to $\beta_1 = -1, \beta_2 = -0.5, \beta_3 = 0.5, \beta_4 = 1$. In 1-PL model, the discrimination parameter is constrained to be the same for all items and we set it equal to 1. To test the proper functioning of the algorithm, we changed the last 25 response patterns of the initial dataset, forcing them to be equal to the most atypical response pattern (i.e., *0-0-1-1*). Such a response pattern is the least plausible one (having a score of 0 on the easiest items and a score of 1 on the most difficult ones); it appears only twice in the initial synthetic dataset.

We determine the *basic* set choosing the *best* among $H = 10$ sub-samples of size $p = 100$, randomly extracted from the initial sample. We choose this subset according to best fit given by (1). At each step of FS, we select the best candidate among response patterns in the non-basic set (which progressively decreases) to enter in the basic. Plots in Figure 1 shows the limited information goodness-of-fit statistic and the corresponding asymptotic p-values at each of the $n-p = 400$ steps of the forward search. It is clear that even when the search is initialized from a subset that yields large values for the M_2^* fit statistic, the *basic* set is incremented so that the fit is improved and stabilized for the larger part of the search; it substantially deteriorates in the last steps. The same, of course, holds for p-value: it reached very

high values after 100 steps of the FS and started to rapidly decrease in the last steps. The 25 atypical responses *0-0-1-1* enter in the last 30 steps of the search.

Figure 2 represents the forward plot for the adjusted residuals $r_{adj}(i)$ for each response pattern $i = 1, \ldots, 2^k$, given by $r_{adj}(i) = (\pi_i - \hat{\pi}_i)/\sigma_i$ where σ_i is the estimated standard deviation for $\pi_i - \hat{\pi}_i$. Pattern *0-0-1-1* reports a very high residual even in the last FS steps. Moreover, also the response pattern *0-0-0-1* (it appears only seven times in the initial synthetic dataset) has residuals far from 0 for all the FS steps.

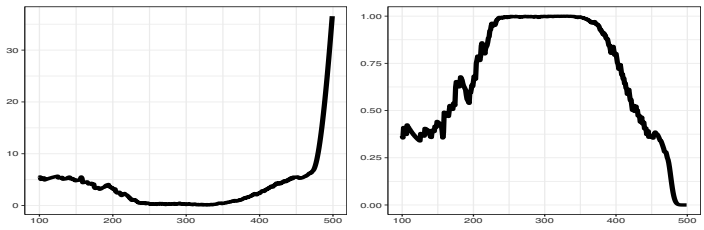


Fig. 1 Limited information statistics (left) and corresponding p-values (right). In the X axis the size of the basic set is indicated.

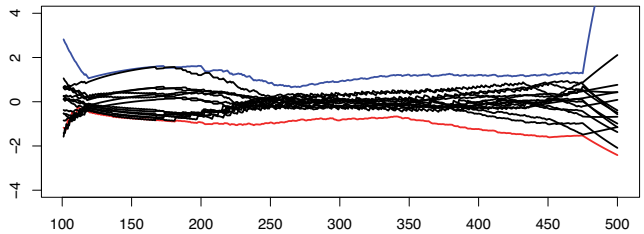


Fig. 2 Adjusted residuals for each response pattern, along the FS. The more extreme residuals are found for *0-0-1-1* (blue) and *0-0-0-1* (red). In the X axis the size of the basic set is indicated.

4 Conclusion and future directions

We explored the topic of atypical response pattern detection using the Forward Search in the IRT modelling approach. Such method was applied to the simplest IRT model, the Rasch model where the discrimination parameter is constrained across items to be equal to 1. Even if this work is still in progress, the good results obtained on simulated data generated from a 4-items Rasch model encourage further developments, which will be performed on 2-PL model and polytomous models. They also established the first promising application of FS algorithm to IRT models.

References

Atkinson, A. C., Riani, M., Cerioli, A. et al. (2004), *Exploring multivariate data with the forward search*, Vol. 1, Springer.

Atkinson, A. C., Riani, M. & Riani, M. (2000), *Robust diagnostic regression analysis*, Vol. 2, Springer.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985), 'Appropriateness measurement with polychotomous item response models and standardized indices', *British Journal of Mathematical and Statistical Psychology* **38**(1), 67–86.

Felt, J. M., Castaneda, R., Tiemensma, J. & Depaoli, S. (2017), 'Using person fit statistics to detect outliers in survey research', *Frontiers in Psychology* **8**(MAY), 1–9.

Hadi, A. S. (1992), 'Identifying multiple outliers in multivariate data', *Journal of the Royal Statistical Society: Series B (Methodological)* **54**(3), 761–771.

Hong, M. R. & Cheng, Y. (2019), 'Robust maximum marginal likelihood (RMML) estimation for item response theory models.', *Behavior research methods* **51**(2), 573–588.
URL: *http://link.springer.com/10.3758/s13428-018-1150-4 http://www.ncbi.nlm.nih.gov/pubmed/30350024*

Hong, M., Steedle, J. T. & Cheng, Y. (2020), 'Methods of detecting insufficient effort responding: Comparisons and practical recommendations', *Educational and psychological measurement* **80**(2), 312–345.

Liu, Y. & Liu, H. (2021), 'Detecting noneffortful responses based on a residual method using an iterative purification process', *Journal of Educational and Behavioral Statistics* **46**(6), 717–752.

Mavridis, D. & Moustaki, I. (2009), 'The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data', *Journal of Computational and Graphical Statistics* **18**(4), 1016–1034.

Mislevy, R. J. & Bock, R. D. (1982), 'Biweight estimates of latent ability', *Educational and psychological measurement* **42**(3), 725–737.

Patton, J. M., Cheng, Y., Hong, M. & Diao, Q. (2019), 'Detection and treatment of careless responses to improve item parameter estimation', *Journal of Educational and Behavioral Statistics* **44**(3), 309–341.

Petropoulou, M., Salanti, G., Rücker, G., Schwarzer, G., Moustaki, I. & Mavridis, D. (2021), 'A forward search algorithm for detecting extreme study effects in network meta-analysis', *Statistics in medicine* **40**(25), 5642–5656.

Schuster, C. & Yuan, K.-H. (2011), 'Robust estimation of latent ability in item response models', *Journal of Educational and Behavioral Statistics* **36**(6), 720–735.

Wainer, H. & Wright, B. D. (1980), 'Robust estimation of ability in the rasch model', *Psychometrika* **45**(3), 373–391.

Waller, M. I. (1974), 'Removing the effects of random guessing from latent trait ability estimates', *ETS Research Bulletin Series* **1974**(1), i–50.

A novel estimation procedure for robust CP model fitting

Perfezionamento della procedura per la stima robusta del modello CP

Valentin Todorov and Violetta Simonacci and Michele Gallo and Nikolay Trendafilov

Sommario The usual way of parameter estimation in CANDECOM/PARAFAC (CP) is an alternating least squares (ALS) procedure that yields least-squares solutions and provides consistent outcomes but at the same time has several deficiencies, like sensitivity to the presence of outliers in the data, slow convergence, and susceptibility to degeneracy conditions. A number of works have addressed these weaknesses, but to our knowledge, there is no outlier-robust procedure that is highly computationally efficient at the same time, especially for large data sets. We propose a robust procedure based on an integrated estimation algorithm, alternative to ALS, which guards against outliers and is computationally efficient at the same time.

Sommario *Il metodo comunemente usato per la stima dei parametri nel modello CP è una procedura dei minimi quadrati alternati (ALS). Questo algoritmo fornisce soluzioni ai minimi quadrati e produce risultati stabili ma, allo stesso tempo, registra diverse carenze come la sensibilità alla presenza di valori anomali nei dati, convergenza lenta, e suscettibilità a condizioni di degenerazione. Numerosi lavori hanno affrontato queste debolezze ma, per quanto ne sappiamo, non esiste una procedura robusta che sia allo stesso tempo altamente efficiente dal punto di vista computazionale, specialmente per grandi insiemi di dati. Proponiamo una procedura robusta basata su un algoritmo di stima integrato, alternativo all'ALS, che protegge dai valori anomali ed è computazionalmente efficiente allo stesso tempo.*

Valentin Todorov
United Nations Industrial Development Organization (UNIDO), VIC, Vienna, e-mail: valentin@todorov.at

Violetta Simonacci
University of Naples Federico II, Italy e-mail: violetta.simonacci@unina.it

Michele Gallo
University of Naples-L'Orientale, Naples, 80134, Italy e-mail: mgallo@unior.it

Nikolay Trendafilov
University of Naples-L'Orientale, Naples, 80134, Italy e-mail: ntrendafilov@unior.it

Key words: ALS, ATLD-ALS, robustness, outliers, computational efficiency

1 Introduction

The standard multivariate analysis addresses data sets represented as two-dimensional matrices. In recent years, an increasing number of application areas like chemometrics, computer vision, econometrics and social network analysis involve analysis of data sets that are represented as multidimensional arrays and multiway data analysis becomes popular as an exploratory analysis tool. Different techniques exist to analyze such multi-way data but CANDECOMP/PARAFAC (CP) is one of the most popular. The usual way of parameter estimation in CP is an alternating least squares (ALS) procedure which yields least-squares solutions and provides consistent outcomes. Together with these desirable features, the ALS procedure suffers several major flaws which might be particularly problematic for large-scale problems: slow convergence and sensitiveness to degeneracy conditions such as over-factoring, collinearity, bad initialization and local minima. Furthermore, it is well-known that algorithms which rely on least squares easily break down in the presence of outliers. The issue of non-robustness of the ALS procedure was addressed by [2] and software is available in the R package **rrcov3way**. The other issues were addressed in a number of works proposing algorithms more efficient than ALS. However, often these do not provide stable results because the increased speed might come at the expense of accuracy. An integrated algorithm was proposed in [4] which seems to combine improved speed and stability. The purpose of this work is to develop further this algorithm by adding capabilities for dealing with outliers present in the data.

2 The CP model, the ALS algorithm and its robust version

The CP model [1, 3] decomposes the 3-way data array $\underline{\mathbf{X}}(I \times J \times K)$ with a generic element x_{ijk} into three loading matrices $\mathbf{A}$ $(I \times R)$, $\mathbf{B}$ $(J \times R)$, $\mathbf{C}$ $(K \times R)$ with R components (using the same number for each mode). The CP model can be written formally as

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C} \otimes \mathbf{B})^\top + \mathbf{E}_A, \tag{1}$$

where $\mathbf{X}_A$ and $\mathbf{E}_A$ are are the original array and the error array unfolded with respect to mode A and the symbol $\otimes$ represents the *Kronecker product* between two matrices. To estimate the optimal component matrices the residual sum of squares

$$||\mathbf{E}_A||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2 = \sum_{i=1}^{I}||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 = \sum_{i=1}^{I} RD_i^2 \tag{2}$$

is minimized. The residual distance (RD) for observation i is thus given by

$$RD_i = ||\mathbf{x}_i - \hat{\mathbf{x}}_i|| = \sqrt{\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2} \tag{3}$$

and the estimation is equivalent to the minimization of the sum of the squared distances. With ALS the component matrices are estimated one at a time, keeping the estimates of the other component matrices fixed, i.e. we start with initial estimates of **B** and **C** and find an estimate for **A** conditional on **B** and **C** by minimizing the objective function. Estimates for **B** and **C** are found analogously. The iteration continues until the relative change in the model fit is smaller than a predefined constant.

The idea of a robust version of CP is to identify enough ”good“ observations and to perform the classical ALS on these observations. This is repeated until no significant change is observed. Finally, a reweighting step is carried out to improve the efficiency of the estimators. In order to identify the "good" observations a robust version of principal component analysis on the unfolded array is used. We will call this procedure R-ALS in the rest of the paper. It is obvious that the robust procedure will be much more time consuming than the classical one, repeating many times the ALS optimization. Therefore, any improvement of the performance of the parameter estimation procedure will contribute to the improvement of the performance of the complete robust procedure.

3 The ATLD and INT2 algorithms

The alternating trilinear decomposition (ATLD) proposed by [6] seems to be the most efficient method among the proposed alternatives to ALS. It is based on the use of three loss functions with different response surfaces. However, these advantages are obtained at the cost of unstable results and non-least-squares solutions. To cope with these issues [4] proposed a multi-optimization procedure in which ATLD is followed by ALS estimation steps which is demonstrated to be quite effective. The robust procedure R-ALS described in Section 2 is entirely based on ALS and thus suffers the slow convergence and other disadvantages of this algorithm. We propose to replace ALS by INT2 thus obtaining a new robust estimation procedure which we will call R-INT2. As before it starts by robust principal components to identify any outlying points and then iterates using the INT2 algorithm until no significant change is observed. After convergence a reweighting step with INT2 is conducted which produces the final solution.

4 Simulation study

The performance of the newly proposed algorithm R-INT2 for robust estimation of trilinear CP models will be demonstrated in a brief simulation study comparing

classical CP, R-ASL and R-INT2. First of, all we want to verify that R-INT2 works well on data sets with and without contamination by identifying the outliers at least as good as R-ALS retrieving solutions with good statistical quality. At the same time we want to verify that the convergence is improved significantly and thus the computational time is reduced.

These two aspects will be illustrated on three-way data generated as in [4], [5] and [2]. The three-way arrays have $I = 50$ observations, $J = 100$ variables and $K = 10$ occasions and the number of factors is $R = 2$. The loadings matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are generated as randomly multivariate normal distributed $N_R(\mathbf{0}, \boldsymbol{\Sigma}_R)$ where $\boldsymbol{\Sigma}_R$ is a diagonal matrix with (10, 2) on the diagonal.

Such data does not contain any contamination to start. Different configurations of outliers can then be considered. In this work, because of the space constraints, only the so-called bad leverage points will be generated. An example of such outliers is shown in the outlier map illustrated in the left panel of Fig. 1 - these are the observations lying in the upper right quadrant. Regular observations lie in the lower left quadrant. Other types of outliers, not exemplified here, include good leverage points (lower right quadrant) and residual outliers (upper left quadrant). For more details about the generation of the different types of outliers see [2, page 158]. In the simulation study, apart from the case with no contamination, two other cases will be studied - with 10% and 20% of bad leverage points. For each setup, in order to account for minor statistical fluctuations 20 replicates were computed yielding the following three quantities which will represent how close the obtained estimates are to the original data: the mean square error (MSE), the angle between the estimated subspace and the true subspace spanned by the B-loadings and analogously, for the C-loadings. Also, the computation time and the number of iterations were recorded. The mean squared error (MSE) is given by

$$MSE = \frac{1}{w} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} w_i (x_{ijk} - \hat{x}_{ijk})^2 \tag{4}$$

with $w = \sum_{i=1}^{I} w_i$ and $w_i = 0$ if the i-th observation is outlier or $w_i = 1$ otherwise. Thus the MSE will be computed only for the regular observations. The angle between the estimated subspace and the original one is given by

$$maxsub = \max_{\mathbf{b}_1} \min_{\mathbf{b}_2} \arccos(\mathbf{b}_1^\top \mathbf{b}_2) \tag{5}$$

This subspace angle has to be as small as possible and is reported in radians. We use the function `subspace()` from the **R** package **pracma** to compute *maxsub*.

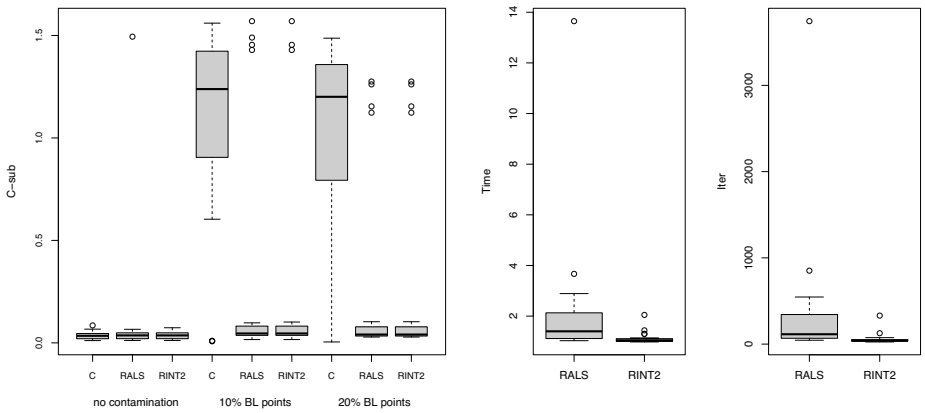


Figura 2 Angle of C-loadings (left) of classical CP (C), robust CP with ALS (R-ALS) and robust CP with INT2 (R-INT2) for simulation settings with no outliers, 10% and 20% bad leverage points and computational time and number of iterations (right) for 20% bad leverage points.

All three estimators perform equally well on clean data both in terms of MSE and maxsub (see Fig. 1, right panel and Fig. 2, left panel). However, when outliers are added to the data (10% and 20%) the classical CP is influenced - the MSE increases and the quality of the fit of the loadings decreases. These effects are even more pronounced when the outlier fraction is increased to 20%. There is no much difference in the performance of the two robust methods in terms of MSE and maxsub, howe-

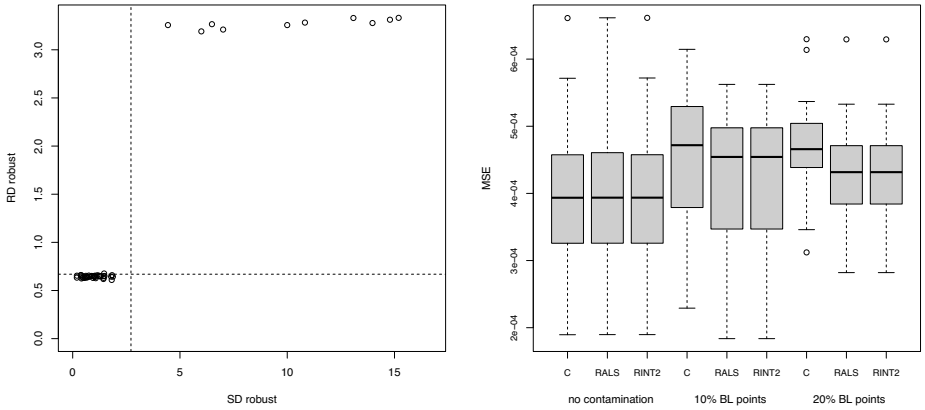


Figura 1 Example of an outlier map for simulated data with 20 percent bad leverage points (left) and MSE values of classical CP (C), robust CP with ALS (R-ALS) and robust CP with INT2 (R-INT2) for simulation settings without contamination and with bad leverage points (right).

ver, if we look at the right panel of Fig. 2 which presents their performance in terms of computational time and number of iterations the gain in performance is obvious. The median time of R-INT2 is more than 30% lower than that of R-ALS which also has much higher variance. This is due to the reduced number of iterations, as it is seen in the right part of the same Figure.

5 Summary and conclusions

We combine the robust procedure for CP proposed by [2] with the highly efficient estimation algorithm INT2 proposed by [4] in order to obtain a fast estimation and robust to outliers CP modeling technique. The conducted simulation study demonstrates the advantages of the new procedure in terms of computational time and at the same time shows that the robustness properties and the statistical efficiency have not been affected. Future work should bring a thorough investigation of the properties of the algorithm, comparison to the many existing fast alternatives and studying the possibilities for combination with other computational algorithms.

References

[1] Carroll J, Chang J (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrica 35(3):283–319

[2] Engelen S, Hubert M (2011) Detecting outlying samples in a parallel factor analysis model. Analytica Chemica Acta 705:155–165

[3] Harshman RA (1970) Foundations of the PARAFAC procedure: Models and conditions for an ”explanatory“ multi-modal factor analysis. Tech. Rep. 10, UCLA

[4] Simonacci V, Gallo M (2020) An ATLD—ALS method for the trilinear decomposition of large third-order tensors. Soft Computing 18

[5] Tomasi G, Bro R (2006) A comparison of algorithms for fitting the PARAFAC model. Computational Statistics & Data Analysis 50(7):1700–1734

[6] Wu HL, Shibukawa M, Oguma K (1998) An alternating trilinear decomposition algorithm with application to calibration of HPLC-DAD for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. Journal of Chemometrics 12

A robust approach for functional ANOVA with application to additive manufacturing

Un approccio robusto per l'ANOVA funzionale con un'applicazione all'additive manufacturing

Fabio Centofanti, Bianca Maria Colosimo, Marco Luigi Grasso, Alessandra Menafoglio, Biagio Palumbo, Simone Vantini

Abstract A new robust nonparametric functional analysis of variance method is presented, which reduces the weights of outlying functional data on the results of the analysis. It is implemented through a permutation test based on a statistic obtained via a functional extension of the classical robust M-estimator. Its favourable properties are shown in an additive manufacturing application based the laser powder bed fusion process.

Abstract *In questo lavoro, viene presentato un nuovo metodo robusto nonparametrico per l'analisi della varianza funzionale, in grado di mitigare l'eventuale distorsione dovuta alla presenza di dati anomali. Il metodo si basa su un test di permutazione con una statistica test ottenuta tramite un'estensione funzionale del classico stimatore M robusto. Le ottime proprietà statistiche del metodo proposto vengono illustrate mediante un'applicazione nell'ambito della produzione additiva basata sul processo di fusione laser a letto di polvere.*

Key words: functional data analysis, functional analysis of variance, robustness, additive manufacturing

Fabio Centofanti, Biagio Palumbo
Dept. of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125, Naples, Italy
e-mail: fabio.centofanti@unina.it,biagio.palumbo@unina.it

Bianca Maria Colosimo, Marco Luigi Grasso
Dept. of Mechanical Engineering, Politecnico di Milano, Milan, Italy.
e-mail: biancamaria.colosimo@polimi.it,marcoluigi.grasso@polimi.it

Alessandra Menafoglio, Simone Vantini
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy.
e-mail: alessandra.menafoglio@polimi.it,simone.vantini@polimi.it

1 Introduction and methods

The development of data acquisition methods allow the analysis of complex systems in several operating conditions. To address this increasing complexity new statistical tools are needed. To this end, functional data analysis (FDA) [15, 9, 10], has been introduced and model observation units as functions defined on a compact domain. Several industrial applications, which apply FDA methods, have appeared in the literature so far [14, 18, 13, 19, 1, 4, 2, 5].

In this setting, the functional analysis of variance (FANOVA), which is extension of the classical (non-functional) analysis of variance problem, consists in the identification of significant differences in group functional means belonging to a sample with varying experimental conditions. As in the classical setting, FANOVA methods tend to be highly sensitive to the presence of data points that differ significantly from other observations, i.e., *outliers*. FANOVA methods rely in fact on statistics that combine in a quadratic fashion the functional mean, which are known to be highly sensitive to the presence of outliers.

In this work, we consider the robust FANOVA method (RoFANOVA), proposed by Centofanti et al. [3], that is able to test, in a nonparametric fashion, the differences among group functional means. The RoFANOVA approach is based on the scale equivariant functional M-estimator, which extends the M-estimator [17]. It is estimated through the FuNMAD estimator of dispersion, which is the functional extension of the normalized median absolute deviation [12], and it is obtained numerically through the standard iteratively re-weighted least-squares algorithm [17]. Several versions of the scale equivariant functional M-estimator are available by using the Huber, Tukey, Hampel, and optimal families of loss functions [3].

The significance of the coefficients in the two-way FANOVA model is tested through a nonparametric permutational approach based on the functional extension of the robust F-statistic proposed by Schrader et al. [16], which is a robust version of the classical F-test statistic. This test statistic is obtained as the difference between the standardized residual sum of dispersions under the reduced and the full model and measures the discrepancy between residuals of the model under the null and the alternative hypotheses. The distribution of the considered statistic under the null hypothesis is estimated through a permutational approach that consists in permuting the data without constraints [11].

2 An additive manufacturing application

In this section, the RoFANOVA approach is applied to a metal additive manufacturing process known as laser powder bed fusion (L-PBF) [7, 6, 8]. The video image frames, acquired during the process, are transformed into a functional format to obtain the *spatter intensity*, defined as the amount of spatters observed in any given region of the bi-dimensional video-frame space. Six specimens are produced by varying the energy density provided by the laser to the material, for six production layers. For

each layer, the laser moves along a predefined path that consists of parallel scan lines and whose orientation changes layer by layer. The aim of the analysis is to assess whether the energy density level, layer, and interaction significantly influences the spatter intensity. The cross-sections of the spatter intensity function are represented in Figure 1 at different energy density levels and in different layers.

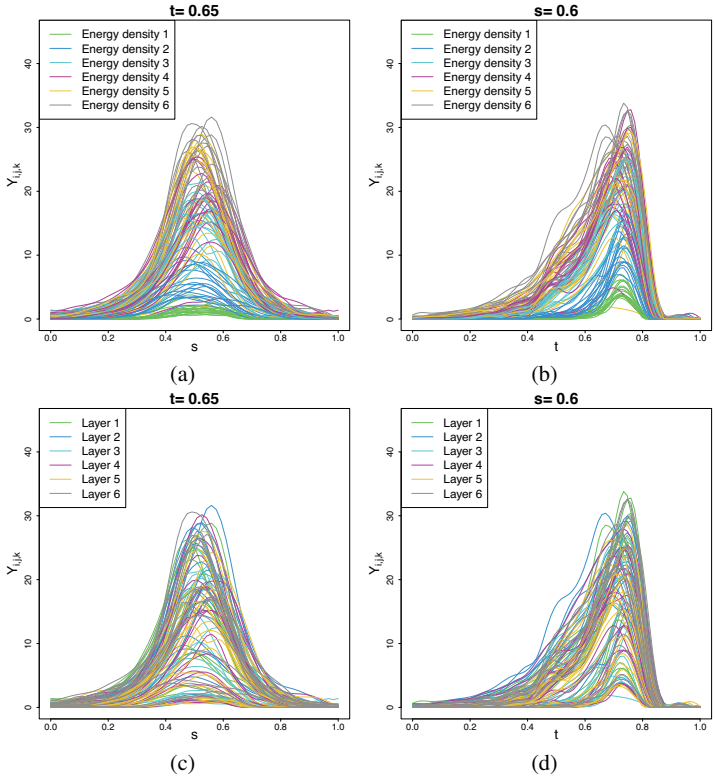


Fig. 1 Cross-sections of the spatter intensity function ($Y_{i,j,k}$), for different energy density levels ((a) and (b)) and different layers ((c) and (d)).

In these setting, the RoFANOVA test can be suitably applied. Several loss functions are considered and tuned to achieve the 95% asymptotic efficiency at the normal model. All tests agree that the energy density and the layer as well as their interaction effects are significant. Moreover, the energy density is increasing with the intensity of the spatter, and different layers show different increasing patterns which may be related with different scan directions. Thus, the influence on the spatter intensity of the energy density, the production layer and their interaction is found significant, differently from competing FANOVA methods that do not consider significant the interaction between energy density and layer [3].

References

1. Capezza, C., Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart for monitoring ship CO_2 emissions. Quality and Reliability Engineering International (2021)
2. Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: Functional clustering methods for resistance spot welding process data in the automotive industry. Applied Stochastic Models in Business and Industry **37**(5), 908–925 (2021)
3. Centofanti, F., Colosimo, B.M., Grasso, M.L., Menafoglio, A., Palumbo, B., Vantini, S.: Robust functional anova with application to additive manufacturing (2021)
4. Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart. Technometrics **63**(3), 281–294 (2021)
5. Colosimo, B., Grasso, M., Garghetti, F., Rossi, B.: Complex geometries in additive manufacturing: A new solution for lattice structure modeling and monitoring. Journal of Quality Technology (2021). DOI 10.1080/00224065.2021.1926377
6. Colosimo, B.M., Grasso, M.: On-machine measurement, monitoring and control. Precision Metal Additive Manufacturing p. 102 (2020)
7. Colosimo, B.M., Huang, Q., Dasgupta, T., Tsung, F.: Opportunities and challenges of quality engineering for additive manufacturing. Journal of Quality Technology **50**(3), 233–252 (2018)
8. Grasso, M., Remani, A., Dickins, A., Colosimo, B., Leach, R.: In-situ measurement and monitoring methods for metal powder bed fusion: An updated review. Measurement Science and Technology **32**(11) (2021). DOI 10.1088/1361-6501/ac0b6b
9. Horváth, L., Kokoszka, P.: Inference for functional data with applications. Springer Science & Business Media (2012)
10. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
11. Manly, B.F.: Randomization, bootstrap and Monte Carlo methods in biology, vol. 70. CRC press (2006)
12. Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust statistics: theory and methods (with R). John Wiley & Sons (2019)
13. Menafoglio, A., Grasso, M., Secchi, P., Colosimo, B.: Profile monitoring of probability density functions via simplicial functional pca with application to image data. Technometrics **60**(4), 497–510 (2018). DOI 10.1080/00401706.2018.1437473
14. Noorossana, R., Saghaei, A., Amiri, A.: Statistical analysis of profile monitoring. John Wiley & Sons (2012)
15. Ramsay, J.O.: Functional data analysis. Wiley Online Library (2005)
16. Schrader, R.M., Mc Kean, J.W.: Robust analysis of variance. Communications in Statistics-Theory and Methods **6**(9), 879–894 (1977)
17. Sinova, B., Gonzalez-Rodriguez, G., Van Aelst, S., et al.: M-estimators of location for functional data. Bernoulli **24**(3), 2328–2357 (2018)
18. Wang, K., Tsung, F.: Using profile monitoring techniques for a data-rich environment with huge sample size. Quality and Reliability Engineering International **21**(7), 677–688 (2005). DOI 10.1002/qre.711
19. Wells, L., Megahed, F., Niziolek, C., Camelio, J., Woodall, W.: Statistical process monitoring approach for high-density point clouds. Journal of Intelligent Manufacturing **24**(6), 1267–1279 (2013). DOI 10.1007/s10845-012-0665-2

Modeling unconditional M-quantiles in a regression framework

Modelli di regressione quantile per la stima di M-quantili marginali

Luca Merlo, Lea Petrella and Nicola Salvati

Abstract In this paper we develop a quantile regression model for estimating unconditional M-quantiles in the presence of covariates. Extending the paper by Firpo et al (2009), we assess the impact of small changes in the explanatory variables across the entire unconditional distribution of the dependent variable by running a mean regression of the recentered influence function on the covariates. The proposed method is applied on data originating from the May 1985 Current Population Survey conducted by the US Census Bureau.

Abstract *In questo articolo sviluppiamo un modello di regressione quantile per la stima di M-quantili marginali della variabile dipendente in presenza di covariate. Estendendo il lavoro di Firpo et al (2009), stimiamo l'effetto marginale delle variabili esplicative sull'intera distribuzione marginale della variabile dipendente tramite un'analisi di regressione della funzione d'influenza sulle covariate. La metodologia proposta viene applicata sui dati raccolti dal Current Population Survey del 1985 condotto dall'US Census Bureau.*

Key words: Expectiles, Influence function, M-estimation, RIF regression

Luca Merlo
Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Roma, e-mail: luca.merlo@uniroma1.it

Lea Petrella
Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Roma, e-mail: lea.petrella@uniroma1.it

Nicola Salvati
University of Pisa, Via Cosimo Ridolfi, 10, 56124 Pisa, e-mail: salvati@ec.unipi.it

1 Introduction

Quantile Regression (QR), as proposed by Koenker and Bassett Jr (1978), has proven to be a powerful tool to explore conditional distributions in many empirical applications. However, if one is interested in how the whole unconditional distribution of the dependent variable responds to changes in the covariates, using the well-known QR would yield misleading inferences (see Firpo et al 2009 and Borah and Basu 2013). Motivated by this interest, Firpo et al (2009) proposed the Unconditional Quantile Regression (UQR) approach for modeling unconditional quantiles of a dependent variable as a function of the explanatory variables. This method builds upon the concept of Recentered Influence Function (RIF) which originates from a widely used tool in robust statistics, namely the Influence Function (IF) discussed in Hampel et al (2011). The RIF of a distributional statistic ν is obtained by adding back the statistic to the IF and it can be thought of as the contribution of an individual observation on ν. In the regression framework where covariates are available, Firpo et al (2009) proposed to replace the dependent variable with the RIF to model the unconditional quantiles of the response and evaluate the effect of changes in the law of the covariates on unconditional quantiles. When the interest of the research is concentrated on the entire distribution of a response variable, in addition to the classical QR, a possible alternative is represented by the M-quantile regression (MQR) approach proposed by Breckling and Chambers (1988). This method provides a "quantile-like" generalization of the mean regression based on influence functions, combining in a common framework the robustness and efficiency properties of quantiles and expectiles (Newey and Powell 1987), respectively.

In this article, we extend the UQR of Firpo et al (2009) to the M-quantile regression framework by using the well-known Huber loss (Huber 1964). Specifically, we develop the Unconditional M-quantile Regression (UMQR) to model the M-quantiles and expectiles of the unconditional distribution of the response variable. In order to analyze the impact of changes in the distribution of explanatory variables on the entire unconditional distribution of the response, we regress the RIF of the proposed model on the covariates.

2 Methodology

Let Y denote a scalar random variable with absolutely continuous distribution function F_Y. The M-quantile of order $\tau \in (0,1)$ of Y is defined as the solution, $\theta_\tau \in \mathbb{R}$, of the following estimating equation:

$$\int \psi_\tau(y-\theta_\tau) dF_Y(y) = 0, \tag{1}$$

where $\psi_\tau(u) = \mid \tau - \mathbf{1}_{(u<0)} \mid \psi(u/\sigma_\tau)$, with ψ being the first derivative of a convex loss function ρ and σ_τ is a suitable scale parameter. In this work, we consider the

popular Huber influence function (Huber (1964)):

$$\psi(u) = u\mathbf{1}_{(|u|\leq c)} + c\,\text{sign}(u)\mathbf{1}_{(|u|>c)}, \tag{2}$$

where c denotes a tuning constant bounded away from zero that can be used to trade robustness for efficiency in the model fit, with increasing robustness when c is chosen to be positive and close to 0 and increasing efficiency when c is chosen to be large and positive. In particular, M-quantiles nicely include quantiles when $c \to 0$, $\psi(u) = \text{sign}(u)$, and expectiles when $c \to \infty$, $\psi(u) = u$.

To build the UMQR model, it follows from Firpo et al (2009) and Hampel et al (2011) that the RIF of the M-quantile θ_τ is defined as:

$$RIF(y;\theta_\tau) = \theta_\tau + IF(y;\theta_\tau) = \theta_\tau + \frac{\psi_\tau(y-\theta_\tau)}{\int \psi'_\tau(y-\theta_\tau)dF_Y(y)}, \tag{3}$$

where $IF(y;\theta_\tau)$ is the IF of θ_τ and $\psi'(u) = \mathbf{1}_{(|u|<c)}$ is the derivative of ψ in (2). In a regression framework when covariates $\mathbf{X} \subset \mathbb{R}^k$ are available, from (3) we define the UMQR model as follows:

$$\mathbb{E}[RIF(Y;\theta_\tau) \mid \mathbf{X} = \mathbf{x}] = \theta_\tau + \mathbb{E}\Big[\frac{\psi_\tau(y-\theta_\tau)}{\int \psi'_\tau(y-\theta_\tau)dF_Y(y)}\Big|\mathbf{X} = \mathbf{x}\Big]. \tag{4}$$

Our objective is to identify how small changes in the distribution of $\mathbf{X}$ affect the M-quantile of the unconditional distribution of Y. From (4) and Firpo et al (2009), the unconditional effect of the τ-th M-quantile, that we denote Unconditional M-quantile Partial Effect, α_τ, is formally defined as:

$$\alpha_\tau = \int \frac{d\mathbb{E}[RIF(Y;\theta_\tau) \mid \mathbf{X} = \mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}) = \frac{1}{s_\tau}\int \frac{d\mathbb{E}[\psi_\tau(Y-\theta_\tau) \mid \mathbf{X} = \mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}), \tag{5}$$

where $F_{\mathbf{X}}$ is the distribution function of $\mathbf{X}$ and $s_\tau = \int \psi'_\tau(y-\theta_\tau)dF_Y(y)$. As suggested by Firpo et al (2009), we can estimate α_τ in (5) via a mean regression of the $RIF(Y;\theta_\tau)$ as dependent variable onto $\mathbf{X}$ by using a two-step procedure. Specifically, an estimate $\widehat{\theta}_\tau$ of θ_τ is obtained by solving (1) via Iterative Reweighted Least Squares, substitute $\widehat{\theta}_\tau$ in (3) and then regress the $RIF(Y;\widehat{\theta}_\tau)$ on $\mathbf{X}$.

3 Application

We investigate the impact of unions on the distribution of hourly wages using a sample of 534 observations originating from the 1985 Current Population Survey conducted by the US Census Bureau. We fit the proposed model at different M-quantile levels, $\tau = (0.1, 0.5, 0.9)$, and compare the results with standard conditional M-quantile regressions. The tuning constant c in (2) has been set to 1.345 and 100 which, in the second case, allows us to obtain the Unconditional Expectile Regres-

sion (UER). The results in Table 1 highlight that the estimated effect of union status is very different on the conditional and unconditional distributions of log wages, especially in the upper tails. Specifically, by looking at the conditional M-quantile and expectile regressions (MQR and ER), the effect of unions is positive and monotonically declines as the M-quantile level increases. On the contrary, unions have a positive but increasing effect on the unconditional distribution of wages and their impact is not statistically significant at $\tau = 0.9$. This demonstrates that if one is interested in the overall effect of unions on wage inequality, the proposed unconditional regression should be used to obtain such unconditional effect.

Variable	MQR			UMQR			ER			UER		
τ-th level	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Union	**0.271**	**0.192**	**0.156**	**0.201**	**0.234**	0.074	**0.273**	**0.204**	**0.161**	**0.181**	**0.204**	0.076
	(0.061)	(0.050)	(0.063)	(0.043)	(0.050)	(0.078)	(0.064)	(0.052)	(0.065)	(0.040)	(0.048)	(0.070)
Female	**−0.153**	**−0.255**	**−0.276**	**−0.164**	**−0.255**	**−0.279**	**−0.147**	**−0.229**	**−0.249**	**−0.151**	**−0.229**	**−0.251**
	(0.047)	(0.039)	(0.048)	(0.048)	(0.041)	(0.055)	(0.049)	(0.040)	(0.050)	(0.046)	(0.040)	(0.051)
Married	**0.127**	**0.094**	0.036	**0.114**	**0.103**	0.013	**0.111**	0.080	0.029	0.089	0.080	0.013
	(0.050)	(0.042)	(0.052)	(0.051)	(0.045)	(0.059)	(0.052)	(0.042)	(0.053)	(0.047)	(0.043)	(0.054)
Education	**0.073**	**0.101**	**0.105**	**0.072**	**0.100**	**0.116**	**0.076**	**0.096**	**0.103**	**0.069**	**0.096**	**0.106**
	(0.009)	(0.008)	(0.010)	(0.008)	(0.008)	(0.015)	(0.010)	(0.008)	(0.010)	(0.007)	(0.008)	(0.013)
Experience	**0.007**	**0.012**	**0.012**	**0.007**	**0.011**	**0.013**	**0.007**	**0.011**	**0.011**	**0.007**	**0.011**	**0.012**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)

Table 1 M-quantile and expectile regression results at $\tau = (0.1, 0.5, 0.9)$. Parameter estimates are displayed in boldface when significant at the 5% level.

References

Borah BJ, Basu A (2013) Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. Health Economics 22(9):1052–1070

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75(4):761–771

Firpo S, Fortin NM, Lemieux T (2009) Unconditional quantile regressions. Econometrica: Journal of the Econometric Society 77(3):953–973

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) Robust statistics: the approach based on influence functions, vol 196. John Wiley & Sons

Huber PJ (1964) Robust estimation of a location parameter. Annals of Mathematical Statistics 35(1):73–101

Koenker R, Bassett Jr G (1978) Regression quantiles. Econometrica: Journal of the Econometric Society pp 33–50

Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica pp 819–847

Model-based clustering

Bayesian mixtures of semi-Markov models

Misture Bayesiane di modelli semi-Markov

Rosario Barone and Andrea Tancredi

Abstract In this paper we propose a clustering technique for continuous-time semi-Markov models in order to take account of groups of individuals having similar process realizations. In fact fitting standard parametric models in presence of heterogeneity between population groups may produce biased inferences for relevant process feautres. To model individual heterogeneity we consider a Dirichlet process mixture (DPM) of semi-Markov continuous-time models. We also consider the case of discretely observed trajectories of continuous time processes, providing an algorithm which clusterize the observations after having reconstructed the continuous-time paths between the observed points. Full MCMC inference is performed with an application to a real dataset.

Abstract *In questo articolo proponiamo un metodo di clustering per modelli semi-Markov a tempo continuo al fine di tenere conto di gruppi di individui con realizzazioni del processo simili. Infatti, l'utilizzo di modelli parametrici standard in presenza di eterogeneità può produrre inferenze distorte. Per modellare l'eterogeneità individuale consideriamo una mistura di processi di Dirichlet (DPM) di modelli semi-Markov a tempo continuo. Consideriamo anche il caso di processi a tempo continuo osservati in punti discreti, presentando un algoritmo che consente di raggruppare le osservazioni dopo aver simulato le traiettorie tra i punti discreti. nel lavoro verrà presentato un algoritmo MCMC e un'applicazione con dati reali.*

Key words: Dirichlet process prior, Multi-state models, MCMC, Time series clustering.

Rosario Barone
University of Rome Tor Vergata, e-mail: rosario.barone@uniroma2.it.

Andrea Tancredi
Sapienza University of Rome, e-mail: andrea.tancredi@uniroma1.it.

1 Introduction

In a panel data set with individuals observed in time, clustering techniques may be useful for finding groups of similar individuals. With exception for the Markov case [1], all the mixtures approaches for multi-state models are related to finite mixtures and consider completely observed processes [2]. In this paper we tackle the clustering problem for the class of semi-Markov processes. In general, note that inference for semi-Markov models may present computational difficulties when observations are at discrete time points, so that the process is not completely observed. In fact, the likelihood function is not available and approximation methods are required. [3] reconstruct the likelihood function by simulating the trajectories between the observed points with a Metropolis-Hastings step based on Markovian proposals drawn from the uniformization algorithm [4]. Here, we extend their approach by considering a Bayesian nonparametric mixture model for both completely and discretely observed continuous time semi-Markov models. More specifically, by defining infinite mixtures of semi-Markov, with a Dirichlet process (DP) prior [5] on the mixing measure, we get a Dirichlet Process Mixture (DPM) [6] of semi-Markov models.

2 Semi-Markov multi-state models

Let us consider a continuous time process $Y(\cdot) = \{Y(t), t \geq 0\}$ with discrete state space $\mathscr{S} = \{1, \dots S\}$. We assume that the process $Y(\cdot)$ is semi-Markov. This is equivalent to say that the instantaneous transition rates $q_{rs}(t, \mathscr{F}_t)$, conditionally on the past history of the process, depend only on the time spent in the current state, i.e.

$$q_{rs}(t, \mathscr{F}_t) = \lim_{\delta t \to 0} \frac{P\{Y(t+\delta t) = s | X(t) = r, T^* = t-u\}}{\delta t}$$

where T^* denotes the entry time in the last state assumed before time t. Hence, semi-Markov models can be obtained by defining the transition functions $q_{rs}(u)$ and setting

$$P\{Y(t+\delta t) = s | Y(t) = r, T^* = t-u\} = \begin{cases} q_{rs}(u)\delta t + o(\delta t) & s \neq r \\ 1 - \sum_{l \neq r} q_{rl}(u)\delta t + o(\delta t) & s = r \end{cases}$$

Notice that a semi-Markov process $Y(t)$ can be also defined as the result of a state sequence generated by a Markov chain with transition probabilities p_{rs} and sojourn times having distribution functions F_{rs}, that is depending only on the departure and arrival states. To specify the functions $q_{rs}(u)$ we can also proceed directly by fixing the transition probabilities p_{rs} and the conditional sojourn distributions F_{rs}. By doing so, the resulting hazard functions turn out to be $q_{rs}(u) = p_{rs}F'_{rs}(u)/(1 - F_r(u))$ where $F_r(u) = \sum_{l \neq r} p_{rl}F_{rl}(u)$. Several parametric distributions can be proposed for $q_{rs}(u)$ or $F_{rs}(u)$. Assuming for example cause-specific hazards proportional to those of a distribution on positive values with parameters depending only on the initial

state, i.e. $q_{rs}(u) = p_{rs}q(u;\phi_r)$, the transition probabilities are p_{rs}, and the density of trajectory y on the interval $[0,T]$ can be generally written as

$$p_\theta(y) = p_\theta(s,z) = \left(\prod_{i=1}^{n} p_{s_{i-1}s_i} q(z_i - z_{i-1};\phi_{s_{i-1}}) e^{-\int_0^{z_i - z_{i-1}} q(u;\phi_{s_{i-1}},)du}\right) \times e^{-\int_0^{T-z_n} q(u;\phi_{s_n})du},$$

where $z = (z_1,\ldots,z_n)$ is the sequence of jump times, $s = (s_1,\ldots,s_n)$ is the sequence of visited states and $\theta \in \Theta$ is the vector with all the process parameters.

3 Dirichlet Process Mixture of semi-Markov models

In this section we introduce the notation of the DPM model. Let $p_\theta(y)$ be the probability density function of a semi-Markov trajectory $y(t) = (s,z)$. Let G be a probability distribution defined on the parameter space Θ. We define the density function of an infinite mixture of semi-Markov models p_G with respect to the mixing measure G as

$$p_G(s,z) = \int p_\theta(s,z)dG(\theta).$$

By assuming a $DP(M,G_0)$ on the mixing measure G, we get a DPM of semi-Markov models. Let $y_i(t) = (s,z)_i$, for $i = 1,\ldots,N$, be N fully observed paths on $[0,T_i]$. Note that N represents the number of sample individuals. We may rewrite the model in a hierarchical form:

$$\begin{aligned} y_i(t)|\theta_i &\overset{ind}{\sim} p_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned}$$

4 MCMC sampling

We now outline the MCMC algorithm. Let $\mathbf{y} = (\mathbf{s},\mathbf{z})$ be the set of N observed trajectories of a continuous time multi-state model defined on the state space $\mathscr{S}$. Let θ_h^*, $h = 1,\ldots,k$ denote the $k \leq N$ unique values among the N individual parameters and let $\Psi_h = (i : \theta_i = \theta_h^*)$. An important property of the Dirichlet process is in fact to produce discrete distributions for G so that the trajectories y_i are clustered on the base of the unique values generated by G. We represent the clustering by an equivalent set of cluster membership indicators, $\psi_i = h$ if $i \in \Psi_h$, i.e. the i-th observation belongs to the h-th cluster. Let $\mathbf{y}_h^*$ indicate the trajectories arranged by clusters, i.e. the set of trajectories belonging to the h-th cluster. Let the multiset $\rho_n = \{\Psi_1,\ldots,\Psi_k\}$ be a random partition of the trajectories $\{1,\ldots,N\}$. The DPM implies a model on

a random partition ρ_n of the experimental units. Let $\pi(\rho_n)$ be the Pólya urn prior on the random partition ρ_n. The posterior conditional densities $\pi(\psi_i = h|\psi_{-i}, \mathbf{y})$ are derived as follows:

$$\pi(\psi_i = h|\psi_{-i}, \mathbf{y}) \propto \begin{cases} N_h^- \, p((s,z)_i|\mathbf{y}_h^{*-}) & \text{for } h = 1,\dots,k^- \\ M p((s,z)_i) & \text{for } h = k^- + 1 \end{cases},$$

where N_h^- is the number of paths in the h-th cluster with exclusion of the i-th observation, M represents the precision parameter of the DP, k^- indicates the unique values of θ with exclusion of the i-th observation, $p((s,z)_i) \equiv \int p_{\theta_h^*}((s,z)_i) G_0(d\theta_h^*)$ and $p((s,z)_i|\mathbf{y}_h^{*-}) = \int p_{\theta_h^*}((s,z)_i) d\pi(\theta_h^*|\mathbf{y}_h^{*-})$. However, with the semi-Markov kernel density there is not conjugate centring measure G_0. Therefore, approximation methods are required in order to evaluate the integrals.
Since there is no conjugacy between the base measure G_0 and the semi-Markov kernel density, we adopt the approach of [7], by defining a valid Markov chain algorithm by drawing $(N-k)$ auxiliary values for θ from the centring measure G_0. More in detail, if $N_h > 1$, the posterior model on the random partition ρ_n is defined as:

$$\pi(\psi_i = h|\psi_{-i}, \mathbf{y}) \propto \begin{cases} N_h^- \, p_{\theta_h^*}((s,z)_i) & \text{for } h = 1,\dots,k^- \\ \frac{M}{k^-+1} p_{\theta_{k^-+1}^*}((s,z)_i) & \text{for } h = k^- + 1 \end{cases}, \tag{1}$$

while if $N_h = 1$, $\psi_i = h$ is imputed to form a singleton cluster. Therefore, with probability $(k-1)/k$ leave ψ_i unchanged, otherwise remove ψ_i from the h-th cluster, relabel the θ_h^* to comply with the no-gaps rule, and then update ψ_i. By considering $y_i = (s,z)_i$, $i = 1,\dots,N_h$, the density of the data lying inside the h-th mixture component conditional on the partition and on the kernel parameters is

$$\pi(\mathbf{y}|\theta_h, \psi) = \prod_{i=1}^{N_h} p_{\theta_h}((s,z)_i|\psi_i = h),$$

by Bayes theorem we get the posterior density of the h-th mixture component parameters:

$$\pi(\theta_h|\psi, \mathbf{y}) \propto \pi(\mathbf{y}|\theta_h, \psi) \cdot G_0(\theta_h). \tag{2}$$

4.1 Discretely observed semi-Markov processes

We now extend the DPM of semi-Markov models to the the case of discretely observed trajectories, that is when the exact jump times are unknown and the density function $p_\theta(y)$ is not available. In particular we use the algorithm proposed by [3], which allows to reconstruct the trajectories between the discretely observed points for each observed individual via a Metropolis-Hastings step based on a Markovian approximation of the semi-Markov process. By indicating with $Q(s,z|\theta,x)$ the conditional distribution of a whole sample path conditionally on an observed set of

points summarized by the vector x, the MCMC algorithm for the discretely observed case may be summarized as follows:

- *Trajectory reconstruction*: for $i = 1\ldots,N$ draw $(s,z)_i \sim Q(s,z|\theta_i,x_i)$;
- *Clustering*: for $i = 1\ldots,N$ draw ψ_i from (1);
- *Updating cluster parameters*: for $h = 1,\ldots,k$ draw θ_h from (2).

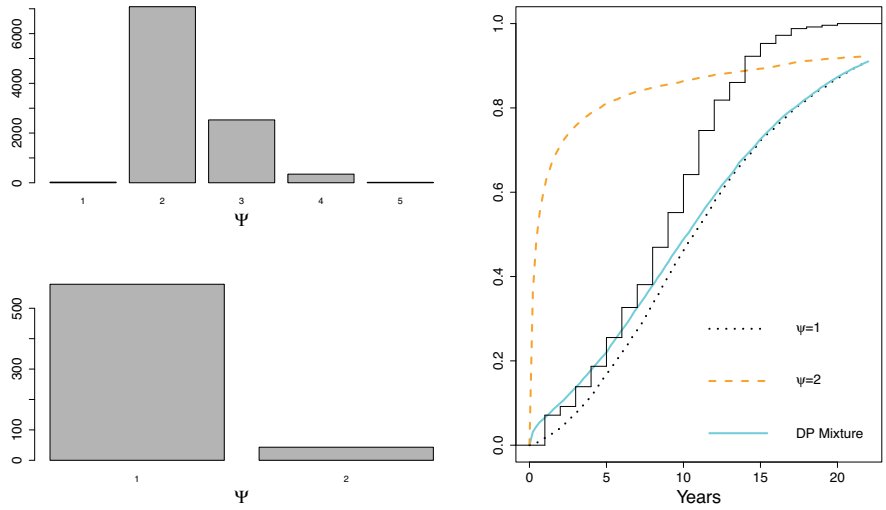
5 Application

As a real data application, we analyze the progression of coronary allograft vasculopathy (CAV) [8] with a data set available with the R package `msm`, see [9]. The data provides the disease status (CAV-free (1), mild CAV (2) and moderate or severe CAV (3)) observed approximately each year after transplant for a set of 622 subjects followed up until their most recent visit if alive at the end of the observation period or until death (state (4)). Death times are exactly observed. Moreover, since data comprises apparent transitions from higher to lower states, which are in fact the results of a misclassification since the deterioration of the arterial walls is an irreversible process, we recode all the reverse transitions as remaining in the higher of the two states and permit transitions only to the adjacent states or to the death state. To specify the semi-Markov model we assume Weibull sojourn times. The parameter set is $\theta = (p,\gamma,\alpha)$ where γ and α are the vector with the rate and shape parameters of the Weibull sojourn times and p is the matrix with the transition probabilities. For the Dirichlet process we chose a precision parameter $M = 1$ and defined the centering measure to be the product between Dirichlet distributions for the rows of p, Gamma distributions for the rate parameters and log Normal distributions for the sahpe parameters, that is $G_0 \equiv \mathrm{Dir}_{\mathscr{S}-1} \times \mathrm{Gamma}_{\mathscr{S}-1} \times \log\mathrm{N}_{\mathscr{S}-1}$. We run the MCMC sampler for 10000 iterations with a burnin of 2000. The model gives evidence of the existence of two groups of observations with different features inside the sample. In Table 1 and Figure 1 we show the results by reporting some posterior summaries for the proposed model.

Table 1 CAV data: DPM of semi-Markov.

	ψ	γ_1	γ_2	γ_3	α_1	α_2	α_3	p_{12}	p_{14}	p_{23}	p_{24}
$E(\cdot\|Y,\psi)$	1	0.13	0.24	0.25	1,47	1.46	1.10	0.71	0.29	0.72	0.28
$SD(\cdot\|Y,\psi)$	1	0.01	0.02	0.03	0.10	0.18	0.12	0.04	0.04	0.08	0.08
$q_{0.025}(\cdot\|Y,\psi)$	1	0.12	0.19	0.20	1.30	1.16	0.88	0.64	0.21	0.56	0.12
$q_{0.975}(\cdot\|Y,\psi)$	1	0.14	0.29	0.30	1.65	1.85	1.34	0.79	0.36	0.88	0.44
$E(\cdot\|Y,\psi)$	2	6.27	1.42	1.30	0.66	0.89	4.45	0.71	0.29	0.16	0.84
$SD(\cdot\|Y,\psi)$	2	3.62	1.45	1.01	0.19	1.32	4.45	0.36	0.36	0.23	0.23
$q_{0.025}(\cdot\|Y,\psi)$	2	0.16	0.02	0.08	0.39	0.17	0.37	0.02	0.00	0.00	0.13
$q_{0.975}(\cdot\|Y,\psi)$	2	13.08	4.30	3.81	1.10	4.95	15.85	1.00	0.98	0.87	1.00

Fig. 1 CAV data: on the top left panel the maximum number of mixture components observed for each iteration and on the bottom left the distribution of the observations across the estimated components; on the right panel the cumulative posterior predictive distributions of the death time.



References

1. Luo, Yu, David A. Stephens, and David L. Buckeridge. Bayesian clustering for continuous-time hidden Markov models. *Canadian Journal of Statistics*, 2021.
2. Sylvia Frühwirth-Schnatter and Christoph Pamminger. Model-based clustering of categorical time series. *Bayesian Analysis*, 2010
3. Rosario Barone and Andrea Tancredi. Bayesian inference for discretely observed continuous time multi-state models. *preprint arXiv: 2202.12447*, 2022.
4. Asger Hobolth and Eric A Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204–1231, 2009.
5. Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.
6. Albert Y Lo. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357, 1984.
7. Steven N MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.
8. Linda D Sharples, Christopher H Jackson, Jayan Parameshwar, John Wallwork, and Stephen R Large. Diagnostic accuracy of coronary angiography and risk factors for post–heart-transplant cardiac allograft vasculopathy. *Transplantation*, 76(4):679–682, 2003.
9. Christopher H Jackson. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38(8):1–29, 2011.

Specification of informative priors for capture-recapture finite mixture models

Specificazione di a-priori informative per modelli cattura-ricattura a misture finite

Pierfrancesco Alaimo Di Loro[1], Gianmarco Caruso[2], Marco Mingione[2], Giovanna Jona Lasinio[2], Luca Tardella[2]

Abstract Many models involve binomial terms depending on parameters $p \in [0,1]$ representing probabilities. That is the case of capture-recapture experiments, where capture and survival of each individual at different occasions are modelled as Bernoulli trials with unknown probabilities. In most actual data applications, the population of interest typically exhibits unaccounted heterogeneity, presumably depending on its partitioning into a finite set of sub-populations, each one having its parameter value. If the sub-population labels are unknown, Finite Mixture Models (FMM) can be exploited to recover the unknown labels and all other model components jointly. Nevertheless, the naive application of finite mixture models within the Bayesian machinery is affected by the so-called *label-switching* problem. The group-specific parameters are assigned ordering constraints to identify their relative roles to overcome this issue. That is usually achieved by specifying conditionally uniform densities that respect such constraints, preventing the possibility to shape the prior according to available prior knowledge. In this work, we propose two flexible classes of joint priors based on manipulating Beta distributions. The idea is to specify a joint prior that retains the flexibility to induce the desired marginal behaviour while still guaranteeing the desired ordering.

Abstract *Diversi modelli includono componenti Binomiali dipendenti da parametri $p \in [0,1]$, che rappresentano probabilità. Questo è il caso di modelli cattura-ricattura, dove la cattura e la sopravvivenza di ciascun individuo in ciascuna occasione sono visti come esperimenti Bernoulliani con probabilità incognite. Nella maggior parte delle applicazioni reali, la popolazione di interesse presenta ulteriore eterogeneità, presumibilmente dovuta al partizionamento della popolazione in sottopopolazioni, ciascuna avente il proprio valore del parametro. Se le etichette della sottopopolazione sono incognite, i modelli a mistura finita possono essere usati per stimare congiuntamente le etichette e le altre componenti del modello. Tuttavia, la semplice applicazione di questi modelli in un contesto Bayesiano soffre del cosiddetto problema di label-switching. Per risolvere questo problema, i parametri gruppo-specifici vengono sottoposti a vincoli di ordinamento. Questo si ottiene tipicamente attraverso la specificazione di distribuzioni Uniformi condizionate che rispettano l'ordinamento, ma che non permettono l'elicitazione di distribuzioni a priori basate su un'eventuale conoscenza pregressa del fenomeno. In questo lavoro, si propongono due classi flessibili di distribuzioni a priori congiunte, basate sulla distribuzione Beta. L'idea è quella di specificare una a priori congiunta che mantenga una flessibilità tale da indurre marginalmente il comportamento desiderato, garantendo allo stesso tempo l'ordinamento.*

Key words: Finite mixture, Bayesian Statistics, Prior elicitation, Capture-recapture, Binomial model

1 Introduction

The implementation of capture-recapture methods in a Bayesian context often involves the specification of Beta distributions to model the prior beliefs on the capture and survival probabilities of the individuals.

[1]LUMSA, Via Pompeo Magno 28, Rome, Italy, e-mail: p.alaimodiloro@lumsa.it
[2]University of Rome "La Sapienza", P.le Aldo Moro 5, Rome, Italy.

The Beta distribution is a very common candidate as a prior for such parameters p: it is compactly supported on $[0,1]$. It can be shaped to represent non-informative or informative settings by manipulating its two shape parameters. However, real data often exhibit additional heterogeneity concerning standard modelling assumptions, as not all the individuals in the population behave in the same way. For instance, the population may be partitioned into a finite set of G sub-populations, or groups, (e.g. by gender, residency pattern, size, etc.), each one having its own parameter value, say p_g, $g = 1,\dots,G$ [McLachlan et al., 2019]. When the group labels are unknown, the partition must be estimated to get reliable inferences. In many applied sciences, the partition is often recovered in advance and subsequently passed to the model as an input [Pace et al., 2021]. Nevertheless, it is usually preferable to perform the classification and the estimation simultaneously to quantify the uncertainty of both better. In this respect, capture-recapture models can be embedded in a Finite Mixture (FMM) setting [Pledger, 2000, 2005; Böhning et al., 2005]. Each unit is uniquely assigned a latent label representing its membership to one of the G possible groups. The posterior distribution of the unknown labels is recovered jointly with all other model components, properly propagating the uncertainty of all the unknowns.

The naive application of finite mixture models within the Bayesian machinery is affected by the so-called *label-switching* problem. That is, at each Markov Chain Monte Carlo (MCMC) step, the groups may interchange their relative role [Jasra et al., 2005]. Under genuine multimodality, an effective and trivial solution is to uniquely identify the components by including prior information about their marginal and relative behaviour, as by imposing ordering constraints that define the group-specific parameters' positions in a hierarchy [Diebolt and Robert, 1994; Stephens, 2000]. When the parameters are constrained in the $[0,1]$ set, the corresponding joint prior could be derived as the product of conditionally specified uniform densities that respect such constraints. It inevitably jeopardizes the possibility of eliciting informative priors whenever the information is available. In this work, we propose two flexible classes of joint priors for this scenario, based on the manipulations of the Beta distribution. The idea is to specify a joint prior that guarantees the desired ordering constraints and retains the flexibility to include prior information in each component's marginal.

2 The modeling framework

For the sake of brevity, we here report the modelling framework, in which we included our proposals when the number of sub-populations is $G = 2$. Remark that this can be easily generalized to $G > 2$ cases simply by re-iterating the conditioning process.

Let $\boldsymbol{y}_1,\dots,\boldsymbol{y}_n$ be n realizations of an l-dimensional random vector with distribution $\boldsymbol{Y} \sim f(\cdot\,|\,\boldsymbol{p})$, where $\boldsymbol{p} = (p_1, p_2)$ and

$$f(\boldsymbol{y}\,|\,\boldsymbol{p}) = \sum_{k=1}^{2} w_k \cdot f(\boldsymbol{y}\,|\,p_k).$$

This is a finite mixture distribution with 2 components and prior weights w_k, $k = 1,2$. This setting can be more efficiently represented by augmenting the space with a latent indicator variable $\zeta \in \{1,2\}$ that denotes from which of the two densities each observation come from. The component weights w_k are the prior probabilities that each unit belongs to one group or the other $P(\zeta_i = k) = w_k$, $i = 1,\dots,n$, $k = 1,2,$. It represents the relative proportion of observations coming from the two groups. Bayesian inference is usually pursued through the following hierarchical specification:

$$\begin{aligned}
&\boldsymbol{Y}\,|\,\boldsymbol{p},\zeta,\boldsymbol{w} \sim f(\cdot\,|\,p_\zeta)\\
&\zeta\,|\,\boldsymbol{w} \sim MN_2(\boldsymbol{w}), \quad \boldsymbol{p} \sim \pi_{\boldsymbol{p}}(\cdot)\\
&\boldsymbol{w} \sim Dir(\boldsymbol{\alpha}),
\end{aligned}$$

where $MN_2(\cdot)$ is the two-components multinomial distribution, $Dir(\cdot)$ the Dirichlet distribution and $\pi_{\boldsymbol{p}}(\cdot)$ is the joint prior on the group specific probabilities p_1, p_2. In order to avoid the label-switching problem, the prior on $\boldsymbol{p}$ shall envision an ordering constraint to solve the symmetry in the posterior distribution and make the two components identifiable. For

instance, we could enforce $p_1 < p_2$ by specifying the joint prior in a conditional fashion:

$$\pi_{\boldsymbol{p}}(p_1, p_2) = \pi_{p_1}(p_1) \cdot \pi_{p_2|p_1}(p_2), \tag{1}$$

where $\pi_{p_2|p_1}(\cdot)$ guarantees $p_2 > p_1$, e.g. has a support depending on p_1. When p_1 and p_2 are probabilities (i.e. $\in (0,1)$), the most striaghtforward solution is to have a Beta prior on p_1 and then specify $\pi_{p_2|p_1}(\cdot) = Unif(\cdot \,|\, p_1, 1)$. In practice, any other specification of $\pi_{p_2|p_1}(\cdot)$ complying with the ordering constraint could be valid. Nevertheless, there is very few literature that explores the marginal distribution induced on $p_2 \sim \pi_{p_2}(\cdot)$ obtained by integrating p_1 out from Eq. 1.

In Sec. 3 and Sec. 4, we investigate some alternative specifications that allow for controlling the marginal expected value and variance, and possibly also the form of the two marginals. These tools can be exploited to embed prior information in the estimation process, which can, in turn, favour the proper identification of the two components.

3 The Beta and truncated Beta

It is possible to derive the marginal prior distribution of p_2, when $p_1 \sim Beta(\alpha_1, \beta_1)$ and $p_2|p_1 \sim tBeta(\alpha_2, \beta_2, p_1, 1)$, where $tBeta(\cdot,\cdot,l,u)$ denotes the *truncated Beta* in (l,u). However, the expression of the marginal $\pi_{p_2}(\cdot)$ is not trivial unless we set $\alpha_2 = 1$, for which we obtain:

$$\pi_{p_2}(p_2) = \frac{B(\alpha_1, \beta_1 - \beta_2)}{B(\alpha_1, \beta_1)} \beta_2 (1-p_2)^{\beta_2 - 1} F_{Beta(\alpha_1, \beta_1 - \beta_2)}(y), \quad \text{with } \beta_1 > \beta_2.$$

In particular, when $\alpha_1 = \beta_2 = k$ and $\beta_1 = k+1$, i.e. $p_1 \sim Beta(k, k+1)$ and $p_2|p_1 \sim tBeta(1, k, p_1, 1)$, then we obtain the following convenient result for the marginal:

$$p_2 \sim Beta(k+1, k), \quad k > 0. \tag{2}$$

Notice that, in this particular case, the marginal distribution induced on p_2 is symmetrical with respect to the distribution of p_1 around the vertical line $p^* = 0.5$.

4 The Beta and restricted Beta

The *restricted Beta* (also called *4-parameters Beta*) is a Beta random variable which has been shifted and scaled to lie on a different domain (l,u).

In other words, if $X \sim Beta(\alpha, \beta)$ and $Z = l + X \cdot (u-l)$, then $Z \sim rBeta(\alpha, \beta, l, u)$. The expected value and variance of Z are:

$$\mathbb{E}[Z] = \frac{\alpha}{\alpha+\beta} \cdot (u-l) + l = \frac{u\alpha + l\beta}{\alpha+\beta}, \qquad \mathbb{V}[Z] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \cdot (u-l)^2$$

The restricted Beta can be exploited to specify a compelling joint prior for p_1 and p_2. If $p_1 \sim Beta(\alpha_1, \beta_1)$ and $p_2 \,|\, p_1 \sim rBeta(\alpha_2, \beta_2, p_1, 1)$ then the corresponding marginal expected value and variance of p_2 can be derived as a function of mean and variance of p_1 using the *Tower Law*:

$$\begin{aligned} \mathbb{E}[p_2] &= \frac{\alpha_2}{\alpha_2+\beta_2} + \mu_1 \frac{\beta_2}{\alpha_2+\beta_2} \\ \mathbb{V}[p_2] &= \sigma_1^2 \cdot \frac{\beta_2^2}{(\alpha_2+\beta_2)^2} \left(1 + \frac{\alpha_2}{\beta_2(\alpha_2+\beta_2+1)}\right) + (1-\mu_1)^2 \cdot \frac{\alpha_2\beta_2}{(\alpha_2+\beta_2)^2(\alpha_2+\beta_2+1)} \end{aligned} \tag{3}$$

where $\mu_1 = \mathbb{E}[p_1]$ and $\sigma_1^2 = \mathbb{V}[p_1]$. Thus, once chosen α_1 and β_1 to comply with some prior information on μ_1 and σ_1^2, we can elicit the values of α_2 and β_2 to respect prior knowledge on μ_2 and σ_2^2 simply by solving a linear system based on Eq. (3).

5 Application in Capture-Recapture analysis

Here, we illustrate the component-specific parameters' prior specification in the context of Capture-Recapture methods.

Capture-Recapture methods are widely employed in estimating the size of elusive populations whose units are subject to multiple captures across several occasions. Here, we assume that the population of interest is *closed* and individuals belonging to the population are *captured independently* one another [Otis et al., 1978]. When this is the case, even when the sampling occurred at J on different occasions, the overall capture frequencies of each individual are sufficient to make inference on the unknown population size N.

Let y_i be the capture frequency of the i-th unit across the J occasions, for $i = 1,\dots,N$, and let D be the total number of distinct *observed* individuals (i.e. captured at least once). The unknown populations size N can range from D to infinity. Following Royle and Dorazio [2012], we can bound the parameter space of N with a large (at will) upper bound M and augment the data sample with null capture frequencies $(y_{D+1},\dots,y_M) = (0,\dots,0)$ (representing fake individuals) in order to exploit classical MCMC sampling in a Bayesian framework. This is achieved by introducing a collection of i.i.d. latent variables $z_1,\dots,z_M$ that indicate whether the individual i belongs ($z_i = 1$) or not ($z_i = 0$) to the true population of N individuals. It is straightforward to see that $N = \sum_{i=1}^{M} z_i$. The advantage of this kind of modelling lies in the possibility to perform Bayesian inference without the need to jump between parameter spaces of varying dimensions, and this avoids the implementation of application-specific *Reversible Jump MCMC* [King and Brooks, 2002].

In particular, a uniform prior distribution in the discrete set $\{0,\dots,M\}$ is induced on N if we consider the hierarchical specification $z_i \,|\, \psi \sim Bern(\psi)$, $i = 1,\dots,M$ and $\psi \sim Unif[0,1]$, where ψ is the probability that an individual of the augmented dataset of size M is a member of the true population.

With the purpose of modelling the capture heterogeneity between individuals, this framework can be embedded in a FMM setting by considering finite mixture of two binomial distributions [Pledger, 2000] for the counts y_i

$$y_i|z_i \sim w \cdot Bin(J, p_1 \cdot z_i) + (1-w) \cdot Bin(J, p_2 \cdot z_i),$$

where w is the probability that individual i belongs to the first mixture component and p_g ($g = 1,2$) is the capture probability of the g-th component. Notice that $y_i = 0$ almost surely when $z_i = 0$ so that the previous model corresponds to a finite mixture of zero-inflated binomial distributions.

5.1 Simulation experiment

This section shows comparative merits of alternative prior specifications in dealing with label switching and exploiting prior information on group-specific parameters. We simulated $k = 1,\dots,50$ alternative capture-recapture datasets from closed populations with 2 groups. The simulation scheme mimics the model described in Sec. 5. Captures have been organized in $J = 10$ different occasions on a *super-population* of size $M = 500$, with a probability to be included in the actual population of $\psi = 0.3$, for all datasets. This yields an expected value of $\mathbb{E}[N_k] = 150$ across all datasets. Individuals have been allocated to the two groups evenly, setting $w = 0.5$. The group-specific capture probabilities have been set to $p_1 = 0.2$ and $p_2 = 0.4$. Under the closure hypothesis, the capture histories of each dataset have been collapsed into the vector of individual overall capture frequencies $\boldsymbol{y}_k = \left[y_1,\dots,y_{D_k}\right]$, where D_k is the number of distinct observed individuals of the simulated set k. The following five priors on the group-specific capture probabilities are considered, which correspond to those represented in the upper panels of Fig. 1:

A) $p_1 \sim Unif[0,1]$ and $p_2 \sim Unif[0,1]$;
B) $p_1 \sim Beta(1.08, 4.32)$ and $p_2 \sim Beta(3.44, 5.16)$, with hyperparameters centered on the true values $\mathbb{E}[p_1] = 0.2$ and $\mathbb{E}[p_2] = 0.4$ with $\mathbb{V}[p_1] = 0.2/10$ and $\mathbb{V}[p_2] = 0.4/10$;
C) $p_1 \sim Unif[0,1]$ and $p_2|p_1 \sim Unif[p_1,1]$, naive constrained prior inducing an improper marginal prior on p_2;

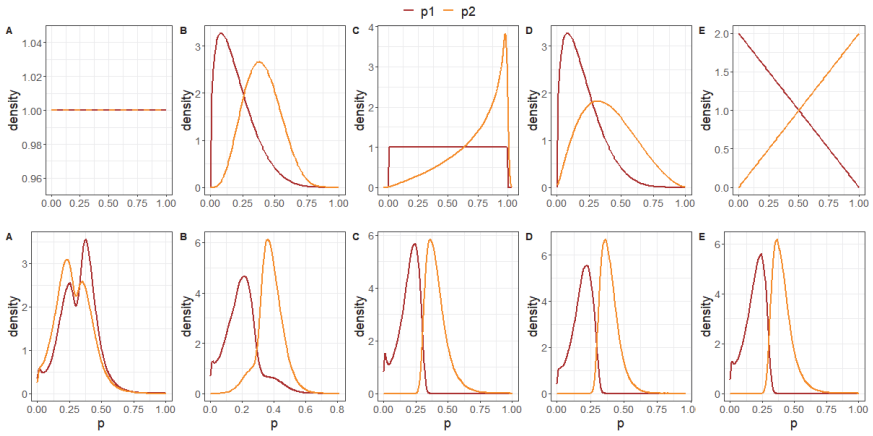


Fig. 1: Comparison between the marginal prior (*upper panels*) and posterior (*lower panels*) distributions for the parameters p_1 and p_2 for the five considered settings (A-E).

D) $p_1 \sim Beta(1.4, 5.6)$ and $p_2|p_1 \sim rBeta(0.826, 2.478, p_1, 1)$, inducing a marginal prior distributions on p_2 with the same means and variances of those considered in setting B;
E) $p_1 \sim Beta(1,2)$ and $p_2|p_1 \sim tBeta(1,1,p_1,1)$ which induces a marginal prior $p_2 \sim Beta(2,1)$ favoring repulsion between the two components.

The model in Sec. 5 is then coded and estimated in JAGS [Plummer et al., 2003] separately on each dataset and for each prior specification. Posterior samples of the capture probabilities p_1 and p_2 corresponding to different datasets have been merged (by prior setting) to analyze their overall behavior. The lower panels in Fig. 1 show their distribution, highlighting how they seem to be correctly identified whenever an ordering constraint between p_1 and p_2 is present (cfr. C-E). On the other hand, when independent priors (cfr. A-B) are considered, the posterior samples are affected by the label switching issue. Notably, regardless of the prior specifications, posterior distributions on p_1 and p_2 tend to be similar in settings C to E. Finally, the differences between the estimated posterior mean $\hat{N}_k$ and its true value N_k have been evaluated for each dataset $k = 1, \ldots, 50$. The overall performances of the different prior setting have been compared in terms of *average Predictive Interval Width* (PIW), *Root Mean Squared Error* (RMSE) and *Deviance Information Criterion* (DIC). Results are reported in Table 1, which highlights how the *constrained informative* prior (i.e. setting D) provides the best performances in all indicators, shortly followed by Setting E and A. Notably, including prior information without introducing constraints to avoid the label switching issue does not yield any improvement (on the contrary it worsens the results).

	A	B	C	D	E
PIW	46	49	51	41	43
RMSE	6.7	7.1	7.5	5.9	6.3
DIC	785.6	787.8	787.7	776.9	787.8

Table 1: 95% Prediction interval width (PIW), Root Mean Squared Error (RMSE) for N, and DIC for the five considered settings (A-E).

6 Final discussion

In light of the above, we can state that constrained priors on capture probability parameters are effective in avoiding the label switching issue, as shown in Sec. 5.1. In particular, if prior information is elicited correctly, it can improve the estimation and predictive performances of the model under the condition that constraints to avoid label switching are present (setting D).

Such a proposal's effectiveness must be tested on a real data example for which the finite mixture assumption is suitable. The simulation study could then be extended to a number of groups $G > 2$ in order verify the robustness of the procedure. Eventually, the whole prior setting could be extended to the more general case where the population is open, including also entrance and survival probabilities that my be assumed to be group-specific or not.

References

D. Böhning, E. Dietz, R. Kuhnert, and D. Schön. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14(1):29–43, 2005.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2): 363–375, 1994.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

R. King and S. Brooks. Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika*, 89(4):785–806, 2002.

G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135, 1978.

D. S. Pace, C. Di Marco, G. Giacomini, S. Ferri, M. Silvestri, E. Papale, E. Casoli, D. Ventura, M. Mingione, P. Alaimo Di Loro, et al. Capitoline dolphins: residency patterns and abundance estimate of tursiops truncatus at the tiber river estuary (mediterranean sea). *Biology*, 10(4):275, 2021.

S. Pledger. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, 56(2):434–442, 2000.

S. Pledger. The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics*, 61(3):868–873, 2005.

M. Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

J. A. Royle and R. M. Dorazio. Parameter-expanded data augmentation for bayesian analysis of capture–recapture models. *Journal of Ornithology*, 152(2):521–537, 2012.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Clustering multivariate categorical data: a graphical model-based approach

Clustering di dati categoriali multivariati: un approccio basato su modelli grafici

Francesco Rettore, Michele Russo, Luca Zerman, Federico Castelletti

Abstract Clustering multivariate data using mixture models is a well-studied topic in statistics. In this contribution we propose a Bayesian framework for clustering categorical data which makes use of graphical models to account for possible group-specific dependence relations between variables. Our mixture model formulation allows to simultaneously infer a clustering structure of the units and the network of dependencies between variables.

Abstract *Il clustering di dati multivariati attraverso modelli mistura è un problema ampiamente studiato nella letteratura statistica. In questo contributo proponiamo un modello bayesiano per il clustering di dati categoriali che attraverso l'impiego di modelli grafici consente di considerare relazioni di dipendenza tra variabili che siano specifiche di ciascun gruppo. Il modello mistura proposto permette quindi di individuare gruppi latenti di osservazioni e al tempo stesso di stimare la struttura di dipendenza tra variabili del sistema.*

Key words: Model-based clustering, Categorical data, Graphical model

Francesco Rettore
Politecnico di Milano, e-mail: francesco.rettore@mail.polimi.it

Michele Russo
Politecnico di Milano, e-mail: michele6.russo@mail.polimi.it

Luca Zerman
Politecnico di Milano, e-mail: luca.zerman@mail.polimi.it

Federico Castelletti
Università Cattolica del Sacro Cuore, e-mail: federico.castelletti@unicatt.it

1 Introduction

Clustering individuals is a pervasive issue in statistics, with applications in several domains, and primarily social sciences. Traditionally, the two main clustering frameworks are represented by *distance-based* techniques that are opposed to *model-based* methods. The former implement a suitable measure of distance between observations and assign units to the same cluster whenever these are "close" in terms of the adopted metric; in this setting the most popular heuristic method is represented by the original k-means algorithm of [2] together with several of its extensions. Differently, model-based clustering techniques are probabilistic models based on *mixtures*, where each mixture component corresponds to a given cluster; see for instance [3].

In this work we assume that *multivariate categorical* observations are available. Categorical data are widely diffuse in many fields, and in particular social studies, where the information on individuals' preferences is collected through surveys; these consist of a battery of items, most of which are typically qualitative. In addition, the so-obtained data can be widely heterogeneous, as subjects differ by individual characteristics such as age and gender, among others. Most importantly, these features are expected to reflect the very way each given subject responds to the survey's items. When groups are known beforehand, one can perform independent analyses on each sub-dataset. However, this choice may produce *unreliable* inferential results due to the small group-sample sizes. On the other hand, modelling the whole dataset as a unique single group can produce *biased* results since it does not account for possible heterogeneity in the sample. In addition, finding clusters of individuals sharing similar features in terms of preferences w.r.t. the items can be of interest in itself, for instance to develop personalized marketing strategies.

In this contribution we propose a model-based framework for clustering categorical observations. Our method fully accounts for dependence relations between variables/items that are encoded in the available data. Specifically, we represent conditional independence relations by using a graphical model-based approach [1]. Graphical models are probabilistic models for a collection of random variables which provide a powerful tool to impose dependence relations between variables in the joint density through a graph structure $\mathscr{G}$. The latter is made up of a set of *nodes* V, each corresponding to a variable in the system, and a set of *edges* E, representing dependence relations between nodes. We consider a Bayesian *non-parametric mixture model* based on a Dirichlet Process prior [7], where each component of the mixture corresponds to a suitable categorical graphical model. Our model formulation allows to simultaneously learn dependence relations between variables and infer a clustering structure among individuals represented by latent groups of units sharing the same set of dependence statements.

The rest of this contribution is organized as follows. In Section 2 we introduce graphical models for categorical data under a Bayesian framework. In Section 3 we describe our mixture of categorical graphical models, while in Section 4 we provide computational details relative to a Markov chain Monte Carlo (MCMC) scheme for posterior inference on clustering and graphs.

2 Categorical graphical models

Let $(X_1,\ldots,X_q)$ be a categorical random vector such that for each $j=1\ldots,q$, $X_j \in \mathscr{X}_j$, the set of levels of the categorical variable X_j. Let also $\mathbb{X}$ be an (n,q) dataset whose rows are i.i.d. realizations of the random vector $(X_1,\ldots,X_q)$, namely $\mathbf{x}^{(i)} = \left(x_1^{(i)},\ldots,x_q^{(i)}\right)^\top$, for $i=1,\ldots n$. It follows that each $\mathbf{x}^{(i)} \in \mathscr{X} := \bigotimes_{j=1}^q \mathscr{X}_j$, the product space generated by the levels of the q categorical variables. We then let $\pi(\mathbf{x}) \in (0,1)$ be the probability to observe a given configuration $\mathbf{x} \in \mathscr{X}$. Starting from the data matrix $\mathbb{X}$, we can compute for each $\mathbf{x} \in \mathscr{X}$ the corresponding *configuration count* $n(\mathbf{x})$ defined as

$$n(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{1}\{\mathbf{x}^{(i)} = \mathbf{x}\}. \tag{1}$$

In addition, for a given subset $S \subseteq \{1,\ldots,q\}$ and each $\mathbf{x}_S \in \mathscr{X}_S$, the corresponding *marginal configuration count* can be computed as

$$n(\mathbf{x}_S) = \sum_{i=1}^{n} \mathbf{1}\big\{\mathbf{x}^{(i)}(S) = \mathbf{x}_S\big\} = \sum_{\mathbf{x}\in\mathscr{X}} n(\mathbf{x})\mathbf{1}\big\{\mathbf{x}(S) = \mathbf{x}_S\big\}, \tag{2}$$

where $\mathbf{x}(S)$ is the sub-vector of $\mathbf{x}$ with components indexed by S. Similarly, we can define the *marginal probability*

$$\pi(\mathbf{x}_S) = \sum_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})\mathbf{1}\big\{\mathbf{x}(S) = \mathbf{x}_S\big\}. \tag{3}$$

By assuming independence among the n observations, we can write the likelihood function as

$$\begin{aligned} p(\mathbf{N} \mid \theta) = \prod_{i=1}^{n}\prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{\mathbf{1}\{\mathbf{x}^{(i)}=\mathbf{x}\}} &= \prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{\sum_{i=1}^n \mathbf{1}\{\mathbf{x}^{(i)}=\mathbf{x}\}} \\ &= \prod_{\mathbf{x}\in\mathscr{X}} \pi(\mathbf{x})^{n(\mathbf{x})}, \end{aligned} \tag{4}$$

where θ is the model parameter collecting the probabilities $\{\pi(\mathbf{x}), \mathbf{x}\in\mathscr{X}\}$, and $\mathbf{N}$ denotes the collection of counts (data) $\{n(\mathbf{x}), \mathbf{x}\in\mathscr{X}\}$.

Let now $\mathscr{G} = (V,E)$ be a decomposable undirected graph (UG) on the set of nodes $V = \{1,\ldots,q\}$. A decomposable UG is uniquely characterized by its set of *cliques* and *separators*; see for instance [1]. More importantly, as we associate each categorical variable X_j to a node in $\mathscr{G}$, the conditional independencies encoded in $\mathscr{G}$ are imposed to the joint density of $(X_1,\ldots,X_q)$ and the likelihood function factorizes as

$$p(\mathbf{N} \mid \theta, \mathscr{G}) = \frac{\prod_{C\in\mathscr{C}} p(\mathbf{N}_C \mid \theta_C)}{\prod_{S\in\mathscr{S}} p(\mathbf{N}_S \mid \theta_S)}, \tag{5}$$

where $\mathscr{C}$ and $\mathscr{S}$ denote, respectively, the sets of cliques and separators of $\mathscr{G}$. Also, $\mathbf{N}_C$ and θ_C represent the contingency tables of marginal configuration counts and probabilities, $\{n(\mathbf{x}_C), \mathbf{x}_C \in \mathscr{X}_C\}$ and $\{\pi(\mathbf{x}_C), \mathbf{x}_C \in \mathscr{X}_C\}$ respectively; see also [4].

We complete the model specification by assigning a prior to θ. Specifically, we assume that conditionally on $\mathscr{G}$, θ follows a *Hyper Dirichlet* (HD) distribution with hyperparameter $\mathbf{A} = \{a(\mathbf{x}), \mathbf{x} \in \mathscr{X}\}$, namely

$$p(\theta \mid \mathscr{G}) = \frac{\prod_{C \in \mathscr{C}} p(\theta_C \mid \mathscr{G})}{\prod_{S \in \mathscr{S}} p(\theta_S \mid \mathscr{G})}, \tag{6}$$

where for each $S \in \mathscr{S}$, $\theta_S \mid \mathbf{A}_S \sim \mathrm{Dir}(\mathbf{A}_S)$, with $\mathbf{A}_S = \{a(\mathbf{x}_S), \mathbf{x}_S \in \mathscr{X}_S\}$ and

$$a(\mathbf{x}_S) = \sum_{\mathbf{x} \in \mathscr{X}} a(\mathbf{x}) \mathbf{1}\{\mathbf{x}(S) = \mathbf{x}_S\}; \tag{7}$$

similarly for each $C \in \mathscr{C}$; see [4] for full details. Finally, a prior $p(\mathscr{G})$, for any $\mathscr{G} \in \mathbb{S}_q$, the set of all decomposable UGs on q nodes, can be assigned through a Beta-Binomial distribution on the number of edges in the graph; see for instance [6].

The HD distribution provides a conjugate prior for the model parameter θ; accordingly, the posterior distribution of θ is still HD and the *marginal likelihood* of $\mathscr{G}$, $m(\mathbf{N} \mid \mathscr{G}) = \int p(\mathbf{N} \mid \theta, \mathscr{G}) p(\theta \mid \mathscr{G}) \, d\theta$, is available in closed form. The latter feature is essential for the implementation of the MCMC scheme for posterior inference on graph structures and clustering introduced in Section 4.

3 Mixture model formulation

We extend the previous framework to a *mixture* of categorical graphical models, thus allowing for possible heterogeneous dependence relations among subjects that are linked to a latent clustering structure of the data. We base our model formulation on a *Dirichlet Process* (DP) prior [7], by assuming

$$\begin{aligned} \mathbf{x}^{(i)} \mid (\theta_i, \mathscr{G}_i) &\sim p\big(\mathbf{x}^{(i)} \mid \theta_i, \mathscr{G}_i\big), \\ (\theta_i, \mathscr{G}_i) \mid M &\sim M, \\ M &\sim DP(M_0, \alpha), \end{aligned} \tag{8}$$

where in particular $DP(M_0, \alpha)$ represents the Dirichlet Process with base distribution M_0 and concentration parameter α. We then take $p(\theta, \mathscr{G}) = p(\theta \mid \mathscr{G}) p(\mathscr{G})$ as the baseline measure M_0, with $p(\theta \mid \mathscr{G})$, $p(\mathscr{G})$ defined as in Section 2.

A well known equivalent representation of the previous model can be obtained by taking the limit as K goes to infinity of a finite mixture model with K components (clusters) of the form

$$\begin{aligned} \mathbf{x}^{(i)} \mid c_i, \{\theta_i, \mathscr{G}_i\}_{k=1}^K &\sim p\big(\mathbf{x}^{(i)} \mid \theta_{c_i}, \mathscr{G}_{c_i}\big), \\ (\theta_k, \mathscr{G}_k) &\sim M_0, \\ c_i \mid \mathbf{p} &\sim \text{Discrete}(p_1, \ldots, p_K), \\ \mathbf{p} &\sim \text{Dir}(\alpha/K, \ldots, \alpha/K). \end{aligned} \tag{9}$$

In particular, each $c_i \in \{1, ..., K\}$ is a random variable indexing the cluster associated with the i-th observation; for each $k = 1, \ldots, K$, $(\theta_k, \mathscr{G}_k)$ are instead cluster-specific parameters.

The latter representation allows for the implementation of a *collapsed* MCMC scheme where parameters θ_k's are integrated out and the algorithm approximates a marginal posterior distribution over the space of graphs and cluster indicators $c_1, \ldots, c_n$.

4 Computational details

Our sampling scheme for posterior inference relies on Algorithm 2 of [5], which applies to conjugate models whereas model-dependent parameters can be integrated out. Accordingly, the target is represented by the joint posterior

$$p\left(K, \{c_i\}_{i=1}^n, \{\mathscr{G}_k\}_{k=1}^K, \mid \mathbb{X}\right), \tag{10}$$

where K is the (random) number of clusters.

Given an initial state where the data $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are divided into K clusters (equivalently, the indicator variables $c_1, \ldots, c_n$ have been fixed) and graphs $\mathscr{G}_1, \ldots, \mathscr{G}_K$ are associated with each of the K components, an MCMC algorithm can be implemented by reiterating the following two steps.

As a first step we update the indicator variables $c_1, \ldots, c_n$ and, implicitly, the number of clusters K through a Gibbs sampling scheme which sequentially samples each c_i $(i = 1, \ldots, n)$ from its full conditional distribution. Specifically,

if $c_i = c_j$ for some $j \neq i$,

$$P(c_i = k \mid \mathbf{c}_{-i}, \mathbb{X}, \mathscr{G}_1, \ldots, \mathscr{G}_K) \propto \frac{n_{-i,k}}{n - 1 + \alpha} \int p(\mathbf{x}^{(i)} \mid \theta_k, \mathscr{G}_k)\, dH_{-i,k}(\theta_k \mid \mathscr{G}_k);$$

if $c_i \neq c_j \; \forall j \neq i$,

$$P(c_i \neq c_j \forall j \neq i \mid \mathbf{c}_{-i}, \mathbb{X}, \mathscr{G}_1, \ldots, \mathscr{G}_K) \propto \frac{\alpha}{n - 1 + \alpha} \int p(\mathbf{x}^{(i)} \mid \theta^*, \mathscr{G}^*)\, dM_0(\theta^* \mid \mathscr{G}^*),$$

where in particular:

- $n_{-i,k} = \sum_{j \neq i} \mathbf{1}\{c_j = k\}$, i.e. the number of indicator variables (excluding c_i) that are equal to k;

- $H_{-i,k}$ denotes the posterior distribution of θ_k based on the prior M_0 and given the data $\{\mathbf{x}^{(j)}, j \neq i, c_j = k\}$ and graph $\mathscr{G}_k$;
- $\mathscr{G}^*$ corresponds to an empty cluster (graph) which is randomly sampled from the baseline measure on $\mathbb{S}_q$.

Since the two integrals correspond, respectively, to posterior and prior predictive distributions and the categorical HD model is conjugate, we can provide closed-form expressions for both the two terms (we omit details).

Finally, conditionally on the cluster-indicator variables $c_1, \ldots, c_n$, each graph $\mathscr{G}_k$, $k = 1, ..., K$, can be updated as follows. Consider first a partition of the n observations into K clusters, $\{C_1, \ldots, C_K\}$, where

$$C_k = \{\mathbf{x}^{(i)} : c_i = k\}, \quad k = 1, \ldots, K.$$

We then update $\mathscr{G}_k$ through a Metropolis-Hastings step where a new graph $\mathscr{G}'_k$ is proposed from a suitable proposal distribution $q(\mathscr{G}'_k \mid \mathscr{G}_k)$ and then accepted with probability

$$\alpha_k = \min\left\{1, \frac{m(C_k \mid \mathscr{G}'_k)}{m(C_k \mid \mathscr{G}_k)} \cdot \frac{p(\mathscr{G}'_k)}{p(\mathscr{G}_k)} \cdot \frac{q(\mathscr{G}_k \mid \mathscr{G}'_k)}{q(\mathscr{G}'_k \mid \mathscr{G}_k)}\right\},$$

where $m(C_k \mid \mathscr{G}_k)$ denotes the marginal likelihood of graph $\mathscr{G}_k$ given the data C_k.

References

1. Lauritzen, S. L.: Graphical Models, Oxford University Press (1996)
2. Hartigan, J. A., Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **1**, 100–108 (1979)
3. Fraley. C., Raftery, A. E.: Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association **97**, 611–631 (2002)
4. Dawid, A. P., Lauritzen, S. L.: Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. The Annals of Statistics **21**, 272–1317 (1993)
5. Neal, R. M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics **9**, 249–265 (2000)
6. Carvalho, C. M., Scott, J. G.: Objective Bayesian model selection in Gaussian graphical models. Biometrika **96**, 497–512 (2009)
7. Ferguson, T. S.: A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics **1**, 209–230 (1973)

The Gaussian mixture model-based clustering for the comparative analysis of the Healthcare Digitalization Index in the Italian local health authorities

La cluster analysis basata sul modello di mistura di Gaussiane per l'analisi comparativa dell'Indice di digitalizzazione sanitaria nelle ASL italiane

Margaret Antonicelli[1], Michele Rubino[2], Filomena Maggino[1]

Abstract In recent years, the COVID 19 pandemic has highlighted the limitations of the health system which, with the digital transformation, can ensure benefits for patients and physicians. Digitization is a complex process that depends on numerous variables that can have a strong territorial connotation considering the organizational aspects of the Italian health system. This study aims to build a digitization index of Italian local health authorities, based on 6 parameters, with which we will carry out a comparative analysis between the regions through a Gaussian mixture model-based cluster analysis.

Abstract *In questi ultimi anni, la pandemia da COVID 19 ha evidenziato i limiti del sistema sanitario che con la trasformazione digitale può assicurare benefici per medici e pazienti. La digitalizzazione è un processo complesso che dipende da*

[1]1.Margaret Antonicelli, Department of Statistic, Sapienza University of Rome, e-mail: margaret.antonicelli@uniroma1.it.
1.Filomena Maggino, Department of Statistic, Sapienza University of Rome, e-mail: Filomena.maggino@uniroma1.it.
2.Michele Rubino, Department of Management, Lum Giuseppe Degennaro University, e-mail: rubino@lum.it.

numerose variabili che possono avere una forte connotazione territoriale considerando gli aspetti organizzativi del sistema sanitario italiano. Questo studio si propone di costruire un indice di digitalizzazione delle Aziende sanitarie locali italiane, basato su 6 parametri con il quale effettueremo un'analisi comparativa tra le regioni attraverso e di una cluster analysis basata sul modello di mistura di Gaussiane.

Keywords: Model based clustering, Healthcare Digitalization Index, complexity, territorial differences, Italian Local Health Centers.

1 Introduction

The COVID 19 pandemic highlighted the limits of the Italian health system and the need to accelerate its digital transformation processes. Healthcare represents a fundamental strategic area on which to build the development and restart of the country to respond quickly and effectively to events that impact the health of citizens (Nunez et al., 2021). At the same time, digital transformation is an important element in driving change in the national health system to improve relationships between patients and physicians. In fact, digitization is an enabling element for new healthcare based on data and information, the interoperability of systems, and services that are usable and accessible to citizens (Tortorella et al, 2020).
Some recent studies have shown that Italian healthcare is lagging in the digital transformation process and that services are inadequate (Patrone et al., 2018). The Italian Ministry of Health highlighted that the digitalization process of the health system is fragmented and inhomogeneous and that a central role in the management of relations between citizens and territories is played by local health authorities. On the above premises, this study aims to answer the following research questions:

RQ1: What is the digitization level of the Italian local health authorities? Are there differences between the different areas of the country or aspects worthy of attention?

To address these questions, this study aims to build a digitization index of Italian local health centers, based on 6 parameters, with which we will carry out a comparative analysis between the regions through a Gaussian mixture model-based cluster analysis.
The remainder of the paper is organized as follows. Section illustrates the methodology. Section 3 presents and discusses the research findings. Finally, implications and concluding remarks are given in Section 4.

2 Methodology

2.1 Research method

This study is based on two distinct and integrated methodologies: the construction of the health digitalization index (HDI) through the recovery of 6 parameters through web scraping analysis; the use of the clustering approach based on Gaussian mixture models to conduct the comparative analysis. The research was conducted by analysing the 101 Italian local health authorities active in December 2019 based on the public registers released by the Ministry of Health.

Referring to the HDI, 4 fundamental parameters are identified in the literature: presence of an electronic record; booking of online services; collection of online reports; online payment of health services. Each of these four parameters was measured through a fictitious variable, which was assigned the value 1 if these services were provided by the hospital, therefore present and active on the site, 0 if it was not provided. Given the complexity of the phenomenon, the organizational aspects are fundamental for its explanation; therefore, in addition to what has already been outlined, two other parameters were selected in order to construct the indicator: specialized IT personnel [5, 6] expressed as a percentage of qualified personnel out of total employees; chatbot for patient assistance, where the value 1 was assigned if this service was provided by the hospital, therefore present and active on the site, 0 if it was not provided [7, 8, 9]. We collected all data extracted from various sources using the vivid functionality of the scrapy web crawler using scripts written in Python 3.6 and further analysed according to customer needs where the data is stored in the company's database [10].

The data extraction process consists of 3 phases: 1) coding: the basic web crawling script used shows the scanned and stored data in the product database from a hospital site, in this case Reddit with the XPath method involved to find the details of each element of the Frequent Research; 2) testing: the project was tested using the various components as defined above and executed on the browser. The extraction carried out is completely relevant and the analysis carried out is estimated; 3) results: the overall results of the project are useful for understanding. Once all the parameters have been detected through web scraping, a further fundamental process is the construction of an index that evaluates and, above all, summarizes the complexity of the phenomenon of "digitization" in the healthcare sector. To this end, the most suitable methodology was considered for the benefit of the doubt, an application of the Date Envelopment Analysis (DEA). This method sees the composite indicator as the ratio between the performance of the single unit with that of the benchmark, ensuring that the optimal set of weights guarantees each associated unit the best possible position compared to all the others. The optimal weights are obtained as follows:

$$CI_c^* e^x = max_{w\,cq} \frac{\sum_{q=1}^{Q} I_{cq} w_{cq}}{max_{kc\,(1\ldots c)} \sum_{q=1}^{Q} I_{cq} w_{cq}}, \qquad \forall c = 1$$

where the weights are placed non-negative and I_{cq} represents the normalized score of the q-th simple indicator (q = 1, ..., Q) for the unit c (c = 1, ... C) and w_{cq} the

corresponding weight. The resulting composite indicator will therefore vary between 0 (the lowest performance) and 1 (the benchmark).

2.2 The Gaussian mixture model-based clustering

The Gaussian Mixture Model (GMM) assumes that the data comes from a finite mixture of G p-multivariate normal distributions, that is, that the population density can be expressed as:

$$f\ (y, \theta, \tau) = \sum_{k=1}^{G} \tau_k f_k\ (y, \theta_k)$$

where f_k and θ_k are the density and the parameters of the k-th component of mixture, and τ_k is the probability that the observation $\underline{y}$ belongs to the k-th component

$$(\ \tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1)$$

and f_k is a multivariate Normal density function, where θ_k is the parameter given by $\underline{\mu_k}$, Σ_k

$$l\ (\theta, \tau, z_i\ |x) = \sum_{1=1}^{n} \log \left[\sum_{k=1}^{G} z_{ik} \tau_k f_k\ (y_i | \theta_k) \right]$$

where z_i means that we have an unobservable vector z for each observation. These values must be estimated together with $\underline{\theta}$ (where θ_k represents in this case the "joint parameter" ($\underline{\mu_k}$, Σ_k) and $\underline{\tau_k}$ (k=1, … G).
An EM algorithm is used to estimate unknown parameters and unobservable variables. The criterion used to select the number of components (clusters) of the Gaussian mixture is the Bayesian Information Criterion (BIC):

$$BIC = 2l\ (\hat{\theta}, \hat{\tau}, \hat{z}_i\ |x) - (\#parameters) * \log(n)$$

in which

$$l\ (\hat{\theta}, \hat{\tau}, \hat{z}_i\ |x)$$

is the maximized value of the log-likelihood, (#parameters) is the number of parameters estimated, and *n* is the sample size. The BIC index is generally defined with the opposite sign compared to the definition used in this package (and therefore, smaller index => better).

3 Result and discussion

To proceed with the analysis, the 6 parameters that make up the HDI are grouped. It is useful, in this phase, to carry out evaluations of a descriptive nature; to this end, we proceed with the calculation of Pearson's linear correlation coefficients (the correlation character is scaled based on the size of the absolute correlation) and their level of significance with confidence level $\alpha = 0.05$ are reported (Table 1).

Table 1 - Pearson's linear correlation coefficients

	1	2	3	4	5	6
1 - Presence of electronic folder	1					
2 - Online service reservations	0,65*	1				
3 - Online report collection	0,59**	0,64	1			
4 - Online payment of healthcare service	0,48	0,77***	0,28	1		
5 - Specialized IT employees	0,51	0,38*	0,56**	0,35	1	
6 - Chatbot for assistance to patients	0,69**	0,72**	0,27*	0,39*	0,81***	1

The next step is the grouping of the 101 Italian local health authorities analysed using a model-based approach using Gaussian mixture models. The choice to use this approach depended on the fact that a unsupervised approach, where the number, shape and size of clusters are achieved through inferential statistics methods (BIC coefficient). Among the possible distributive hypotheses on the data, we focus on the multivariate Gaussian density mixture for its ability to approximate the density function of any unknown distribution.
The cluster analysis was conducted using the R Mclust software package, in which it was decided to use the variable volume model, equal shape and equal orientation and 3 components (clusters). In the Table 2 reports the indicators' means of each component and the number of units.

Table 2 - Means of components and number of units.

Parameters	Cluster 1	Cluster 2	Cluster 3
Presence of electronic folder	0,69	0,78	0,51
Online service reservations	0,84	0,87	0,63
Online report collection	0,52	0,81	0,39
Online payment of healthcare service	0,49	0,66	0,35
Specialized IT employees	0,41	0,84	0,28
Chatbot for assistance to patients	0,35	0,47	0,19
Number of Unit	47	25	29

The analyses carried out show a very strong territoriality of the phenomenon analysed. In particular, the territorial health units with average levels of digitalisation have returned to the first cluster; these are essentially located in central Italy, in Liguria, Sardinia and in part of Puglia. As regards the second cluster, which includes healthcare companies with high levels of digitization, these are concentrated exclusively in the north, with the sole exception of Rome. Finally, the third cluster groups the health units with medium-low levels of digitization, essentially located in southern Italy and Sicily.

4 Conclusion

The HDI is a suitable indicator for comparing the levels of digitalization between the various Italian ASLs. The classification of the various health bodies through clustering based on Gaussian mixture models, given the complexity of the phenomenon analysed, made it possible to make a reading and the results obtained much easier and more understandable and, consequently, allowed to make the comparison between the various territorial entities highlighting the substantial territoriality of the phenomenon that sees the entities of Northern Italy in a much better situation than those of the Southern regions.

References

1. Núñez, A., Sreeganga, S. D., & Ramaprasad, A., Access to Healthcare during COVID-19. International Journal of Environmental Research and Public Health, 18(6), 2021
2. Tortorella, G.L., Fogliatto, F.S., Mac Cawley Vergara, A., Vassolo, R., & Sawhney, R., Healthcare 4.0: trends, challenges and research directions. Production Planning & Control, 31(15), 1245-1260, 2020
3. Italian Minister of Health, 2021. Atto di indirizzo, Accessed on 10.06.2021 https://www.salute.gov.it/imgs/C_17_pubblicazioni_3036_allegato.pdf, 2021
4. Patrone, C., Lattuada, M., Galli, G., & Revetria, R., The role of Internet of Things and digital twin in healthcare digitalization process. In The World Congress on Engineering and Computer Science (pp. 30-37). Springer, 2018.
5. Rudaleva, I.A., Kabasheva, I.A., Balashova, E.Y., Khisamutdinov, A.N., Personnel typologization and barriers to innovative activity in the course of introducing digital technologies (by the Example of Healthcare Sector), CEUR Workshop Proceedings 2830, pp. 274-286, 2021
6. Shulzhenko, E., Holmgren, J., Gains from resistance: rejection of a new digital technology in a healthcare sector workplace, New Technology, Work and Employment 35(3), pp. 276-296, 2020
7. Lee, H., Kang, J., Yeo, J., Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: Development and deployment, Journal of Medical Internet Research 23(5), e27460, 2021
8. Martinengo, L., Lo, N.Y.W., Goh, W.I.W.T., Car, L.T., Choice of behavioral change techniques in health care conversational agents: Protocol for a scoping review, JMIR Research Protocols 10(7), e30166, 2021
9. Tuncel, F., Mumcu, B., Tanberk, S., A chatbot for preliminary patient guidance system. SIU 2021 - 29th IEEE Conference on Signal Processing and Communications Applications, Proceedings 9478023, 2021
10. Thomas, D.M., Mathur, S., Data Analysis by Web Scraping using Python. Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019 8822022, 450-454, 2019

Student performance evaluation

Rasch model *versus* Rasch Mixture model: strengthens and limits in identifying factors affecting students' performance in mathematics

Il modello di Rasch a confronto con il modello di mistura di Rasch: punti di forza e limiti nell'individuare i fattori che influenzano la performance degli studenti in matematica

Clelia Cascella

Abstract The Rasch model has been widely used in educational research to measure students' performance and to identify factors affecting it. Despite its multiple strengths, the Rasch model shows some limitations in exploring the relationship between students' performance and their (e.g., socio-demographic) characteristics, thus calling for the employment of alternative analytical solutions. In the current paper, I employed the Rasch Mixture Model to analyse data collected by INVALSI (at grade 5, in 2017), to (i) identify students' answers profiles, and (ii) understand if/how answers profiles are associated with students' characteristics. Similarities and differences between the Rasch model and the Rasch Mixture Model have been presented and discussed.

Abstract *Il modello di Rasch è stato diffusamente utilizzato nella ricerca educativa per misurare la performance scolastica degli studenti e identificare i fattori che la influenzano. Nonostante i suoi molteplici punti di forza, il modello di Rasch mostra alcuni limiti nell'esplorare la relazione tra la performance degli studenti e le loro caratteristiche (ad esempio, sociodemografiche), dunque suggerendo la necessità di implementare soluzioni analitiche alternative. In questo articolo, ho utilizzatoil modello di mistura di Rasch per analizzare i dati raccolti da INVALSI (al grado 5, nel 2017), per individuare i profili di risposta degli studenti e osservare se/come tali profili siano associati alle caratteristiche degli studenti. Somiglianze e differenze tra il modello di Rasch ed il modello di mistura sono state presentate e discusse.*

Key words: Rasch model, Mixture Rasch model, Mathematics, students' profiling

1 Introduction

Mathematics has been listed by the European Commission as one of the key competences needed by all for personal fulfilment and development, employability, and active citizenship. Understanding what factors can affect students' performance in mathematics is thus considered as a priority for both the national and the international agenda.

So far, most educational research has employed the Rasch model (RM) (Rasch, 1960), the first Item Response Theory (IRT) model, to estimate students' *ability*, a technical term used to refer to the latent trait measured via an achievement test (Andrich & Marais, 2019), consisting of a set items (i.e. questions). The Rasch model assumes that: (i) the latent trait is (at least, predominantly) unidimensional, (ii) the latent trait is measured via a set of items whose probability of being successfully encountered is assumed to be locally independent (i.e., the observed responses to each item are independent of responses to all other items, conditional upon the underlying trait); and, that (iii) the conditional probability of giving a correct answer follows a monotone non decreasing function of students ability (Hambleton & Swaminathan, 1985).

Within such an analytical framework, the Differential Item Functioning (DIF) analysis (Osterlind & Everson, 2009) has been extensively used to test for *measurement invariance* (Engelhard, 2009, 2013) of students' ability, measured across sub-groups of students, *matched on ability* and grouped by sex, ethnicity, socioeconomic status, or any other characteristic of interest. DIF analysis is used to test for measurement invariance because the RM assumes that the probability of giving a correct answer to an item is given by student's *relative ability*, that is his/her intrinsic ability (e.g., in mathematics) compared with item difficulty. DIF occurs when item's difficulty changes across sub-groups of students: if the same item is perceived as more or less difficult by students matched on ability ability, then the grouping variable is somehow associated with student's probability of encountering that item successfully. Such a relationship between the grouping variable and the probability of encountering an item successfully violates Rasch model assumptions as the RM assumes that the probability of encountering an item successfully is a function of students' relative ability, and that no other factor (sex, ethnicity, or any other variable) can affect it: any association between variable other than student's relative ability and the probability of encountering an item successfully is not modelled within the framework of the RM and can threat measurement invariance (Wright & Stone, 1979).

DIF analysis is thus a powerful tool to test for data-model fit, but it shows some serious drawbacks, especially in large-scale assessment, as it assumes homogeneity within sub-groups of students. For example, in a DIF analysis carried out by students' sex, results may "lead to a stereotypical view of an item as being more advantageous to all members of the group (males or females) favoured by DIF item, while ignoring the true heterogeneity within each group" (Cohen & Bolt, 2005): so, the same item could differently work in (and thus results as easier or as more difficult for) sub-groups of students within the male or the female group. Moreover,

DIF analysis can account for just one variable at time, without considering the possible intersectionality between variables. Results based on DIF analysis that ignores such an intra-group variability may thus mislead data interpretation.

In the current paper, I dealt with such a problem by employing the Rasch Mixture Model (RMM) (Rost, 1990), "a generic approach for modelling data that is assumed to stem from different groups (or clusters [or components]) but group membership is unknown" (Frick, Strobl, Leisch, & Zeileis, 2012a).

The analysis of large-scale assessment data collected in Italy has been used to present and discuss the differences between RM and RMM, and present one of the possible uses of RMM in educational research.

2 Methodology

Students' ability in mathematics is the latent trait investigated in present study. I measured it by analysing data collected by the Italian national institute for the evaluation of educational system (*Tr. Istituto Nazionale per la Valutazione del Sistema di Istruzione e formazione* - INVALSI). Data were collected by administering a mathematics achievement test[1] consisting of 13 multiple-choice items, 9 True/False items and 24 open-ended items, covering the subareas of (i) "space and shape" (roughly geometry), (ii) "change and relationship" (algebra), (iii) "quantity" (arithmetic), and (iv) "uncertainty" (probability), in a range of difficulty from those that require simple mathematical operations to those that require complex thinking. This achievement test was administered, in 2017, at both census (491,658 students) and sample level (35,000 students), in primary school (grade 5), (on average 10-years old students)[2].

Since the RMM is very sensitive to sample size and the model can over-extracts latent classes (when it is employed to analyse very large data sets), I tentatively extracted (from the INVALSI sample) five, smaller samples (of 25,000, 20,000, 15,000, 10,000 and 5,000 students each). Then, I comparatively analysed them by employing the RMM. The iterative procedure based on different sample sizes has been briefly described in the results section.

INVALSI data have been preliminary analysed via the Rasch (i.e. 1- parameter logistic) model, and via either 2- or 3- parameter logistic IRT models[3] to explore data-model fit as, if data fit the 2- or 3-PL IRT model better than 1-PL model, then applying a RMM may result in an overextraction of latent classes (Alexeev,

[1] The administered mathematics achievement test is available on-line at https://invalsi-areaprove.cineca.it/docs/file/05_Matematica_Fasc_1_2017.pdf

[2] A description of the sample is available on-line at https://www.dropbox.com/scl/fi/oj9xvzwxbj4utu7h7rssi/Sample-description.xlsx?dl=0&rlkey=0n97gnsghphyt5os99o9fq50b

[3] The 2- and the 3-PL models allow for the estimation of item discrimination and guessing, respectively: in the former item discrimination is allowed to vary item by item (as apposed to the Rasch model where item discrimination is supposed to be the same for each item); whereas the latter also estimates guessing, i.e. the probability of correctly answering an item by accident (Andrich & Marais, 2019).

Templin, & Cohen, 2011). Therefore, first, I checked for data model fit by performing a fit analysis within the framework of the 1-, 2- and 3-PL model, and then I analysed INVALSI data by using the RMM.

RMM modelling challenges Rasch estimates' invariance by allowing parameter to vary among latent classes (Rost, 1990), rather than assuming trait unidimensionality (Rasch, 1960). Mixture Rasch thus allows the exploration of a "mixture of latent sub-populations that are qualitatively different but within which a measurement model based on a continuous latent variable holds. In this modelling framework, one can characterize students by both their location on a continuous latent variable as well as by their latent class membership" (De Ayala & Santiago, 2017). Exploring the possible existence of latent classes can be precious because using a single or the same parameter estimation for all groups may result in a loss of information as it does not account for the heterogeneity in the population analysed (von Davier & Carstensen, 2007).

In the present study, the model with the lowest BIC has been presented in the results section. All the analyses have been carried out in R, by using both eRM (Mair & Hatzinger, 2007) and psychomix (Frick, Strobl, Leisch, & Zeileis, 2012b).

3 Results

The preliminary fit analysis (not reported here) showed that data fit both the 2- and the 3-PL IRT model slightly better than 1-PL model. Since that data fit both the 2- and the 3-PL model a bit better than the 1-PL model, then the Rasch Mixture Model may extract a number of latent classes greater than the actual number of clusters in the data. Nonetheless, the difference in BIC between the 1-, 2- and 3-PL model is small. An empirical analysis of INVALSI data via the MRM has thus been performed.

Results based on different sample sizes (25,000, 20,000, 15,000, 10,000 and 5,000) confirmed that the larger the sample size, the large the noise in the data: results based on 25,000, 20,000 and 15,000 students showed between 7 and 11 latent classes with just 1 latent class whose posterior probabilities were either high or low. In the ideal case, posterior probabilities of the observations for *each* latent class are either high or low, yielding a U-shape in all panels (Frick et al., 2012a).

In contrast results based on smaller sample sizes (both 10,000 and 5,000 cases) showed that the optimal model (i.e., that showing the lowest AIC and BIC) was that with 5 latent classes (Figure 1). In this case, the rootogram of posterior probabilities shows that in 3 out of 5 latent classes posterior probabilities were either high or low, yielding in a U-shape in 2 out of 5 classes (class 3 and 4). The remaining three classes (i.e., 1, 2, and 5) showed a decreasing shape, thus raising concerns about their interpretability. These three classes may be "not real", but over-extracted (Alexeev e 2011). Nonetheless, it is worth noting that this just an interpretative hypothesis: determining the best fitting model is still a challenging, open problem in the existing literature that calls for further research.

Item difficulty changes by class. Figure 3 shows a selection of items whose estimated difficulty largely changes by class. By comparing item distribution by class, on average, estimated item difficulty is relatively lower in class 3 than in the other classes, with some exceptions (such as item the D12 and the item D20, whose estimated difficulty is sharply higher than that in most of the other latent classes). In line with both the national and the international literature, students' probability of belonging to class 3 is higher for Italian males, whose age is that expected for grade 5 (i.e., neither retained nor in advance student), and from high-SES families.

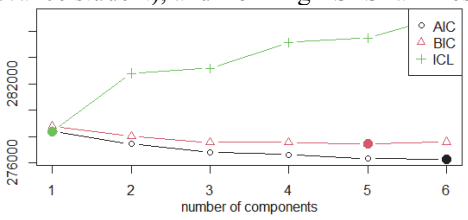


Figure 1. *Number of components extracted*

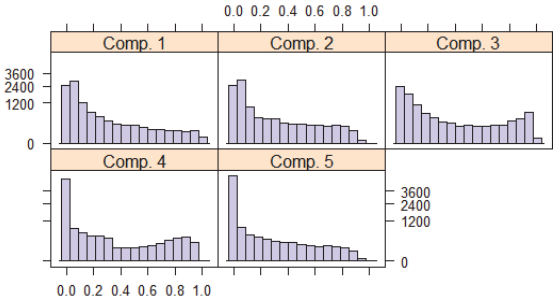


Figure 2. *Rootogram of posterior probabilities*

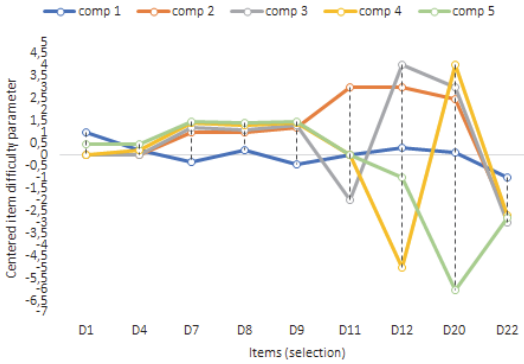


Figure 3. *Item profiles for the 5-component (selection of items showing the highest difference in terms of estimated difficulty between latent classes)*

4 Discussion and conclusions

Identifying factors affecting students' performance in mathematics is a priority for both the national and the international agenda. Existing methods to address such an

aim (such as DIF analysis within the framework of the Rasch analysis) show some technical criticalities.

The present paper aimed to show the functionality of an alternative analytical approach based on the RMM. As opposed to the RM (Rasch, 1960/1980), the RMM (Rost, 1990) does not assume data unidimensionality, but it accounts for data heterogeneity by allowing parameter to vary among latent classes (Rost, 1990).

In the present study, I anaysed INVALSI data by employing the Rasch Mixture Model to (i) detect latent classes that differ in patterns of item difficulty; and, (ii) associate students' characteristics to the probability of belonging to each of the latent classes extracted and, thus, to items' difficulty.

Studying such an association can solicit new research questions. For example, why items D12 was perceived as more difficult by students whose probability of belonging to the latent class 3 is higher (even though, on average, students' ability in mathematics is higher than in the other latent classes)? Why such a gap (in items difficulty) is larger between class 3 and class 4 even though all the other items difficulty were similar (and thus students seem to be matched on mathematics ability)? Do students' characteristics play a role in explaining such a gap?

5 References

1. Alexeev, N., Templin, J., & Cohen, A. S.: Spurious latent classes in the mixture Rasch model. Journal of Educational Measurement **48**(3), 313–332 (2011).
2. Andrich, D., & Marais, I.: A course in Rasch measurement theory. Measuring in the Educational, Social and Health Sciences. Springer. Singapore (2019).
3. Cohen, A. S., & Bolt, D. M.: A mixture model analysis of differential item functioning. Journal of Educational Measurement **42**(2), 133–148 (2005).
4. De Ayala, R. J., & Santiago, S. Y.: An introduction to mixture item response theory models. Journal of School Psychology **60**, 25–40 (2017).
5. Engelhard, G.: Item and Person Functioning for Students With Disabilities. Educational and Psychological Measurement **69**(4), 585–602 (2009).
6. Engelhard, G.: Invariant measurement using Rasch models in the social, behavioral, and health sciences. New York: Routledge (2013).
7. Frick, H., Strobl, C., Leisch, F., & Zeileis, A.: Flexible Rasch mixture models with package psychomix. Journal of Statistical Software **48**, 1-21 (2012).
8. Hambleton, R. K., & Swaminathan, H.: Item Response Theory: Principles and Applications. Netherlands: Springer (1985).
9. Mair, P., & Hatzinger, R.: Extended Rasch modeling: The eRm package for the application of IRT models in R. EPub WU Institutional Repository (March, 2007).
10. Osterlind, S. J., & Everson, H. T.: Differential Item Functioning. Thousand Oaks: SAGE Publications, INC (2009).
11. Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. Copenhagen: Danish Institute for Educational Research
12. Rost, J.: Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. Applied Psychological Measurement **14**(3), 271–282 (1990).
13. von Davier, M., & Carstensen, C. H.: Multivariate and Mixture Distribution Rasch Models. (Springer, Ed.) (2007).
14. Wright, B. D., & Stone, M. H.: Best test design. (MESA Press). Chicago (1979).

Does taking additional Maths classes improve university performance?

A clustered case-control study based on Italian high-school and university data

Martina Vittorietti, Andrea Priulla, Massimo Attanasio

Abstract Several recent studies in educational literature showed how students' skills in maths affect their success at higher levels of education. The aim of this paper is to evaluate the effect of taking additional maths class at high school on first-year performance of Italian university students. However, university performance and the choice of the high-school depend on several factors that make this evaluation challenging. Using information coming from three different sources, we carry out a multilevel propensity score procedure to estimate the average treatment effect between the applied sciences track and the traditional scientific one. After balancing for school- and student-level covariates, the results of a logistic regression model suggest no difference between the two school tracks.

Abstract Studi recenti nell'ambito dell'istruzione secondaria hanno mostrato come le abilità matematiche degli studenti abbiano un effetto sul loro futuro successo scolastico e universitario. L'obiettivo di questo lavoro è valutare se studiare più matematica a scuola è associato ad una migliore performance al primo anno di università. La performance universitaria e la scelta della scuola dipendono tuttavia da molteplici fattori che rendono complessa questa analisi. Utilizzando dati provenienti da tre diverse fonti, applichiamo un multilevel propensity score matching per stimare l'effetto medio di due trattamenti: "il liceo scientifico delle scienze applicate" e il "liceo scientifico tradizionale". Dopo aver bilanciato le caratteristiche della scuola e degli studenti, i risultati di un modello logistico mostrano che non emergono differenze tra i due percorsi scolastici.

Key words: propensity score, fine balance, educational data, multilevel propensity score, optimal network flow

Martina Vittorietti, Andrea Priulla, Massimo Attanasio
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: martina.vittorietti@unipa.it, andrea.priulla@unipa.it, massimo.attanasio@unipa.it

1 Introduction

Mathematics is perceived as the foundation for scientific and technological knowledge that is cherished by societies worldwide. Due to its relevance and application, mathematics is widely regarded as one of the most important school subjects and a central major of the school curriculum [8]. mathematics underpins the study of many disciplines at university, not only in science, technology, engineering and mathematics (STEM) but also in agriculture, pharmacy and economics [10]. In Italy, there is at least one mathematics or mathematics related course in most of the degree courses, and it is often regarded as one of the toughest exams. Several surveys at both international level, such as the PISA tests promoted by OECD, and at national level, such as INVALSI tests, are conducted for investigating mathematical skills in high school. Italy is characterized by a high school system organized in separate tracks, which sorts students according to their academic or vocational orientation. Despite the existence of different tracks, the high school could be described as tripartite, with a 5-year academic oriented generalist education provided by lyceums, with further distinctions in humanities, science activities, languages, pedagogical sciences, 5-year technical schools, and 5-year vocational schools. In 2010, Italy underwent a reform of the secondary school education. One of the main innovations was the introduction of a new school curriculum: the applied sciences curriculum. The main difference with respect to the traditional scientific track was the substitution of the Latin language class with classes in scientific subjects such as mathematics, physics, natural sciences, and the addition of computer science as a new subject. In [12] the author compares the in-going and out-going mathematical skills of the students of the two different curricula, underlying the strengths and the weaknesses of both tracks, but as far as the authors' knowledge there are no results about the academic performances related to this new high school track. Therefore, the main objective of this paper is to compare the academic performances of the students coming from traditional scientific or applied sciences tracks, in order to answer the following question: "Does taking additional maths classes improve the university performance?" Before answering this question, it should be noted that university performance lies above an incredibly intricate net of multi-dimensional factors. Therefore, the question in the present form is ill-posed because of social and cultural factors that affect the choice of studying more mathematics or more correctly the choice of the applied sciences track, that eventually will affect university choice and the university performance too [1]. Among the social and cultural factors, the most relevant are: gender [3, 6], social level, location of the high school [2, 11], socio-economic status of the family [4, 16]. For investigating the difference between the two tracks, a balancing for all the possible factors related to both the school track and the academic performance is needed. Analogously to clinical studies, we investigate the effect of the *treatment* "Studying more maths". It is possible to see this study as a clustered case-control study, in which the cases are the students who attended the applied sciences track, and the control are the students who attended the traditional Scientific track. We propose a multi-level balancing approach of the the socio-demographic characteristics using a multi-level propen-

sity score model. This procedure aims to eliminating the selection bias due to the fact that, in observational studies, the choice of the high school track, is not random, because it is related to socio-economic and cultural factors[4]. The number of ECTS achieved by the two group of students, treated and controls, in their first academic year is then compared and the average treatment effect (ATE) is estimated.

2 Data

The dataset used in this paper is built by linking three distinct administrative sources: aggregate data coming from the National Archive of Schools (ANS-S); micro-data on high school career from the National Evaluation Institute for the School System (INV);

As most of the educational data, the data at hand presents a hierarchical structure. In fact, it is possible to distinguish variables based on two levels:

1. *school-level*: macro-region, university enrollment rate and the Economic, Social, Cultural Status index (ESCS) used by INVALSI [14];
2. *student-level*: gender, ESCS, high school career's regularity.

3 Methods

Educational data often presents a multilevel structure, so, it is common to speak of clustered treatments. In clustered-treatment observational studies, treatments are applied to entire cluster (schools) or to individuals (students or teachers), or to multiple levels simultaneously (schools and students) [13]. Let S be the number of matched pairs of schools $s = 1,\dots,S$. Each school presents both classes belonging to the applied sciences and the traditional scientific track, $j = 1,2$, the treatment and the control respectively, for a total of $2S$ units. Each school track of each school s_j contains $n_{s_j} > 1$ students, $i = 1,\dots,n_{s_j}$. Each pair is matched on a vector of observed pre-treatment covariates $x_{s_{j_i}}$. Let X_{s_j} be the matrix whose rows consists of the $x_{s_{j_i}}$ vectors for each student i in the j-th track of the s-th school with support $\mathbf{X} \subset \mathbf{R}$. A student i in school track j of school s is described by both observed covariates and possibly unobserved covariates $u_{s_{j_i}}$. In our study, treatment assignment occurs at class level as whole classes are assigned to treatment or to control. If the j-th class in pair s receives the treatment $Z_{s_j} = 1$, whereas if it receives the control $Z_{s_j} = 0$, thus $Z_{s_1} + Z_{s_2} = 1$, for each s as each pair contains the treatment classes and the control classes. Following the designed structure, the statistical analysis strategy uses a propensity score model as a matching strategy [9, 15]. In particular, we follow the optimal multilevel matching using network flows proposed in [13], in which first we match the schools on school-related covariates and then we match students using individual socio-demographic characteristics. After the matching procedure, we fit a logit regression model with random effects on the matched data, to evaluate the

effect of the treatment on the probability of achieving at least 37 ECTS. We chose the cut-point of 37 ECTS because empirical evidences suggest that 37 is the best cut-point predictor of university success

4 Results

In Table 1 we assess the balance among the students covariates before and after the matching procedure.

	Unmatched				Matched			
	Treated	**Control**	**SMD**	**P-value**	**Treated**	**Control**	**SMD**	**P-value**
Male %	71,0%	50,0%	0,44	0,00	66,9%	65,0%	0,04	0,17
Student's ESCS	0,39	0,52	-0,14	0,00	0,48	0,52	-0,05	0,08
HS final mark	78,87	81,56	-0,22	0,00	79,41	79,88	-0,04	0,16
INVALSI math result	245,61	245,20	0,01	0,34	248,83	248,42	0,01	0,66
INVALSI italian result	226,04	233,24	-0,22	0,00	231,05	232,26	-0,04	0,18
INVALSI english result	451,84	459,11	-0,13	0,00	460,38	461,73	-0,02	0,40

Table 1 Student covariates comparison in the treatment and control groups before and after the matching.

According to the standardized mean difference (SMD) [5] before the matching procedure, the major imbalance is in terms of gender, while no imbalance is recorded for the INVALSI math results. After the matching procedure all the student covariates are balanced, in fact no SMD is statistically significant.

In Table 2, the results of two logistic models with random effects on the university field of study fitted on unmatched and matched data are reported. The main differences are: before matching, the treatment effect was negative and significant, suggesting that the applied sciences students have worse performance with respect to their peers of the traditional scientific; after matching, the treatment is not anymore effective. Moreover, in Figure 1, it is possible to see how the treatment varies after the matching procedure with respect to the field of study: if on unmatched data the treatment had a positive effect for the STEM courses and negative on no STEM courses, after the matching the effect is reversed.

5 Conclusions

In this paper we compare the university performance of students that attended two different school tracks, to assess the effect of different skills. The main results show that the two groups significantly differ by gender, socio-economic level, final high school grade and scores in English and Italian INVALSI tests, while no difference are observed in maths test scores. After having balanced both school and students

Variable	Unmatched		Matched	
	Estimate	Pvalue	Estimate	Pvalue
Intercept	-5,48	0,00	-5,98	0,00
HS diploma (ref="Scientific")	-0,11	0,05	-0,04	0,56
HS final mark	0,07	0,00	0,07	0,00
Genere (ref="Female")	-0,14	0,00	-0,22	0,01
HS macro-region (ref="North") - Center	-0,61	0,00	-0,59	0,00
HS macro-region (ref="North") - South and Islands	-0,98	0,00	-0,92	0,00
Student's ESCS	0,06	0,00	0,04	0,26
INVALSI math result	0,01	0,00	0,01	0,00
INVALSI italian result	0,00	0,95	0,00	0,16
INVALSI english result	0,00	0,00	0,00	0,54

Table 2 Parameters of logit regression models with random effects before and after matching.

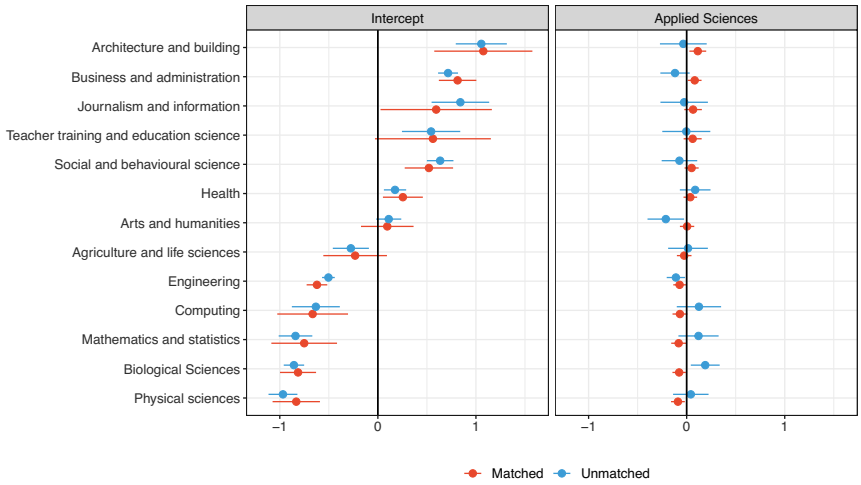


Fig. 1 Comparison of the intercept and the treatment (Applied Sciences) parameters for each university field of study estimated using both the unmatched and the matched data.

covariates, the results of a logit regression model with random effects with respect to the university field of study suggest that there is no difference between the two tracks. However, some considerations are needed. We evaluate the difference among the two different school tracks just with respect to their academic performance. In reality, it already exists a difference in the enrollment rates of the students belonging to the two groups, strongly in favor of the traditional scientific school track. This tunes the matching approach towards a specific direction: it selects the best students of the applied sciences school track. The Italian high school system is still hierarchical and the traditional scientific track is then considered as characterized by a better "social context" compared to the applied science track. Our procedure can not balance for this social context, which is an unobservable variable.

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P.

References

1. Aina, Carmen, Massimiliano Bratti, and Enrico Lippo. "Ranking high schools using university student performance in Italy." Economia Politica 38.1 (2021): 293-321.
2. Azzolini, Davide, and Loris Vergolini. "Tracking, inequality and education policy. Looking for a recipe for the Italian case." Scuola democratica 2 (2014): 0-0.
3. Cascella, Clelia, Chiara Giberti, and Giorgio Bolondi. "An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks." Studies in Educational Evaluation 64 (2020): 100819.
4. Checchi, Daniele, and Luca Flabbi. "Intergenerational mobility and schooling decisions in Germany and Italy: The impact of secondary school tracks." (2007).
5. Cohen, Jacob. "Statistical power analysis for the social sciences." Hillsdale, NJ: Lawrence Erlbaum (1988).
6. Contini, Dalit, Di Tommaso, Maria L., and Mendolia, Silvia. "The gender gap in mathematics achievement: Evidence from Italian data." Economics of Education Review 58 (2017): 32-42.
7. D'Amore, Bruno, et al. "La didattica della matematica: strumenti per capire e per intervenire." Atti del Convegno Nazionale omonimo (2014): 3-4.
8. Hagan, John Ekow, et al. "Students' perception towards mathematics and its effects on academic performance." Asian Journal of Education and Social Studies 8.1 (2020): 8-14.
9. Keele, Luke, Matthew Lenard, and Lindsay Page. "Matching Methods for Clustered Observational Studies in Education." Journal of Research on Educational Effectiveness 14.3 (2021): 696-725.
10. Nicholas, Jackie, et al. "Mathematics preparation for university: entry, pathways and impact on performance in first year science and mathematics subjects." International Journal of Innovation in Science and Mathematics Education 23.1 (2015).
11. Panichella, Nazareno, and Moris Triventi. "Social inequalities in the choice of secondary school: Long-term trends during educational expansion and reforms in Italy." European Societies 16.5 (2014): 666-693.
12. Petolicchio, Annamaria. "La Riforma del Liceo Scientifico: il curricolo Scienze Applicate." FORMAZIONE & INSEGNAMENTO. Rivista internazionale di Scienze dell'educazione e della formazione 14.2 (2016): 187-194.
13. Pimentel, Samuel D., et al. "Optimal multilevel matching using network flows: An application to a summer reading intervention." The Annals of Applied Statistics 12.3 (2018): 1479-1505.
14. Ricci, Roberto. "The Economic, Social, and Cultural Background: a continuous index for the Italian Students of the fifth grade." Mimeo (2010).
15. Rosenbaum, Paul R. "Modern algorithms for matching in observational studies." Annual Review of Statistics and Its Application 7 (2020): 143-176.
16. Xie, Yu, Michael Fang, and Kimberlee Shauman. "STEM education." Annual review of sociology 41 (2015): 331-357.

University dropout and churn in Italy: an analysis over time

Abbandono degli studi universitari e churn in Italia: un'analisi nel tempo

Barbieri Barbara, Porcu Mariano, Salaris Luisa, Sulis Isabella, Tedesco Nicola and Usala Cristian

Abstract This paper analyses dropout and churn in the 1st year at the university by using micro data related to the cohorts of freshman students enrolled in Italian universities between 2010 and 2018 based on ANS data. The analysis focuses on the role that students' education background and the choices related to the field of study have on influencing the outcomes of the university system. Moreover, the paper pays attention on the role that the quality of university facilities and communication have in preventing possible system inefficiencies.

Abstract *Questo lavoro analizza il fenomeno dell'abbandono degli studi universitari e del "churn"(cambio di istituzione universitaria) per gli studenti immatricolati nelle università italiane tra il 2010 e il 2018 utilizzando i microdati dell'ANS forniti dal MUR. L'analisi si concentra sul ruolo che il background scolastico degli studenti e la scelta del gruppo disciplinare hanno sull'influenza dei risultati osservati. Il contributo, inoltre, presta attenzione al ruolo che la qualità delle strutture e della comunicazione universitaria hanno nella prevenzione delle inefficienze del sistema.*

Key words: university dropout; university churn, multilevel models, ANS, university policies

[1] Barbieri Barbara, University of Cagliari; email: barbara.barbieri@unica.it
Porcu Mariano, University of Cagliari; email: mariano.porcu@unica.it
Salaris Luisa, University of Cagliari; email: salaris@unica.it
Sulis Isabella, University of Cagliari; email: isulis@unica.it
Tedesco Nicola, University of Cagliari; email: tedesco@unica.it
Usala Cristian, University of Cagliari; email: cristian.usala@unica.it

1 Introduction and aim

The inefficacy of the Italian university system has been largely discussed in Italy, with main interest after the 3+2 reform toward the monitoring the regularity of students' careers and the identification of good policies and practices to prevent academic dropout [1-4]. On the national scene, in fact, the events of irregularity of university careers and the dropout - the two itineraries with which the university dispersion takes shape - are revealed with a consistency that is much higher than that recorded in the main nations of the continent.

The analysis focuses on the role that the students' educational path and their university choices (in terms of local and non-local universities, and of course of study) have in influencing the indicators of the inefficiency of a university system in terms of dropout and churn. Moreover, the paper pays attention on the role that the quality of university facilities and communication have in preventing both phenomena thanks to the use of CENSIS indicators on contributions, communication, service, facilities available for 2011-2018.

We move from the hypotheses that the inefficiency of tertiary education institutions is determined by factors external and internal to the university system. The former factors are strictly linked to students' previous educational backgrounds [5-7]. The positive or negative impact that the educational background has in determining students' success at the university is a topic of great relevance to make guidance and tutoring services able to provide the best support to students in the transition from secondary school to university. Nonetheless, the internal factors can be related to the peculiarities of the field of study chosen but also to several policies that universities can implement to prevent adverse events.

This contribution also hypothesizes that university policies aimed at improving the functionality of communication and digital services (university websites, official social profiles) and their effectiveness in providing effective support to students, can prevent churn. For this aim, the microdata provided by the MUR relating to cohorts of students enrolled in the period 2010-2018 were analysed and students were classified according to their career after one year as regular, drop out (student who has not renewed enrolment to any Italian university), churn (student who changed university between the first and second year). The three CENSIS indicators related to the quality of university facilities, communication and digital services and grants for 2011-2018 have been used to assess the role that universities policies addressed to enhance these dimensions have in order to prevent inefficiencies of the university system over and above the differences in students' profiles related to demographic characteristics, previous education experiences, the choice of attending a local or non-local universities, and the heterogeneity in the field of studies. Multilevel multinomial logit models have been used to this aim as they allowed us to disentangle the effect that factors which act and individual level and field of study level can play in preventing crunch and dropout.

2 Data

The data here considered are from the ANS – Anagrafe Nazionale Studenti (Italian National Student Archive) for the academic year 2010-2018 (Database MOBYSU.IT) which includes the cohorts of students who enrolled during the same years in a Bachelor degree program of an Italian University. The following covariates have been considered: age, gender, foreigner status, high school grade, high school, macro area of residence, year of enrollment and ISCED field, mobility status (free movers and forced movers) [1]. CENSIS indicators on grants, communication, service, structures available for 2011-2018 where used to integrate information on university contributions, communication, service, facilities.

The analysis covers a period of 9 years (from 2010 to 2018) looking at the careers of an average of 200 thousand enrolled per year for a total of 1.8 million students (Figure 1). Figure 2 reports their distribution according to ISCED field.

Figure 1 – Students by year (2010-2018) (absolute number)

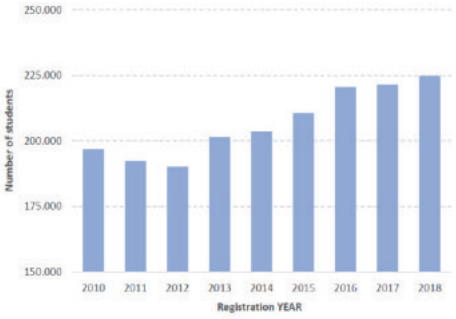
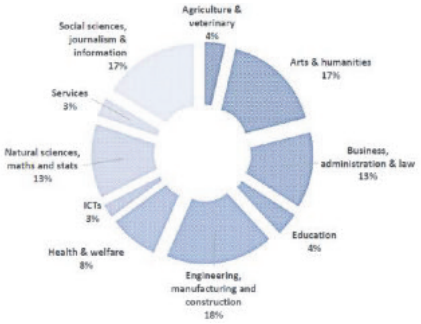


Figure 2 – Students by ISCED field (percentage)



As concerns the careers (Figure 3), it is noted that on average 80.9% of students after the first year continue their studies in the same course and University in which they have enrolled, 14.9% dropout and 4.2% churn. The percentage of student that churn does not show significant variations in time, while dropout has instead decreased, passing from 17.6% in 2010 to 14.2% in 2018 (-3.4%).

Males and females do not behave in the same way with respect to the propensity to dropout, with males dropping out more frequently (17.5%) than females (12.7%). There are also some differences with respect to the diploma, students from the classical and the scientific institute in fact record lower percentages of drop out

[1] Free movers: students who select a non-local university that is located more than 15 minutes of travel time from the closest university to their city (local university) and the closest university offers courses in the field of study where they enrolled.
Forced movers: students who select a non-local university that is located more than 15 minutes of travel time from the closest university to their city (local university) and the closest university does not offer courses in the field of study where they enrolled.

(respectively 7.9% and 8.6%) and higher of churn, and by contrast the students arriving from technical institutes (23.7%) and vocational institutes (29.6%) register higher proportion of drop out.

Finally, as regards ISCED field (Figure 4), the areas of Information Communication & Technologies (ICTs) and of Agriculture & Veterinary that record the highest percentages of drop out. It is interesting to note that for Natural Sciences, Mathematics and Statistics there is a significant weight of churn.

Figure 3 – Carrier outcome by year (percentage)

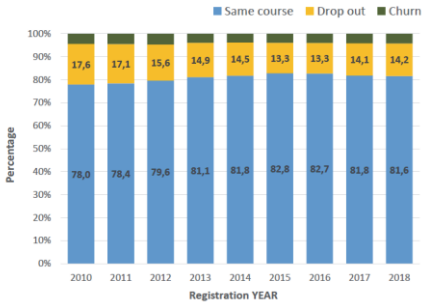
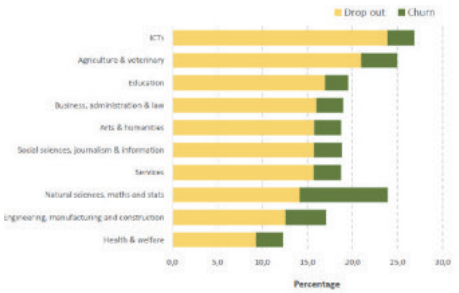


Figure 4 – Drop out and Churn by ISCED field (percentage)



3 Modelling approach and results

A multinomial multilevel logit model has been adopted to study factors mainly associated with the probability of dropout or churn during the first year at the university. The baseline event, model the probability to enrol in the second year at the same university. We considered several factors, which are associated with students' success during the study to shape some lights on internal and external factors, which can influence both outcome and their trend over time. We include three set of predictors which allow us to account for divergences in socio-demographic factors (gender, age, area of residence; foreigner status), educational background (high school, and high school grade), university choices (in terms of local and non-local university and field of study) and university policies in supporting students (CENSIS indicators on grants, communication, services, and structures).

The results of the multinomial logit model reported in Table 1 show that in terms of socio-demographic characteristics female students from the North who select non-local universities have a lower probability of dropout. Moreover, the probability of dropping out is highly associated with previous educational experiences. Namely students from vocational and technical schools are those with the highest risk to dropout during the first year and this event is negatively affected by the high school grade. The sign and the size of the coefficients of the dummy indicators for monitoring divergences across cohorts show a progressive reduction of the dropout

rate between 2011 and 2017 and a turnaround from 2017. The churn rate shows a different behaviour with respect to high school provenience with the highest churn rates associated with students from Scientific High School and who selected non-local universities. The negative sign of the coefficients of the CENSIS indicators shows that the quality of communication, and structures has an average a positive, even if weak, effect in preventing churn and dropout events.

Table 1: Results of the multinomial multilevel model: baseline category Regular

Predictors	DROPOUT		CHURN	
	coeff	95% CI	coeff	95% CI
Cons	3.986	[3.896;4.079]	3.632	[3.411;3.855]
Censis Grants	-0.004	[-0.004;-0.004]	-0.005	[-0.006;-0.004]
Censis Communication	-0.004	[-0.004;-0.003]	-0.010	[-0.011;-0.009]
Censis Services	0.000	[-0.001;0.000]	-0.014	[-0.015;-0.013]
Censis Structures	-0.003	[-0.003;-0.002]	0.007	[0.006;0.008]
Age	0.053	[0.052;0.054]	-0.073	[-0.079;-0.068]
Female	-0.226	[-0.236;-0.216]	0.006	[-0.010;0.023]
Foreigner	-0.016	[-0.039;0.007]	-0.166	[-0.222;-0.109]
Mov_Forc	-0.138	[-0.151;-0.123]	0.452	[0.431;0.473]
Mov_Free	-0.085	[-0.097;-0.073]	0.243	[0.221;0.264]
High School Grade	-0.047	[-0.048;-0.047]	-0.017	[-0.018;-0.017]
High School (Classico)				
Other	0.835	[0.816;0.854]	-0.205	[-0.230;-0.180]
Scientific	-0.025	[-0.044;-0.007]	-0.350	[-0.372;-0.328]
Technical	1.113	[1.095;1.132]	-0.473	[-0.499;-0.446]
Vocational	1.529	[1.507;1.548]	-0.447	[-0.488;-0.409]
Macro Area (Center)				
Islands	0.078	[0.062;0.095]	-0.266	[-0.296;-0.237]
North	-0.314	[-0.326;-0.302]	-0.278	[-0.299;-0.258]
South	0.039	[0.015;0.051]	0.087	[0.055;0.119]
year2011	0.033	[-0.093;-0.058]	0.086	[-0.011;0.052]
year2012	-0.075	[-0.146;-0.111]	0.020	[-0.146;-0.085]
year2013	-0.128	[-0.204;-0.169]	-0.116	[-0.211;-0.148]
year2014	-0.187	[-0.278;-0.242]	-0.180	[-0.169;-0.104]
year2015	-0.260	[-0.292;-0.255]	-0.137	[-0.162;-0.099]
year2016	-0.273	[-0.252;-0.216]	-0.130	[-0.119;-0.057]
year2017	-0.233	[-0.232;-0.196]	-0.088	[-0.177;-0.116]
year2018	-0.214	[2.821;15.900]	-0.146	[1.713;9.871]
Level2				
Variance	6.728	[2.821;15.900]	4.238	[1.713;9.871]
Cov(Drop, Churn)	5.239	[2.121;12.324]		
Bayesian DIC	1975595			

4 Discussion

In line with the scientific literature, the data from this study show that dropout probability is strongly associated with previous educational experiences, particularly for students with vocational and technical backgrounds. Furthermore, the choice of attending a non-local university appears as a factor which prevent dropout especially for females studying in Northern Italy universities.

This study also sought to assess student dropout and churn based on the presence and quality of certain services offered aimed at increasing the efficiency and effectiveness of engagement programs. Our study shows that CENSIS indicators have an average positive, albeit weak, effect in the prevention of dropout and churn events. If the issue of strengthening entry-level training orientation initiatives and the development of orientation services *in itinere*, recovery and didactic tutoring, especially during the first year of studies, certainly remain key issues for limiting dropouts.

References

1. Perchinunno, P., Bilancia, M., & Vitale, D. (2021). A statistical analysis of factors affecting higher education dropouts. *Social Indicators Research*, 156(2), 341-362.
2. Ballarino, G. (2011). Le politiche per l'università. In U. Ascoli (ed.), *Il welfare in Italia* (pp.197-224). Bologna: il Mulino.
3. Tedesco, N. & Salaris, L. (2020). University dropout and mobility in Italy. First evidence on first level degrees, *SIS 2020 - Book of Short Papers*, 9788891910776, PEARSON, pp.1601-1606.
4. Trivellato, P., & Triventi, M. (2011). Differentiated trends in student access and performance during the "Bologna Process". The case of universities in Milan. *Italian Journal of Sociology of Education*, 3(2).
5. Di Pietro, G., & Cutillo, A. (2008). Degree flexibility and university drop-out: The Italian experience. Economics of Ed Di Pietro, G., & Cutillo, A. (2008). Degree flexibility and university drop-out: The Italian experience. *Economics of Education Review*, 27, 546-555.ucation Review, 27, 546-555
6. Belloc, F., Maruotti, A. & Petrella, L. (2011) How individual characteristics affect university students drop-out: a semiparametric mixed-effects model for an Italian case study, *Journal of Applied Statistics*, 38:10, 2225-2239, DOI: 10.1080/02664763.2010.545373.
7. Meggiolaro S., Giraldo A., & Clerici R. (2017). A multilevel competing risks model for analysis of university students' careers in Italy. *Studies in higher education*, vol. 42, pp. 1259-1274, doi: 10.1080/03075079.2015.1087995.

The ANOGI for detecting the impact of education and employment on income inequality

L'ANOGI per studiare l'impatto del titolo di studio e dell'occupazione sulla disuguaglianza dei redditi

Elena Fabrizi and Alessio Guandalini and Alessandra Spagnoli

Abstract The ANOGI proposed by Frick et al. (2006) decomposes the overall inequality measured by Gini index by three components: within, between and overlapping. Based on ADSILC data we analyze the inequality of income, by separating two aspects: inequality due to education and to the career. The overlapping component shows how intertwined the subgroups are, whereas increases in stratification can cause a negative effect on inequality. In this context, its definition has been extended to compare pairs of workers according to their career and education. We demonstrated that workers with higher education are associated to the highest level of stratification.

Abstract L'analisi ANOGI proposta da Frick et al. (2006) permette di scomporre la disuguaglianza misurata dall'indice di Gini in tre componenti: within, between and overlapping. Basandoci sulla banca dati ADSILC abbiamo analizzato le disuguaglianze delle retribuzioni, separando l'effetto della disuguaglianza riconducibile all'istruzione da quella dovuta ad una carriera lavorativa. La componente overlapping definisce l'interconnessione tra i sottogruppi, mentre un aumento della stratificazione può causare un effetto negativo in termini di disuguaglianze. In questo contesto, tale definizione è stata estesa in modo da comparare gruppi di lavoratori con carriere e livelli di istruzione differenti. I nostri risultati indicano che i lavoratori con istruzione elevata sono associati ad elevati livelli di stratificazione.

Key words: Income distribution, decomposition by subgroups, work history pattern, ANOGI

Elena Fabrizi
Università degli Studi di Teramo, e-mail: efabrizi@unite.it

Alessio Guandalini
Istituto Nazionale di Statistica (ISTAT) e-mail: alessio.guandalini@istat.it

Alessandra Spagnoli
Università "La Sapienza" di Roma e-mail: alessandra.spagnoli@uniroma1.it

1 The Gini index decomposition by subgroups

The Gini index (Gini, 1914) constitute a common measure to investigate the inequality in the income distribution. Several expressions of the Gini index have been proposed in the literature (for a deep investigation see, e.g., Giorgi, 1992, 2019).

An interesting expression for the Gini index in the overall population, G, has been proposed by Lerman and Yitzhaki (1984)

$$G = \frac{2\,\mathrm{cov}\,(y, F(y))}{\mu}, \tag{1}$$

that is, G is equal to twice the covariance between the income y and the rank $F(y)$ divided by the mean income, μ.

When a population can be divided in k $(i = 1, \ldots, k)$ partitions, according to one or more characteristics of the individuals in the population, G can be given by (Yitzhaki, 1994)

$$G = \underbrace{\sum_{i=1}^{k} s_i\, G_i\, O_i}_{\text{inequality within}} + \underbrace{G_b}_{\text{inequality between}}. \tag{2}$$

$s_i = (p_i\, \mu_i)/\mu$, is the ratio between the mean of the income y in the subgroup i, μ_i, multiplied by its share, $p_i = N_i/N$, and the mean of y computed on the whole population, μ, G_i is the Gini index within the subgroup i, O_i is the overlapping index of group i with the entire population and G_b is the between subgroup inequality.

1.1 Overlapping versus stratification

The within inequality is directly influenced by the overlapping term:

$$O_i = \frac{\mathrm{cov}_i\,(y, F(y))}{\mathrm{cov}_i\,(y, F_i(y))}. \tag{3}$$

O_i is defined as the ratio between the covariance of the income y of the units belonging to the subgroup i and its rank in the overall income distribution and the covariance of y and the rank within subgroup i.

Overlapping is a measure of the inverse of stratification (Yitzhaki and Lerman, 1991), that represent the extent one subgroup is overlapped by the other. In fact, stratified society can bear less the inequality because, looking at (2), increases in stratification can cause a negative effect on inequality.

The term O_i moves between $1/p_i$ and $(2 - p_i)/p_i$, where:

- $O_i = 1/p_i$ when subgroup i does not form a strata at all.
- $O_i = 1$ when subgroup i is perfectly stratified.

- $O_i = (2 - p_i)/p_i$ when subgroup i in turn is composed of two subgroups. This is a very extreme case, where group i is not a subgroup at all, but rather is composed of two perfect strata.

The value of O_i is larger as the share of the subgroup is small.

However, the overlapping index O_i related to a given subgroup i in (3) can be written in terms of the overlapping index between two groups, i and j,

$$O_i = \sum_j p_j O_{ij} = p_i O_{ij} + \sum_{j \neq i} p_j O_{ji} = p_i + \sum_{j \neq i} p_i O_{ij}$$

where (Yitzhaki, 1994)

$$O_{ji} = \frac{\operatorname{cov}_j (y, F_j(y))}{\operatorname{cov}_i (y, F_i(y))}.$$

represents the overlapping index of subgroup j by group i and provides a measure of the presence of group j units within the group i. In particular, following (Frick et al., 2006, p. 437),

- $O_{ji} = 0$ no member of subgroup j lies in the range of subgroup i. Subgroup i is a "perfect stratum" because its range is not "contaminated" by members of the j group.
- $O_{ji} = 1$, the distributions of subgroup i and j are identical. In fact, $O_{ii} = 1$.
- $O_{ji} = 2$ if all observations of distribution j are in the range of i and are concentrated at the mean of distribution i.

It is important to point out that O_{ji} is not symmetrical, that is the higher O_{ji} the lower O_{ij}.

1.2 Between subgroups inequality

Another essential element is the measurement of the between-group inequality (G_b). The between-group inequality can be determined as

$$G_b = \frac{2 \operatorname{cov} (\mu_i, \bar{F}_{ui}(y))}{\mu},$$

that is twice the covariance between the mean of variable y of each subgroup and the subgroups' mean rank in the whole population divided by μ. However, G_b is not really a concentration index because it can be negative.

When the population is perfectly stratified the between subgroup inequality is called between subgroup Pyatt inequality, G_b^p (Pyatt, 1976, p. 247). In this case, it can be obtained as the covariance between the mean of each subgroup and the subgroups' mean rank,

$$G_b^p = \frac{2\,\mathrm{cov}\left(\mu_i, \bar{F}_i(y)\right)}{\mu},$$

Yitzhaki and Lerman (1991, p. 322) demonstrated that $G_b^p \geq G_b$.

1.3 The analysis of the Gini index (ANOGI)

From expression (2), with simple algebra, Frick et al. (2006) made explicit four components,

$$\begin{aligned} G &= \sum_{i=1}^{k} s_i G_i + \sum_{i=1}^{k} s_i G_i \left(O_i - 1\right) + G_b^p + \left(G_b - G_b^p\right) \\ &= IG + IGO + BG + BGO, \end{aligned}$$

where IG ($0 \leq IG \leq G$) is the within inequality, BG ($0 \leq BG \leq G$) is the between inequality, IGO the overlapping within and BGO ($-BG - IGO - IG \leq BGO \leq 0$) is the overlapping between, BGO is always negative because the overlapping reduce the ability to distinguish between subgroups.

2 Application

2.1 Data description

The ADSILC dataset is provided by two different sources: information from administrative archives, held by the National Institute of Social Security (Inps), and the Italian database of the European Union Survey on Income and Living Conditions (Eu-Silc), collected by the National Institute of Statistics (Istat). Each individual interviewed in It-SILC cross-sectional waves for the 2004-2017 period has been merged with data driven from Inps archives updated to 2019, using the tax code as key. As a result, the longitudinal information provided by the two different sources include information regarding the socio-demographic and economic status of interviewed (collected in It-SILC survey), enriched with working history of individuals (driven by Inps archives).

2.2 Results

In this section we present an overview of the results of the Gini decomposition (ANOGI) for detecting the impact of education and employment on income inequality (see Table1). The total inequality is principally due to the within component.

When *IGO* and *BGO* are included, the inequality due to between differences account, respectively, 6% and 11% for the education and the employment taken separately. The introduction of the interaction term intensify the spread: the between component represent 3% of total inequality.

Table 2 shows that the average income among the subgroups decline driven by contractual stability. The Gini index (last coloumn) shows, as expected, that among civil servant inequality are very limitated. In the rest of the labour market we observe a spread of income, expecially among unstable and workers with high education. The overlapping index is also used to compare pairs of workers according to their career and level of education (Fig. 2.3). Workers with high educational attainment hired with fixed term contract seems to be category the less stratified with respect to the self-employer and atypical workers.

Table 1 Gini index (*G*) and ANOGI components.

Variables	*G*	*IG*	*IGO*	*BG*	*BGO*
Education	0.437	0.424	-0.013	0.194	-0.168
	(0.004)	(0.053)	(0.003)	(0.014)	(0.056)
Employment	0.437	0.418	-0.028	0.205	-0.158
	(0.004)	(0.053)	(0.029)	(0.006)	(0.077)
Employment * Education	0.437	0.435	-0.012	0.204	-0.190
	(0.004)	(0.042)	(0.025)	(0.008)	(0.055)

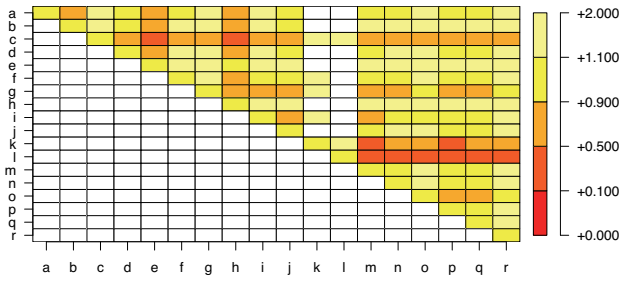
Table 2 Mean income, overlapping index and Gini index of the subgroups according to the employment and the educational level.

	Employement	Education	Mean		O_i		G_i	
			Value	SE	Value	SE	Value	SE
a.	open-ended public	lower sec. degree	28525.39	724.634	0.677	0.019	0.256	0.015
b.	open-ended public	upper sec. degree	31591.67	449.265	0.600	0.014	0.221	0.009
c.	open-ended public	tertiary degree	40606.16	866.526	0.533	0.012	0.315	0.012
d.	open-ended private	lower sec. degree	21391.54	231.744	0.850	0.007	0.304	0.006
e.	open-ended private	upper sec. degree	26549.42	366.187	0.864	0.006	0.348	0.007
f.	open-ended private	tertiary degree	36280.50	811.446	0.804	0.014	0.405	0.012
g.	professionals	lower sec. degree	23961.46	5504.174	1.290	0.072	0.676	0.052
h.	professionals	upper sec. degree	20294.11	1401.532	1.166	0.030	0.542	0.025
i.	professionals	tertiary degree	28959.86	2119.571	1.226	0.018	0.663	0.022
j.	fixed-term workers	lower sec. degree	8078.74	196.362	0.687	0.024	0.482	0.013
k.	fixed-term workers	upper sec. degree	10478.35	436.539	0.890	0.039	0.522	0.014
l.	fixed-term workers	tertiary degree	13036.17	610.536	0.973	0.037	0.499	0.025
m.	self-employed	lower sec. degree	16923.61	378.929	0.893	0.021	0.326	0.014
n.	self-employed	upper sec. degree	20391.33	439.346	0.955	0.013	0.367	0.012
o.	self-employed	tertiary degree	23462.74	1395.353	0.972	0.028	0.393	0.030
p.	atypical	lower sec. degree	24491.23	3105.069	1.261	0.050	0.608	0.041
q.	atypical	upper sec. degree	28936.51	2376.885	1.223	0.027	0.621	0.027
r.	atypical	tertiary degree	22283.25	2108.219	1.244	0.036	0.576	0.022

2.3 Conclusions

Based on ADSILC data we analyze the inequality of income, by separating two aspects: inequality due to education and to the career. Following the ANOGI, the Gini index is decomposed by three components: within, between and overlapping. The latter term shows how intertwined the subgroups are, whereas stratified society can bear the less inequality. Different careers, especially in the private sector, for workers with higher education are associated to the highest level of stratification, according to the overlapping index.

Fig. 1 Overlapping matrix subgroups according to the employment and the educational level.



References

Frick, J. R., J. Goebel, E. Schechtman, G. G. Wagner, and S. Yitzhaki (2006). Using analysis of Gini (ANOGI) for detecting whether two subsamples represent the same universe: The German SOcio-Economic Panel study (soep) experience. *Sociological Methods & Research 34*(4), 427–468.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto veneto di scienze, lettere ed arti 73*, 1203–1248. English traslation In *Metron* 2005, 63(1):3–38.

Giorgi, G. M. (1992). *Il rapporto di concentrazione di Gini: Genesi, evoluzione ed una bibliografia commentata*. Siena: Libreria Editrice Ticci.

Giorgi, G. M. (2019). The Gini concentration ratio: Back to the future. *Rivista Italiana di Economia, Demografia e Statistica 73*(2), 1–18.

Lerman, R. I. and S. Yitzhaki (1984). A note on the calculation and interpretation of the gini index. *Economics Letters 15*(3-4), 363–368.

Pyatt, G. (1976). On the interpretation and disaggregation of gini coefficients. *The Economic Journal 86*(342), 243–255.

Yitzhaki, S. (1994). Economic distance and overlapping of distributions. *Journal of Econometrics 61*(1), 147–159.

Yitzhaki, S. and R. I. Lerman (1991). Income stratification and income inequality. *Review of income and wealth 37*(3), 313–329.

What Causes Juvenile Crime? A Case-Control Study

Cosa Causa la Criminalità Giovanile? Uno Studio Caso-Controllo

Elena Dalla Chiara and Federico Perali[1]

Abstract This work analyses the causes of juvenile crime within a case-control study undertaken in the Italian regions of Veneto and Sicily. We show that family background matters. Parents' education and family income have no effect on crime rates in the Veneto region but are significant risk factors in Sicily. Dropping out of school substantially increases the probability that an adolescent is involved in crime activities. Poor parental relations with children or living in a broken family significantly raises the odds to be in conflict with the law.

Abstract *Questo lavoro analizza le cause della criminalità minorile con uno studio caso-controllo condotto nelle regioni italiane del Veneto e della Sicilia. Lo studio dimostra che il background familiare è importante. L'istruzione dei genitori e il reddito familiare non hanno effetto sui tassi di criminalità nella regione Veneto, ma sono fattori di rischio significativi in Sicilia. L'abbandono scolastico aumenta di molto la probabilità che un adolescente sia coinvolto in attività criminali. Relazioni dei genitori con i figli non buone o il vivere in una famiglia disgregata aumenta significativamente la probabilità che un minore sia in conflitto con la legge.*

Key words: case-control study, juvenile crime, multi-dimensional poverty, single parent

1 Introduction

This research studies the causes of juvenile crime in Italy. We are interested in learning how the income dimension of poverty and other non-material dimensions of

[1] Elena Dalla Chiara, Interdepartmental Center of Economic Documentation (CIDE), University of Verona, email: elena.dallachiara@univr.it
Federico Perali, University of Verona (Italy), Department of Economics and CHILD, email: federico.perali@univr.it

poverty (Dalla Chiara and Perali, 2021) can have a causal role in explaining youth offending. We also inquire whether the identified risk and protective factors may contribute to aggravating or alleviating the incidence of juvenile crime after the occurrence of critical events such as a perduring economic crise or a pandemic emergency (Brandford Wilcox *et al*., 2021; McCarthy *et al*., 2021).

Our young people are leaving the health emergency behind with educational deficits and heavy relational problems within the family. Loneliness, often due to an improper use of social networks, and fear of the other may have affected the personality traits of minors in a delicate developmental phase (Nobili, 2020) thus exposing our young people to the risk of engaging in criminal activities. In general, the crisis has been accompanied by greater poverty and serious gaps in parental control, especially in single parent households, that may transform juvenile crime in a serious public health problem. It is then urgent to design evidence-based policies that may prove effective in preventing and curing juvenile crime in Italy.

To this end, we collected individual and household level data in 2010 about young people who committed an offence before the age of 18 and were still serving their sentences under the supervision of the Juvenile Social Service Offices (USSM), that helped identifying a priori the subjects who had problems with the law, in Veneto and Sicily as representative regions of the social and economic situations in the North and South of Italy. These cases have been compared with a control sample with similar design to better understand what causes juvenile crime in Italy.

2 Methodology

One of the main epidemiological questions that we investigate, controlling for potential confounders, is whether "poverty" causes the risk that young people become offenders, where income or other non-monetary forms of poverty is the exposure and offending is the outcome.

We undertake a case-control study because, differently from, for example, the United States where panel data such as the National Longitudinal Survey of Youth (NLSY) have been collected since 1979 (Levitt and Lochner, 2001), no cohort studies sufficiently large to record the occurrence of a public health problem, such as juvenile crime, is available. Further, the case-control approach allows us to investigate a wide range of exposures. In case-control studies of juvenile crime a group of young offenders ('cases') and a suitable group of subjects without the condition of having experienced problems with the law ('controls') are selected and their exposure to risk factors is ascertained. If the level of exposure among cases and controls is different, it is possible to infer that the exposure may be associated with an increased or decreased occurrence of the outcome of interest.

We administered a questionnaire to young offenders and one to their families to form a sample of 138 cases in Veneto and 178 cases in Sicily to implement a north-south comparison of general policy interest. Though official statistics about population features related to the prevalence of juvenile crime by type of offence are not available, on the basis of sampling statistics typical of rare target populations, as

is the case for the population of young offenders, we maintain that the subsample of young subjects in custody at the Juvenile Social Service Offices (USSM) is representative. The questionnaires have a socio-ecological design collecting information about socio-economic and psychological characteristics, relational aspects and social capital, consumption, income and wealth, intra-household distribution of resources, and time use. We record a wide range of questions relevant to criminal activity, including theft, drug sales, use of force, causing injury to someone or causing property damage. We aggregated these activities in property crimes and crimes of violence that are likely to be determined by social and economic factors accounting also for the fact that youth engaged in one criminal activity are also likely to be involved in other illegal activities.

The integrated design of the sample of cases requires an integrated control sample constructed matching three ISTAT surveys (living standard, consumption and time use surveys) and the Survey on Family Conditions and Social Capital conducted by CISF for the same year of collection of the cases (Dalla Chiara, Menon and Perali, 2019). As control sample we selected families with at least a family member aged 0-19 years with a representative sample of 888 and 313 controls in Veneto and Sicily respectively. The variables in both questionnaires have a common definition.

The socio-ecological design has the objective to identify the causal importance of social factors such as individual characteristics, traits and responsibilities, family background, quality of parenting and their relations with the children, female-headed households, education, youth, and community circumstances along with the degree of "social control" exerted by a community, and economic factors such as income, unemployment, and poverty (Becker, 1968). It was not possible to evaluate some putative factors recorded for the cases such as the cognitive and noncognitive abilities of the young offenders because it was not available in the controls.

The analysis places special emphasis on potential sources of selection bias, confounding and reverse causality. We first analyze odds ratios and then we pose our attention to the estimation of conditional causal odds ratios using logistic regression after adjusting for potential confounding variables (Borgan *et al*., 2018).

The odds ratio is an index defining the cause-effect ratio between two factors. It is defined as the odds of the social problem (disease) between exposed subjects divided by the odds of the problem among non-exposed subjects. If OR=1, the odds of the exposed cases is equal to the odds of the exposed in the control, that is the risk factor does not affect the occurrence of the problem, if OR>1 the risk factor can be a cause of the problem, while if OR<1 the risk factor is a protective factor against the problem. Under the assumption that all members of each group have a comparable risk of becoming an offender, the estimated odds ratio is an average causal effect (Holland and Rubin, 1987) because of the assumed within groups homogeneity. All potential risk factors are analyzed as dummies where 1 is the exposure.

3 Results

Table 1 shows that the estimated odds ratios for the association between crime and potential risk factors vary substantially in the Veneto and Sicily groups. The employment condition of the parents whether they are jointly employed, or only the father or the mother works are significant protection factors, especially in Sicily. The working status is highly associated with the level of education and family income. A low level of parents' education is a risk factor, especially high in Sicily, affecting also children's school attendance and performance as signalled by the OR levels of the variables "dropout" and "low parents' education and dropout" in both regions. Being a single parent (90% in the case sample are mothers as compared to 85% of single parent in the control sample) is also a critical risk factor.

Income poverty is a significant risk factor for all types of crime (violent, property, drug) in both regions. Social capital in the form of trust relations with family members (bonding) is significantly more important than the bridging type of social capital based on the quality of relationships outside the family in both regions. The poor quality of parental relations with children and the scarcity of time spent together is a strong risk factor only for the Veneto region.

Table 1: Odds ratios (OR)

	Veneto			***Sicily***		
Risk factor	***OR***	***95% CI***	***p***	***OR***	***95% CI***	***p***
Parents' occupation						
Parents working full time	0.86	(0.58, 1.26)	0.424	0.42	(0.25, 0.71)	0.000
Working father	0.22	(0.15, 0.33)	0.155	0.16	(0.10, 0.24)	0.000
Working mother	0.73	(0.50, 1.08)	0.589	0.59	(0.38, 0.90)	0.012
Education						
Low parents' education	1.56	(1.06, 2.32)	0.021	5.05	(3.10, 8.42)	0.000
Children's dropout	10.36	(6.48, 16.55)	0.000	8.38	(5.10, 13.91)	0.000
Low parents' educ. & dropout	12.19	(7.25, 20.49)	0.000	11.55	(6.85, 19.74)	0.000
Single parenthood						
Single or no parents	4.28	(2.81, 6.46)	0.000	5.24	(3.02, 9.26)	0.000
Income poverty						
Low family income	4.75	(2.95, 7.56)	0.000	5.10	(3.36, 7.74)	0.000
Violent crime & low income	4.22	(1.45, 10.83)	0.004	5.45	(2.67, 11.42)	0.000
Property crime & low income	6.18	(3.26, 11.38)	0.000	5.09	(2.97, 8.77)	0.000
Drug & low income	3.89	(1.63, 8.56)	0.001	6.82	(2.38, 22.02)	0.000
Social capital & quality of relations						
Trust - Bonding	6.68	(4.31, 10.28)	0.000	9.85	(4.82, 21.64)	0.000
Trust - Bridging	1.77	(1.18, 2.70)	0.004	2.24	(1.46, 3.47)	0.000
Relational WB - Children	12.69	(7.91, 20.29)	0.000	1.72	(1.01, 2.92)	0.036
Relational WB - Time	3.71	(2.50, 5.47)	0.000	1.66	(1.05, 2.60)	0.025

Logistic regression provides the modelling framework to adjust for confounding factors. We select the variables with the most relevant influence in explaining the odds to be in conflict with the law. Table 2 presents the marginal effects computed as discrete changes from the base level for all crime types and property crimes only

for both Veneto and Sicily. We place a special emphasis on property crime because it has the highest prevalence in both regions. Controlling for potential confounding reveals that living in a broken home significantly raises the probability of involvement for all crime types and property crimes. The marginal effect is relatively higher in Sicily. Parents' education and family income have no effect on either general or property crime rates in the Veneto region. This result is consistent with the findings of Levitt and Lochner (2001) for the United States. On the other hand, these factors significantly raise participation rates for both general and property crimes in Sicily. Dropping out of school substantially increases the probability that an adolescent, mainly male, participates both in all crime types and property crime in both regions, as evinced by the coefficients on dropout. Interestingly, poor relations with children significantly raises the probability to be involved in both general and property crime only in the Veneto region. The quality of time spent together does not play a significant role neither in Veneto nor in Sicily.

A few important considerations can be drawn from this logit regression. First, family background matters. Adolescents raised in families where both parents are present are much less likely to engage in crime. This conclusion agrees with the evidence from the US NLSY (Brandford Wilcox *et al*., 2021; Levitt and Lochner 2001), though we can further qualify that the low quality of relations may cause greater criminal involvement as is the case of the Veneto region where families seem to be relatively more fragile. Parents' education has a significant effect on criminal involvement only in Sicily. Adolescents who drop out of school, a factor that increased dramatically during the pandemic, are significantly more exposed to the risk of getting involved with crime.

Table 2: Logistic regression (marginal effects)

	All crime types				***Property crime***			
Risk factor	***Veneto***		***Sicily***		***Veneto***		***Sicily***	
Single or no parents	0.115	***	0.264	***	0.071	***	0.219	***
	(0.031)		(0.062)		(0.023)		(0.072)	
Low family income	0.061		0.159	***	0.042		0.106	***
	(0.032)		(0.042)		(0.023)		(0.040)	
Low parents' education	0.001		0.166	***	-0.015		0.182	***
	(0.018)		(0.044)		(0.014)		(0.039)	
Children's dropout	0.319	***	0.375	***	0.243	***	0.292	***
	(0.048)		(0.051)		(0.047)		(0.056)	
Relational WB - Children	0.300	***	0.101		0.190	***	0.061	
	(0.051)		(0.060)		(0.048)		(0.059)	
Relational WB - Time	0.049	*	0.016		0.032		0.028	
	(0.023)		(0.047)		(0.017)		(0.044)	
Adjusted R^2	0.328		0.301		0.417		0.315	

* p<0.05, ** p<0.01, *** p<0.001

4 Conclusions

The successful development of young people requires that we protect and nurture a set of interrelated physiological, cognitive, and socio-emotional systems especially during enduring critical events such as the pandemic. What happens to children in early life and adolescents can have long-term consequences and may carry over to the next generation. Depending on the timing, transmission mechanisms, and context, the consequences for children's physical, cognitive, and socio-emotional development may be costly and irreversible. The health crisis, accompanied by a reduction in public expenditures for social services, can affect an adolescent through a variety of settings, including the individual, the family, the school, and the community.

In the recovery from the pandemic will then be crucial for communities to invest on household support programs encouraging positive developments of adolescents and healthier parent-child relationships, especially in the North of Italy. Also, target investments in schools or community programs, where many adolescents at risk of dropping out of school can be reached to mitigate some of the negative effects experienced within the family, are strongly recommended. These programs should offer services of multi-systemic therapy involving family, school, and peers, and functional family therapy aiming at reducing risk factors and recidivism. Diversion policies with the objective of rebuilding family and community ties would have a very positive impact on the child, the victim, and the community. Because diversion is cost-effective, it is non-stigmatising and reduces the likelihood of children reoffending, it should have a high priority in the social agenda of the recovery plan.

References

1. Becker, G.: Crime and punishment: An economic approach. *Journal of Political Economy* 76:169–217 (1968)
2. Borgan, Ø., Breslow, N., Chatterjee, N., Gail, Scott, M., Wild, C. (eds.): Handbook of Statistical Methods for Case-Control Studies, CRC Press (2018)
3. Brandford Wilcox, W., Wang, W., Rowe, I.: Less Poverty, Less Prison, More College: What Two Parents Mean for Black and White Children. Institute for Family studies (2021)
4. Dalla Chiara, E., Menon, M., Perali, F.: An Integrated Database to Measure Living Standards. *Journal of Official Statistics* (2019) doi: 10.2478/jos-2019-0023
5. Dalla Chiara, E., Perali, F.: Relational Well-Being and Poverty in Italy in C. Perna, N. Salvati e F. Schirripa Spagnolo, Book of Short Papers SIS 2021:1057-1062. Pearson (2021)
6. Holland, P. W., Rubin, D. B.: Causal inference in retrospective studies. ETS Research Report Series: 203–231 (1987)
7. Levitt, S.D., Lochner, L.: The Determinants of Juvenile Crime. In: Gruber, J. (eds.) Risky Behavior among Youths: An Economic Analysis, pp.327-374. University of Chicago Press (2001)
8. McCarthy, M., Homely, J., Ogilvie, J.: Initial impacts of COVID-19 on youth offending: An exploration of differences across communities, *Journal of Criminology* (2021) doi: 10.1177/00048658211005816
9. Nobili, L.: Impatto psicologico e comportamentale del COVID-19 sui bambini delle famiglie in Italia, Irccs Giannina Gaslini (2020)

Methods and Applications in Survival analysis

Recursive Partitioning for Survival Data

Partizionamento Ricorsivo per Dati di Sopravvivenza

Ambra Macis

Abstract During the years many machine learning methods have been introduced for analyzing survival data. Among these, survival trees are a useful method for defining homogeneous groups according to their survival probability. In this context there are still some unclear points, both related to theoretical and practical issues in model fitting and performance evaluation. The aim of this contribution is to shed light on some of these points.

Abstract *Negli anni diversi metodi di machine learning sono stati introdotti per l'analisi di sopravvivenza. Tra questi, gli alberi di sopravvivenza sono un metodo utile per definire gruppi omogenei sulla base della loro probabilità di sopravvivenza. In questo contesto ci sono ancora alcuni punti poco chiari legati sia a questioni teoriche che pratiche nell'adattamento del modello e nella valutazione della performance. Lo scopo di questo contributo è di fare luce su alcuni di questi punti.*

Key words: Survival data, Machine learning, Recursive partitioning, Survival Trees

1 Introduction

Survival analysis aims to study the occurrence of a specific event (endpoint) during an observed period of time (follow-up) with particular interest on estimation of the survival probability and of the risk of a group of subjects, usually characterized by a set of covariates.

The main feature of survival data is *censoring*. Censoring occurs when the endpoint of interest has not been observed for a certain subject under study during the

Ambra Macis
University of Brescia - Department of Economics and Management, C.da Santa Chiara 50, 25122 Brescia, Italy e-mail: a.macis@unibs.it

observation time; so, the only known thing about a censored subject is the last time he did not experience the event.

During the years statistical models firstly, and machine learning methods later, have been proposed for analyzing survival data. Among the last ones, survival trees, i.e. tree-based algorithms for censored data, are an interesting tool for defining homogeneous groups according to their survival probability.

Through the years many proposals have been advanced, mainly extending Classification and Regression trees (CART) to survival data (see [1]). Most of these works are based on modifications of the splitting criterion and the pruning algorithm using specific tools for right-censored data. Among the most interesting algorithms there are relative risk trees [4] and conditional inference trees [2, 3]. Other extensions have also been proposed over the years (see Table 1 and [5]). Despite up to now many proposals have been presented, there are still some unclear points about the best models to use and about how to use these tools practically. A guide for simulating right-censored data, fitting survival trees and evaluating their performance with the statistical software *R* can be found in [5].

The main aim of this work is to shed light about some of the available packages for fitting and evaluating survival trees for right-censored data, showing the main issues encountered. Particular attention has been given to relative risk trees (RRTs), conditional inference trees based on the *ctree* algorithm (CITs) and conditional inference trees based on the *SurvCART* algorithm (SCTs).

The article is organized as follows. In the next section a brief presentation of the methodological framework is reported. Then, section 3 shows the available *R* packages and the related issues. The paper ends with an application on real data.

2 Survival Trees and Performance Evaluation

Relative Risk Trees. RRTs [4] are an extension of CART based on the Cox proportional hazard model. They are equivalent to Poisson trees, due to the connection between the proportional hazards full likelihood and Poisson model likelihood.
In details, the algorithm splits the covariate space into regions that maximize the reduction in one-step deviance realized by the split. Once a large binary tree is grown, it is then pruned through the classical cost-complexity pruning algorithm of CART [4]. The node summary is the quantity in (1), that can be interpreted as the ratio between observed and expected number of events in each node under the assumption of no structure in survival times:

$$\theta_k = \frac{\sum_{i \in S_k} \delta_i}{\sum_{i \in S_k} \hat{H_0}^1(t_i)}\,, \tag{1}$$

where S_K is the index set of k^{th} node, δ_i is the event indicator (equal to 1 for events and 0 for censoring) and $\hat{H_0}^1$ is an estimate of the cumulative baseline hazard (see [4]).

Conditional Inference Trees. CITs [2] stand on the conditional inference framework. The splitting procedure is divided in variable and split point selection. Firstly, the global null hypothesis of independence between any of the covariates and the outcome variable is tested through a non-parametric permutation test, and the covariate with the strongest association is chosen as splitting variable. Then, the best cut-off is selected through any splitting criterion, e.g. the log-rank test.

Very recently, an alternative way for growing conditional inference survival trees has been proposed [3]. The *SurvCART* algorithm differs from that of CITs for two main aspects: (i) it incorporates a particular distribution for survival times in the parameter instability tests, and (ii) it allows to also consider censoring heterogeneity. So, SCTs are a powerful tool in presence of *conditionally independent censoring*, i.e. when censoring is independent of the time-to-event distribution only conditional on the set of covariates. To this extent, heterogeneity on both the time-to-event and censoring distributions is tested and the most significant variable is chosen. Then, the best cut-off value is selected, usually through the maximization of log-rank statistic, considering the most heterogeneous distribution. The procedure is then repeated until the null hypothesis of homogeneity cannot be rejected.

For both CITs and SCTs the estimated median survival time and the Kaplan-Meier (K-M) survival curve at each terminal node can be easily obtainable.

Performance Evaluation. The most used measures for evaluating performance of survival methods are the concordance index (C-index), the Area Under the Curve (AUC), the Brier score (BS) and the Integrated BS (IBS). The first two are discrimination measures, while the last two measure at the same time discrimination and calibration.

Both C-index, AUC and BS can be evaluated at a single time point. C-index is a ranking measure based on the comparison of each possible pair of subjects with respect to observed times and predicted survival/hazard and it ranges between 0 and 1. An overall C-index can be obtained through a weighted average; the most used weights are those of Harrel and Uno. AUC also measures the discriminating ability of the model. The main differences with classical ROC analysis are censoring and the fact that the outcome of an observation can change over time.

Finally, BS and IBS quantify through a quadratic loss function the distance between observed and predicted outcome. The lower the value the better the performance. IBS is obtained integrating over the time-period the BS at each time point.

For more details and literature review see [5].

3 Growing Survival Trees and Assessing Their Performance in *R*

Many *R* packages allow to fit survival trees. Among these there are *rpart* (with the homonymous function), *partykit* (with the *ctree* function) and *LongCART* (*SurvCART* function) for fitting RRTs, CITs and SCTs respectively. Many other packages exist (see Table 1); however, these are not shown here for the following reasons:

1. *LTRCtrees*, that allows to extend RRTs and CITs to left-truncated and right-censored data, is exactly based on the *rpart* and *ctree* functions. So, in the examined setting (i.e. right-censored data), it provides identical results to the abovementioned *R* functions;
2. *rocTree* can be used for fitting ROC-guided survival trees. It has been seen that it provides a lot of splits and its computation time is slow. The algorithm can be of interest but it seems that it still needs some improvements (e.g. computation speed, clarity of the documentation).
3. *partDSA* implements the PDS algorithm; unfortunately there is little documentation for survival trees implementation.

For what concerns RRTs, CITs and SCTs the respective functions are easy to implement. However, some doubts emerge concerning the interpretation of the node outcome for *rpart*. Indeed, in the available *R* documentation there are not clear examples about survival trees; it is straightforward to deduce that RRTs are grown as Poisson trees, but it is not exactly clear what the "estimated response rate" means in survival analysis. However, it should represent a measure of risk of the node [4]. The other two algorithms, instead, provide very intuitive results. The only drawback for *ctree* is that it is quite complicate to manage the related object. Indeed, there is little documentation about its structure, making it difficult to extrapolate much information about trees structure (useful, for example when performing simulation studies). On the contrary, *SurvCART* is easily manageable and interpretable.

Once a survival tree is grown, its performance can be evaluated. Some of the most used *R* packages are *pec*, *SurvMetrics*, *Hmisc*, *survAUC* and *riskRegression* (see Table 1). The *pec* package differs from all the others because it requires to transform the fitted tree in a compatible *pec* object through some specific functions; while the others only require a vector of predicted values.

The first issue is that both *pec* and *predict* are not compatible with *SurvCART*. Future research will involve the construction of a new function that allows to evaluate its performance. However, it is still possible to obtain a measure of overall concordance for SCTs extracting, for each observation, the estimated median survival time from the *subj.class* matrix of the *SurvCART* object. Moreover, with *predict* it is not possible to estimate survival probabilities at given time points for all tree algorithms; this issue can be overcome through the use of *pec::predictSurvProb*. However, in some specific cases this function returns NA, making it impossible to evaluate many metrics; future research work will involve this aspect.

Furthermore, also in this step, the main issue concerns little documentation about almost all these functions. Indeed, for many of these, for example, it is not specified if risk or survival predicted values are required. This could imply possible wrong evaluation assessments, since, for example, C-index is based on ranking and survival and risk have an inverse trend. It has been empirically verified, using simulated data, that *Hmisc::rcorr.cens* and *SurvMetrics::Cindex* require survival probabilities, while *survAUC::UnoC* requires a measure of risk. If it is not possible to obtain this measure from trees, a tricky solution is to provide as argument the opposite of survival probabilities, maintaining therefore the right ranking (of course also the viceversa holds). Moreover, in many cases, it is not declared which weights are used

Table 1 Packages and functions for fitting survival trees for right-censored data (extensions of CART) and evaluating their performance

R function (package)	Algorithm
rpart (rpart)	Relative risk trees
ctree (partykit)	Conditional inference trees
partDSA (partDSA)	Partition substitution addition algorithm
LTRCART (LTRCtrees)	RRTs[a] for left-truncated and right-censored data
LTRCIT (LTRCtrees)	CITs[b] for left-truncated and right-censored data
rocTree (rocTree)	ROC-guided survival trees
SurvCART (LongCART)	Conditional inference trees for accounting censoring heterogeneity
R function (package)	**Performance measure**
cindex (pec)	Uno's Concordance index
UnoC (survAUC)	Uno's Concordance index
Cindex (SurvMetrics)	Harrel's Concordance index
rcorr.cens (Hmisc)	Harrel's Concordance index
Score (riskRegression)	Time-dependent AUC[c] with IPCW[d]
AUC.hc (survAUC)	Time-dependent AUC[c]
Brier (SurvMetrics)	Brier Score at a single time-point with Kaplan-Meier IPCW[d]
pec (pec)	Brier Score
IBS (SurvMetrics)	Integrated Brier Score with Kaplan-Meier IPCW[d]
pec (pec)	Integrated Brier Score

[a] Relative risk trees; [b] Conditional inference trees with *ctree* algorithm; [c] Area Under the Curve; [d] Inverse probability censoring weights

for evaluating C-index. After an in-depth analysis it has been verified that *SurvMetrics::Cindex* and *Hmisc::rcorr.cens* evaluate Harrel's C-index; while *pec::cindex* and *survAUC::UnoC* provide Uno's version.

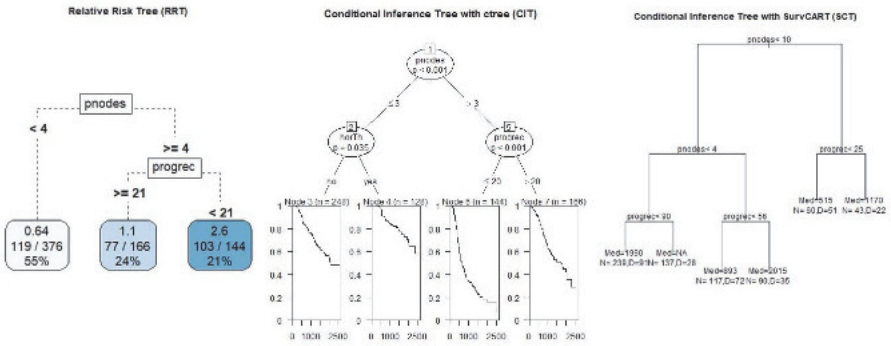
Time-dependent AUC can be evaluated by using *riskRegression::Score* and *survAUC::AUC.hc*; the two functions usually provide similar results.

Finally, BS can be evaluated with *SurvMetrics::Brier* and *pec::pec*; while IBS can be evaluated by using *SurvMetrics::IBS* and pec::pec functions. After the analysis of the *SurvMetrics* functions code (no information was provided in the documentation) it has been verified that it evaluates the index by only using K-M estimator for obtaining the inverse probability censoring weights (IPCWs). Differently, *pec* allows the use of different censoring models.

For both functions an issue related to estimated survival probability was encountered. Indeed, as already hinted, *predictSurvProb* returns NA when only events occurred in a node, implying the impossibility of estimating BS and IBS after that time point. The solution to this problem is deferred to future research work.

A case study. A comparison of the overall performance of RRTs, CITs and SCTs has been carried out on the dataset collected by the German Breast Cancer Study Group [6]. All the three methods identify *pnodes* as the most important variable, followed by *progrec*. Only CIT also identifies *horTH* as a splitting variable. A higher number of positive nodes, lower levels of the progesterone receptor and not receiving an hormonal therapy lead to a higher risk of recurrence. As already hinted, for

SCTs only overall concordance measures could be obtained. In particular, for evaluating C-index for both CIT and SCT, estimated median survival times greater than the length of follow-up were artificially fixed to a maximum value (see [5]). The three methods lead to similar results both in terms of interpretation and goodness-of-fit (Figure 1).



	RRT	CIT	SCT
Harrel's C	0.686	0.672	0.688
Uno's C	0.614	0.553	0.608
IBS	0.165	0.165	

Fig. 1 Performance comparison of RRT, CIT and SCT. Data: German Breast Cancer Study dataset

Acknowledgements I would like to thank Prof. Paola Zuccolotto and Dr. Marco Sandri for their precious help in developing this research work.

References

1. Bou-Hamad I., Larocque D., Ben-Ameur H.: A review of survival trees. Stat. Surv. **5**, 44–71 (2011)
2. Hothorn T., Hornik K., Zeileis A.: Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph. Stat. **15:3**, 651–674 (2006)
3. Kundu M.G., Ghosh S.: Survival trees based on heterogeneity in time-to-event and censoring distributions using parameter instability test. Stat Anal Data Min. **14:5**, 466–483 (2021)
4. LeBlanc M., Crowley J.: Relative risk trees for censored survival data. Biometrics **48:2**, 411–425 (1992)
5. Macis A.: A practical guide for simulations in R with right-censored data. Submitted (2022)
6. Schumacher M., Bastert G., Bojar. H, Hübner K., Olschewski M., Sauerbrei W., Schmoor C., Beyerle C., Neumann RL., Rauschecker HF.: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. J. Clin. Oncol. **12:10**, 2086–2093 (1994)

Detecting survival patterns in a digital learning platform

Rilevare i modelli di sopravvivenza in una piattaforma di apprendimento digitale

Marta Cannistrà, Mara Soncin and Federico Frattini

Abstract In an ever-changing world, having the right competences for the job market represents the key challenge. Hence, the number of digital platform for learning is increasing. Anyway, less is known about the learners' profiles or the time spent on it. The present work aims at filling this gap, by proposing a deep analysis to identify latent subgroups of learners with similar behaviour within an online learning platform. Then, the identified subgroups are described in terms of personal features and survival profiles. From findings, it is observed the emerging of three distinctive latent classes, with very different survival profiles. The analysis allows to better personalize contents according to the type of learner in order to support him/her to conclude the assessment and "survive" the platform, acquiring the right skills for the job market.

Abstract *In un mondo in continua evoluzione, avere le giuste competenze per il mercato del lavoro rappresenta la sfida chiave. Di conseguenza, il numero di piattaforme digitali per l'apprendimento è in aumento. Ad ogni modo, si sa meno sui profili degli studenti o sul tempo impiegato su di esso. Il presente lavoro mira a colmare questa lacuna, proponendo un'analisi approfondita per identificare sottogruppi latenti di studenti con comportamenti simili all'interno di una piattaforma di apprendimento online. Quindi, i sottogruppi identificati sono descritti in termini di caratteristiche personali e profili di sopravvivenza. Dai risultati, si osserva l'emergere di tre classi latenti distintive, con profili di sopravvivenza molto diversi. L'analisi permette di personalizzare al meglio i contenuti in base alla tipologia di discente in modo da supportarlo nel concludere la valutazione e "sopravvivere" alla piattaforma, acquisendo le competenze giuste per il mercato del lavoro.*

Key words: Digital Education; Survival Analysis; Latent Class Analysis

[1] Marta Cannistrà, Politecnico di Milano; email: marta.cannistra@polimi.it
Mara Soncin, Politecnico di Milano; email: mara.soncin@polimi.it
Federico Frattini, Politecnico di Milano; email: federico.frattini@polimi.it

Introduction

The necessity of better aligning job requirements and education is increasingly topical, mainly due to the wide spread of digitalisation, which is transforming all sectors. Hence, job market requires a continuous competences adaptation to stay on track and be competitive. To strengthen this perspective, the Digital Agenda for Europe 2020 described the principles for ensuring the acquisition of digital skills and literacy for all citizens (Levano-Francia et al., 2019). The Future Jobs Report, prepared by the World Economic Forum (2018) predicts that the large number of professions that exist today and in the coming years will require digital skills to be able to perform their work. In this view, educational institutions are called to align their academic offer with labor market needs to increase employability, stay competitive and strengthen the cooperation between university and industry as one of the main drivers of innovation.

Technological advancements enabled the delivery of educational digital platforms that can keep the pace of such a fast and evolving need of knowledge. However, it is documented that dropout rates of e-Learners are substantially higher than in-presence counterparts (Levy, 2007). In some case, the main predictor of learners' retention in online platforms is the general level of satisfaction, which represents a proxy of their engagement. Hence, the actual challenge is to understand learning paths and learners' behaviours in digital platforms, aiming at reaching higher retention and pursuing professional and academic goals of the learners.

The present work aims at contributing to the existing academic literature about learners' dropout in online educational platforms. The research objectives concern (i) the identification of subgroups of learners with similar digital behaviours, and (ii) test whether subgroups differ in their survival curves in the platform. Identifying early signals of disengagement before dropout would be relevant to foster the educational potential of personalised digital platforms.

The personalized digital platform

The digital platform analysed has been developed in 2019 by the Business School of a leading Italian technical university. The "digital mentor" allows learners to get closer to settled professional goals along a continuous, personalised learning pathway. Before starting, every learner is asked to fill a short questionnaire about their actual job, role and competences and set the aspirations for future professional career. According to settled goals, aspirations and market job requirements, the digital mentor, based on AI algorithms, provides tailored suggestions and contents to explore. The typical learning pathway lasts for six months, after which the learner can set new learning goals. The features of the platform include periodical skills assessment, qualified training, challenges, networking with the whole digital community, and job postings from selected companies. The platform proposes three kinds of activities: daily workout, structured as "knowledge "nuggets", learning

pathway, with a long-run perspective, and search bar, where learner may find contents by themselves. The digital mentor allows every learner to reach the settled goals within six months from the registration. This period will be then used as reference to evaluate learner's success (or not).

Data and Methods

Data are extracted from the platform and, after the cleaning processing, the number of observations (learners) is 2403. Variables can be categorised into three groups: (i) platform-learner interaction, (ii) learners' characteristics and (iii) learner survival time within the platform. The latter describes the learning path in the platform: the status defines whether the learner completed the learning pathway after six months from the registration and, if he/she dropped out before, when the event occurs. This information is essential to study the survival function of latent classes, in line with the second research objective.

Two different methodologies are adopted along the study. First, a Latent Class Analysis is used with the aim of detecting hidden patterns in learners' interaction with the platform and classify them into latent subgroups (Muthén & Muthén, 2000). By latent subgroups, we mean clusters of learners, who share underlying common features in terms of important dimensions (in our case, the level of interaction and the type of activity in the platform). Firstly, the latent classes are defined based on indicators about learner-platform engagement. As a result of step 1, each observation is assigned to the group for which the probability of class membership is the highest. The number of classes is selected assessing the goodness of fit of the model through the Bayesian Information Criterion (BIC) and the Lo-Mendell-Rubin (LMR) test (Lo et al., 2001). In step 2, multinomial regression is run to characterise the latent classes by means of personal information of learners.

The second methodology is the survival analysis (Ameri et al. 2016), employed with the aim of studying the survival curves of the identified latent subgroups. The Kaplan-Meyer functions, $S(t_i)$, are estimated to visualise the probability of "surviving" on the platform with time-dynamics. This is a step-function, a non-parametric estimation characterised by discontinuities at points given by time events: $S(t_i) = S(t_{i-1}) * (1 - \frac{d_i}{n_i})$. $S(t_{i-1})$ is the probability of being alive at t_{i-1}, n_i is the number of active learners just before t_i and di is the number of events at t_i. The event is defined when at least one learner drops-out (not survive). The underlying assumption is that the drop-out events are uncorrelated.

Results

The number of latent classes that proxy our data at best is equal to three, as indicated by the lowest value of the BIC, and the platform-learner interaction profile of each class is presented in Figure 1. The first class is composed by a low share of learners (6.0%) and characterised by a high profile of activity. Their activities mainly relate to daily workout sessions, thus they are often logged in as confirmed by the high number of sessions with a relatively short duration. Also, they open and like contents proposed by the digital mentor, and for this reason can be labelled as *Platform-engaged learners*. The second class, represented by the 20.5% of the learners, report a high number of sessions, but their activity is mainly characterised by searching contents by themselves. For this reason, they can be labelled *Self-engaged learners*. Lastly, 73.5% of learners interact very poorly with the platoform. They are the least active and log in few times for long time, thus they can be labelled *Disengaged learners*.

The multinomial regression in Table 2 provides a picture about some of classes' characteristics. The probability of being a *Self-engaged learner* is higher for young learners, compared to the other groups of learners. For all the other characteristics, *Self-engaged* and *Platform-engaged* learners are rather similar, while *Self-engaged* and *Disengaged learners* differ in terms of citizenship (Italians are less likely to be part of the latter) and assessment status – *Disangaged learners* are more likely to have not even started or completed the initial skill assessment. In this respect, *Disengaged learners* tend to be so since their enrolment in the platform. Finally, having followed one or more courses at the Business School does not affect the probability of belonging to a specific class of learners.

In order to address the second research objective, the Kaplan Mayer (KM) curves, depicted in Figure 2, give a visual estimate of how fast the survival probability of each latent class decreases over time. Results suggest that *Disengaged learners* have a completely different survival function compared to the other two classes. The curve registers an initial drop at time = 0, meaning that many learners log in the platform one time to never come back. The other two KM curves are very similar each other, even though *Platform-engaged learners* show an initial drop, too. Interestingly, the survival probability of *Self-engaged learners* is higher than that of *Platform-engaged* learners until the 140th day after registration, when the curves overlap and then switch. It is worth to note that the probability of "surviving" until the end of the learning path (after six months) is quite low (<25%) for all the three classes.

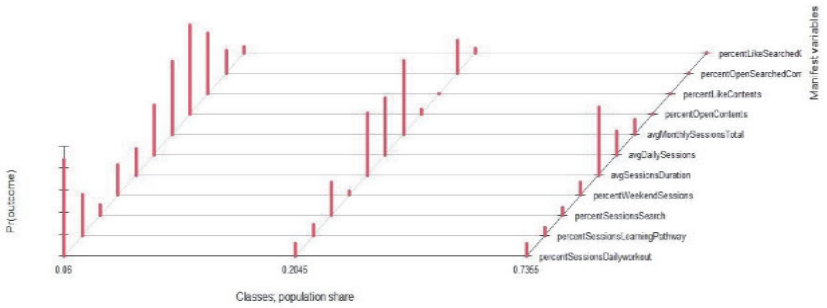


Figure 1: Latent Classes' profiles

Table 2. Multinomial regression to characterise latent classes

	Dependent variable (ref.: Disengaged learners)	
	Platform-engaged learners	Self-engaged learners
age	-0.027*** (0.007)	0.003 (0.011)
Italian	0.304** (0.149)	0.495* (0.257)
learnerTypeStudent	0.748*** (0.136)	0.974*** (0.227)
assessmentStatus: not completed	-0.589*** (0.198)	-1.313*** (0.429)
assessmentStatus: not started	-1.502*** (0.176)	-2.068*** (0.350)
mipEducations_dummy	-0.242 (0.592)	0.029 (1.061)
Constant	-0.779 (0.656)	-3.358*** (1.151)
Akaike Inf. Crit.	2,903.671	2,903.671

Note: *p<0.1; **p<0.05; ***p<0.01. Reference category: Disengaged learners.

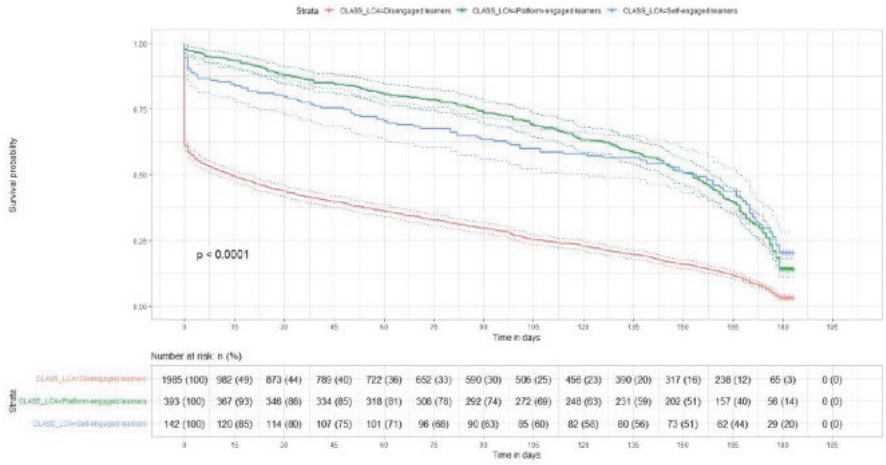


Figure 2: Kaplan-Meyer curves of identified Latent Classes.

Conclusions

The paper explores data derived from a digital learning platform based on AI algorithms to explore patterns of platform-learner interaction and observe the probability to persist on the platform until the end of the learning pathway. The high drop-out rate and risk of disengagement is a critical issue in online education (Levy, 2007; Bañeres Besora & Conesa Caralt, 2017) that may undermine the effectiveness even of highly personalised learning experiences (Moreno-Marcos et al. 2020).
Findings suggest that most learners (73.6%) do not even take a step further after registering, a behaviour observed also among MOOC learners (Agasisti et al., 2021). The paper contributes to the academic debate on the identification of online learners' profiles (Binali et al. 2021), showing that despite the high personalisation of contents, learners show high levels of disengagement, as observed in other online formats (Korableva et al., 2019). In terms of practitioners' implications, findings suggest possible early warning signals that may help to prevent learners' drop-out, as disengaged students tend to show their attitude since the very beginning – and, for instance, are more likely to skip the initial steps of skills assessment.

Citations and References

1. Agasisti, T., Azzone, G., & Soncin, M. (2021). Assessing the effect of Massive Open Online Courses as remedial courses in higher education. Innov. Educ. Teach. Int., 1-10.
2. Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. In Proc. 25th ACM Int. Conf. Inf. Knowl. Manag. (pp. 903-912).
3. Binali, T., Tsai, C. C., & Chang, H. Y. (2021). University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification. Comput. Educ., 175, 104315.
4. Korableva, O., Durand, T., Kalimullina, O., & Stepanova, I. (2019). Studying user satisfaction with the MOOC platform interfaces using the example of coursera and open education platforms. In Proc. 2019 Int. Conf. Big Data Educ. (pp. 26-30).
5. Levano-Francia, L., Sanchez Diaz, S., Guillén-Aparicio, P., Tello-Cabello, S., Herrera-Paico, N., & Collantes-Inga, Z. (2019). Digital Competences and Education. J. Educ. Psychol., 7(2), 579-588.
6. Levy, Y. (2007). Comparing dropouts and persistence in e-learning courses. Comput. Educ., 48(2), 185-204.
7. Lo Y., Mendell N. R. and Rubin D. B. (2001) Testing the number of components in a normal mixture. Biom.88(3): 767–778.
8. Moreno-Marcos, P. M., Munoz-Merino, P. J., Maldonado-Mahauad, J., Perez-Sanagustin, M., Alario-Hoyos, C., & Kloos, C. D. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. Comput. Educ., 145, 103728.
9. Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. Alcohol.: Clin. exp. res., 24(6), 882-891.
10. World Economic Forum. (2018). The Future of Jobs Report 2018. Geneva: World Economic Forum.

An extension of proper Bayesian bootstrap ensemble tree models to survival analysis

Un'estesione degli alberi ensemble con bootstrap Bayesiano proprio all'analisi di sopravvivenza

Elena Ballante

Abstract In this work, a new bagging survival tree model is proposed. An exension of Efron's bootstrap procedure, that is usually implemented in bagging methods, is integrated in the classical bagging survival tree model. The proper Bayesian bootstrap allows to enrich the original feature space with new observations sampled from a prior distribution, that are not already present in the original data.
Empirical results are shown in a simulated study. The proposed model reaches competitive performances with respect to classical survival models (Cox model and survival random forest) in terms of integrated Brier score with higher stability, especially when small sample size are involved.
Abstract *In questo lavoro viene proposto un nuovo modello ensemble basato su alberi decisionali. Un'estensione del bootstrap di Efron è stata integrata all'interno del modello di bagging di alberi di sopravvivenza. L'utilizzo dell'algoritmo di bootstrap Bayesiano proprio permette di arricchire lo spazio delle variabili con nuove osservazioni che vengono campionate da una prior e che non sono già presenti all'interno del dataset originale.*
I risultati empirici sono mostrati grazie all'analisi di dati simulati. Il modello proposto mostra performance competitive rispetto a modelli di sopravvivenza classici (modello di Cox e survival random forest) in termini di Brier score integrato e una maggiore stabilità, soprattutto quando si analizzano dati con una bassa numerosità.

Key words: Survival analysis, Bootstrap, Bayesian nonparametric learning, ensamble models

Elena Ballante
Department of Political and Social Sciences, University of Pavia, Italy e-mail: elena.ballante@unipv.it

1 Introduction

In this paper we extend the Bayesian non-parametric learning applied to ensemble tree modeling defined in [1] to the analysis of right-censored survival data.
As shown in [1], that the application of different resempling techniques, as the Proper Bayesian Bootstrap [2], in ensemble tree modeling can improve the stability of the model with competitive prediction performance. On the basis of the promising results in classification and regression frameworks, time-to-event data analysis can benefit from the introduction of this new model.
Bootstrap resempling techniques are applied to approximate the posterior distribution of a statistic of interest, that for ensemble tree models is the statistical functional $\Phi(F)$, where F is a random distribution function as defined in Proper Bayesian Bootstrap [2] and $\Phi(F)$ is a decision tree.
This methodological proposal inherits the main advantages of Bayesian nonparametric learning such as the flexibility and the computational strength, considering also prior opinions and thus overcoming the main drawbacks of Efron's and Rubin's bootstrap procedures ([3], [4]) used in classical ensemble decision tree models.
Coupled with these considerations, we want to underline that survival data are commonly analyzed using methods that rely on restrictive assumptions, e.g. proportional hazards. Most of the models considered are parametric (or semi-parametric as Cox model), with all the implied assumptions and limitations.
Some nonparametric models are defined for survival data: survival tree models are defined in [5] and survival random Forest is proposed in [6], deploying Efron's bootstrap technique in the algorithm.
Based on these considerations, we underline the importance of developing and applying non parametric innovative models that can lead to higher performances in survival analysis.

The rest of the paper is organized as follow: Section 2 describes the methodological proposal, Section 3 describes the computational setting and the preliminary results of the study and Section 4 contains the conclusions.

2 Proposed model

The purpose of this paper is to give a first definition of Proper Bayesian bootstrap ensemble tree model for survival data analysis. The Proper Bayesian bootstrap defined in [2] is considered and adapted to survival data analysis.
The statistic of interest of ensemble tree models is $\phi(F,\mathbf{X})$, which is a decision tree model based on a distribution F and on observed data $\mathbf{X}$. As well explained in [1], an estimation of the posterior distribution of ϕ can be obtained through bootstrap procedures. Each tree is obtained fitting the model on the drawn weighted dataset and the obtained predictions are an approximation of posterior mean.
To apply the Proper Bayesian bootstrap we define a prior distribution $D(kF_0)$ for F,

which is a Dirichlet process. In order to explain a response variable y given a list of P covariates $x_1,...,x_P$, the parameter F_0 of the Dirichlet process is a joint distribution depending both on $(\mathbf{x},y)$.
The algorithm that describes the sampling method from the posterior of $\phi(F,\mathbf{X})$ is defined in [1] and reported in this work as Algorithm 1.

Input: Training set T
for *b in 1:B* **do**
  Sample $(\mathbf{x}_1^*,y_1^*),...,(\mathbf{x}_m^*,y_m^*)$ from $(k+n)^{-1}(kF_0+nF_n)$;
  Draw $\mathbf{w}^b$ from $D(\frac{n+k}{m},...,\frac{n+k}{m})$;
  Get $\phi^b=\phi(\mathbf{w}^b)$ running weighted tree on the new sample $\mathbf{x}_1^*,...,\mathbf{x}_m^*$
end

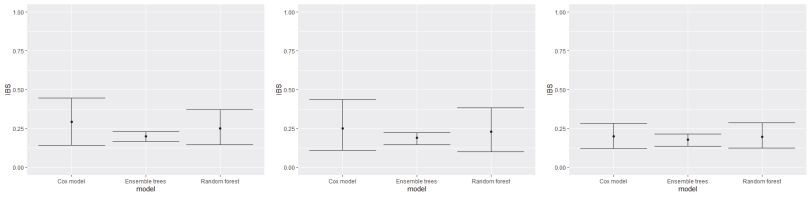
Algorithm 1: Proper Bayesian Bootstrap in ensemble tree models

The bootstrap resample $(\mathbf{x}_1^*,y_1^*)...,(\mathbf{x}_m^*,y_m^*)$ is generated by a mixture of distributions of the prior guess F_0 and the empirical distribution F_n. Since the covariates are assumed to be independent, when a new observation of bootstrap resample is generated from F_0, a new vector of covariates $\mathbf{x}$ is generated from the original prior distributions $F_0(x_k)$, then the new value of the response variable is associated on the basis of the chosen prior distribution.
The main differences introduced for the application to the survival analysis are the two: the response variable associated to the vector of covariates generated from the prior is obtained with a suitable survival model and the aggregation method of the predictions should be performed taking into account the nature of the time-to-event data.
More specifically for this second point, the output of the model is a bootstrap aggregated version of the estimted conditional survival function S for a new observation $\mathbf{x}_{new}$ computed by $\hat{S}_A^B(\cdot|\mathbf{x}_{new})=\hat{S}_{L_A^B(\mathbf{x}_{new})}(\cdot)$, as suggested in [7].
The main element of novelty with respect to classical ensemble procedure, is that the prior F_0 allows to generate new observations, not contained in the training set, which can enrich our prediction model. The obtained bootstrap samples are less dependent one to each other, reducing overfitting and leading to a more stable method, as shown in Section 3. For these reasons we think that the proposed model can guarantee more stable results and, in some settings, better performances than the classical models.

3 Experimental setting and preliminary results

A simulation study is performed in order to show the potential of the method proposed with different sample size and different weights assigned to the prior distribution F_0. The simulated datasets are composed of 5 numerical covariates and a time-to-event target variable where a 20% of censored observations are considered. The simulation of the data is performed by the flexible-hazard method as described

in [8]. The sample size is set as $N = 30$, $N = 50$ and $N = 100$.
The prior for each covariate is set as a uniform distribution on the range of that covariate in the original dataset and the parameter k is set such that the weight $w = \frac{k}{k+n}$ assigned to the prior F_0 is equal to 0.25.
The prior relation between y and $\mathbf{x}$ is evaluated with an exponential regression model.
The proposed model is compared with the most common models in the survival analysis: the Survival Random Forest [6] and the Cox model [9].
Prediction performance was evaluated in terms of Integrated Brier Score (IBS) in a 5-fold cross validation exercise. For each setting, 100 datasets are generated. Mean values and nonparametric confidence intervals of the resulting IBSs are presented.



(a) N=30 (b) N=50 (c) N=100

Fig. 1: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross validation for the 100 simulated datasets in three different setting of sample size.

From the results presented in Figure 1, we can see that IBS average is lower for the proposed method, but not significantly lower considering the width of the confidence interval for classical methods.
The difference between Proper Bayesian Bootstrap ensemble trees and the other two methods becomes less evident increasing the sample size. The higher quantity of information already present in a bigger dataset make the impact of the prior sampled observation less evident.
The most evident result is that the confidence intervals of the proposed model is much smaller than the classical methods, underlying that our proposal is more stable in terms of prediction performance.

4 Conclusion

In this paper we present a new ensemble tree model based on Proper Bayesian Bootstrap and we show its potential with respect to classical survival models. The model proposed shows in a simulated framework higher stability coupled with competitive performances.

The introduction of synthetic data generated from a prior distribution, that were not already present in the original dataset, makes the final ensemble model more stable, especially with small sample size datasets. Further investigation are planned, as a sensitivity analysis on different percentage of censored data, different values of k (the confidence assigned to the prior F_0) and a comparative analysis for the different priors chosen for covariates.
The analysis of performance on real datasets is another fundamental point that will be investigated in a future work.

References

1. Galvani, M., Bardelli, C., Figini, S., Muliere, P.: A Bayesian Nonparamtric Learning Approach to EnsambleModels Using the Proper Bayesian Bootstrap. Algorithms **14**(1), 11 (2021) doi: 10.3390/a14010011
2. Muliere, P., Secchi, P.: Bayesian Nonparametric Predictive Inference and Bootstrap Techniques.Annals of the Institute of Statistical Mathematics. **48**(4), 663–673 (1996)
3. Efron, B.: Bootstrap methods: another look at the Jackknife, Annals of Statistics, **7**(1), 1–26 (1979).
4. Rubin, D. B.:The Bayesian Bootstrap, Annals of Statistics, **9**(1), 130–134 (1981).
5. Gordon, L. and Olshen, R.A.: Tree-structured survival analysis. Cancer treatment reports **69** (10) 1065–1069 (1985)
6. Ishwaran, H., Kogalur, U.B., Blackstone,E.H., Laure, M.: Random Survival Forest. The Annals of Applied Statistics **2**(3), 841–860 (2008) doi: 10.3390/a14010011
7. Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M.: Bagging survival trees. Statist. Med., **23**, 77-91 (2004). https://doi.org/10.1002/sim.1593
8. Harden, J. J. and Kropko, J.: Simulating Duration Data for the Cox Model. Political Science Research and Methods (2018) https://doi.org/10.1017/psrm.2018.19
9. Andersen, P. and Gill, R.: Cox's regression model for counting processes, a large sample study. Annals of Statistics **10**, 1100-1120 (1982).

Modelling time to university dropout by means of time-dependent frailty COX PH models

Modelli di COX tempo-dipendenti con frailty per la modellizzazione del tempo all'abbandono universitario

Mirko Giovio, Paola Mussida, Chiara Masci

Abstract University dropout is studied worldwide, with the aim of identifying its determinants and of preventing it. In this study, we model and predict the time to dropout of university students, following a survival analysis approach. We analyse data about 50,000 students, enrolled in 18 different bachelor of science in Engineering at Politencico di Milano. Given the potential heterogeneity in the dropout determinants over time and the two-levels structure of the data (students nested within degree programmes), we applying a Cox PH model that includes time-varying covariates, to track students careers over time, and a frailty, to take into account the heterogeneity across degree programmes. The strength of the method is twofold: to describe the dropout over time and to identify a good trade-off between a high prediction accuracy and the development of an early warning system.

Abstract *Il dropout universitario è un fenomeno studiato in tutto il mondo, con l'obiettivo di caratterizzarlo e prevenirlo. In questo studio, analizziamo il tempo all'abbandono univeristario, seguendo un approccio di analisi di sopravvivenza. Analizziamo le carriere di 50,000 studenti immatricolati in 18 corsi di studio di Ingegneria del Politencico di Milano. Viste le potenziali differenze nelle determinanti del dropout nel tempo e la struttura annidata dei dati, applichiamo un modello di COX che include covariate tempo-dipendenti, per tracciare la carriera degli studenti nel tempo, e una frailty, per modellizzare la variabilità tra corsi di studio. La forza del metodo proposto è duplice: descrivere le dinamiche del dropout nel tempo e identificare un giusto trade-off tra una previsione accurata e una previsione precoce.*

Key words: Student dropout, Survival analysis, COX frailty model, time-dependent covariates.

Mirko Giovio, Chiara Masci
MOX - Modelling and Scientific Computing, Politecnico di Milano, 20133 Milano (IT). e-mail: mirko.giovio@mail.polimi.it, chiara.masci@polimi.it

Paola Mussida
DEIB, Politecnico di Milano, 20133 Milano (IT). e-mail: paola.mussida@polimi.it

1 Introduction

The Italian Higher Education (HE) system measures a high level of dropout, with many students abandoning their studies during the Bachelor. According to ANVUR, the dropout rate is around 28.2%, with almost 2/3 of them dropping out in the first two years [3]. Many studies aim at individuating personal features of students who are more likely to drop in order to act as soon as possible and partially prevent the phenomenon, and most of them use a classification approach [1, 4]. Recent evidences show that university dropout might be very heterogeneous in time, in its determinants over time and across universities and degree programmes [5]. For these reasons, we propose a survival analysis approach [6] to university dropout in which the target variable is the *time to dropout*. We analyse data about 50,000 careers of students enrolled in 18 different bachelors of science in Engineering at Politecnico di Milano (PoliMI). PoliMI dropout rate in engineering is around 30%, with the majority of students dropping out during the first year [5]. PoliMI collects student-level data that regard (i) personal information at the enrolment, (ii) the career status and career track over time of the student, continuously updated, (iii) and the degree programme in which the student is enrolled. To model the time to dropout, we adopt a Cox Proportional Hazard (PH) model [7] in which the response variable is the time to dropout of each student within three years after the enrolment, if any. Given the hierarchical structure of the observations and in order to model the heterogeneity given to it, we include a random frailty in the Cox PH model. Moreover, we consider the extension of the frailty Cox PH model to handle time-varying covariates, in order to take into account the credits and the grade point average (gpa) of studens, updated each semester, in the prediction of the time to dropout.

Survival analysis has already been applied in the framework of student dropout [2, 8], but, to the best of our knowledge, this is the first study that a frailty Cox PH model with time-varying covariates is adopted is this context. The advantage of this modelling regards the possibility to explore the latent heterogeneity across degree programmes and to evaluate the improvement in the model predictive power due to the update of information semester by semester.

The aim of this study is twofold: to accurately describe the dropout phenomenon and its heterogeneity and to develop an early warning system to predict the time to dropout of the enrolled students. In this perspective, we put a special focus on the effect of the inclusion of student career information over time on the model predictive power in order to identify the best trade-off between a very accurate prediction and the development of an early warning system.

2 Dataset

The dataset of interest is collected by PoliMI and contains information about 56,970 students (careers) enrolled in a bachelor of science in Engineering at PoliMi, between a.y. 2010/2011 and 2020/2021. 32% of the careers are active, 19% are con-

cluded with a dropout and 49% are concluded with a graduation. Our target variable is the time to dropout, expressed in semester, measured from the enrolment. In this analysis, we exclude from the sample those students who dropped out during the first semester[1]. Student-level variables are collected at the enrolment and during the career. As time-invariant variables collected at the enrolment, we consider the gender, the age, the type of previous studies, the high school grade, the family income, the origins and the PoliMi admission score. As time-varying variables, updated each semester, we consider the gpa and the cumulative number of credits obtained. The observed time to dropout is distributed between 1 and 6 semesters, with a median dropout time of 2.2 semesters and a mean dropout time of 3.1 semesters.

3 Methodology

Survival Analysis is the process that studies the time that passes from a particular moment until a certain event occurs [6]. If for an individual the event never occurs (during the follow-up period), the observation is considered as censored. For each individual $i = 1,\dots,N$, we call T_i^* the event time (observed if the event occurs) and C_i the censoring time (observed if the event does not occur).

The target variable is therefore identified by $D = \{(T_i, \delta_i), i = 1,\dots,N\}$ where $T_i = min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$ is the indicator variable that indicates whether the $i-$th observation is censored or not.

Letting T denote the non-negative random variable of survival time with probability density function $f(t)$ and distribution function $F(t) = P(T \leq t)$, we define the *Survival function* $S(t)$ as the probability to survive until time t and the *Hazard function* $h(t)$ as the instantaneous risk of event at time t. In particular,

$$S(t) = P(T > t) = 1 - F(t)$$

and

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Given these quantities, the frailty Cox PH model with time-invariant and time-varying covariates assumes the Hazard function $h_{ij}(t)$ of the $i-$th individual within the $j-$th group to be expressed as follows:

$$h_{ij}(t|X(t)) = h_0(t) \times w_j \times \exp\left(\sum_{p=1}^{P_f} x_{ijp}\beta_p + \sum_{v=1}^{P_v} z_{ijv}(t)\alpha_v\right)$$

[1] In the perspective of modelling and predicting the time to dropout of students observing at least first semester information, we need to exclude first semester dropout. A separate and specific analysis can be conducted on this sample of students, for which the dropout determinants are potentially various and unpredictable.

where

- $h_0(t)$ is the baseline hazard function
- $\mathbf{x}_{ij} \in \mathscr{R}^{P_f}$ is the vector of P_f time-invariant covariates relative to the $i-$th observation within $j-$th group and $\beta \in \mathscr{R}^{P_f}$ is the vector of coefficients
- $\mathbf{z}_{ij}(t) \in \mathscr{R}^{P_v}$ is the vector of P_v time-varying covariates relative to the i-th observation within $j-$th group and $\alpha \in \mathscr{R}^{P_v}$ is the vector of coefficients
- $w_j \sim Gamma(1,\theta)$ is the shared frailty of $j-$th group.

4 Results

In the application of the frailty Cox PH model to the PoliMi dataset, the student dropout represents the event of interest and the Survival probability function $S(t)$ indicates the probability of not dropout until time t. We divide the sample into a training (70%) and a test set (30%) and we run four frailty Cox PH models: *Model 1* considering only covariates measured at the enrolment, *Model 2* considering covariates measured at the enrolment and at the end of first semester, *Model 3* considering covariates measured until the end of second semester (i.e., including time-varying covariates measured at the end of first and second semesters) and *Model 4* considering covariates measured until the end of the career (i.e., including time-varying covariates measured at the end of each semester). Figures 1 and 2 report the results for *Model 4* in terms of estimated fixed-effects coefficients and frailty distribution, respectively.

In terms of students characteristics (Figure 1), results show that students with low income level or with scholarship (LS) have on average a lower risk of dropout with respect to students with a medium income level; commuters have on average a higher risk of dropout with respect to students living in Milan, while offsite students have on average a lower risk of dropout; students coming from technical highschools have on average a lower risk of dropout with respect to students coming from a scientific school; students who enroll at university at an older age have on average a lower risk of dropout with respect to other students; the higher the number of credits and the gpa, the lower the risk of dropout. The C-index results to be 0.9146 for the training set and 0.9133 for the test set.

By looking at Figure 2, we observe that the baseline hazard functions differ across degree programmes with four degree programmes having a statistically significant higher risk of dropout (net to the effect of student-level covariates) with respect to the average and four of them having a statistically significant lower risk of dropout. The estimated variance of the frailty w_j is $\hat{\theta} = 0.02027$ (p-value = 0.014), confirming a significant difference between the baseline hazard functions of different degree courses.

The analysis of the model predictive power reveals that, with respect to the baseline model, as expected, including the career information over time (by including time-varying covariates in the Cox model) increases the predictive accuracy.

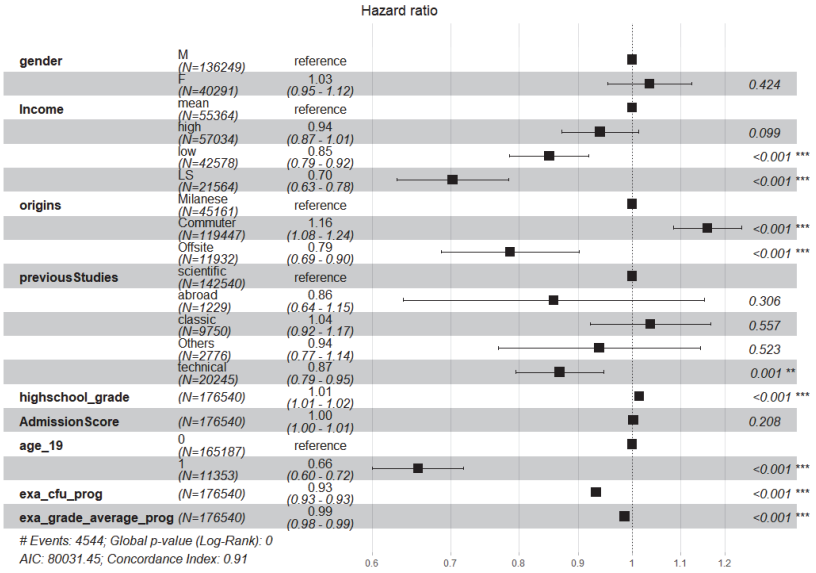


Fig. 1 Frailty cox PH model with time-varying covariates updated each semester until the end of the career (*Model 4*).

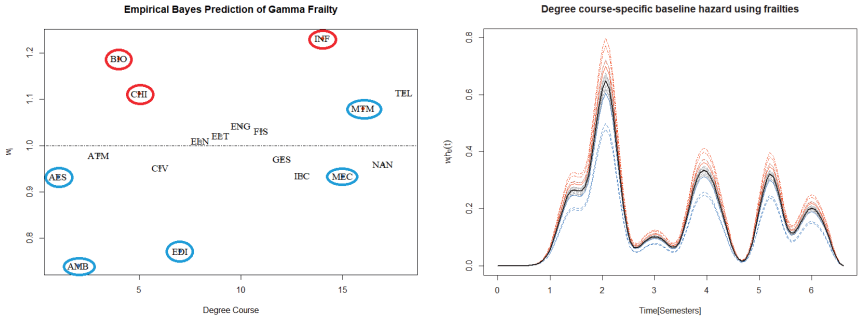


Fig. 2 Estimated frailty distribution in *Model 4*. Left side figure reports the 18 estimated frailties, one for each degree programme. Red dots represent degree programmes with a frailty significantly higher than 1, while blue dots represent degree programmes with a frailty significantly lower than 1. Right side figure reports the baseline hazard for each degree programme.

Nonetheless, while we observe a big gap in the predictive accuracy between *Model 1*, *Model 2* and *Model 3*, the inclusion of the career information regarding later semesters only slightly increases the predictive accuracy, suggesting that first year information is already very powerful in predicting time to dropout.

5 Conclusions

In this work, we apply Frailty Cox PH models with time-varying covariates to model and predict time to dropout of Politecnico di Milano engineering students, within three years after the enrolment. The Frailty Cox PH model with time-varying covariates allows to estimate the association between student characteristics and their risk of dropout over time, to estimate the effect of different degree programmes on the dropout risk of their students and to quantify the advantage, measured in terms of predictive power, of the inclusion of the student career information over time as time-varying covariates of the model.

The survival analysis approach results to be very informative when applied to university dropout data. Future work directions regard the enlargement of the follow-up window and the development of a time-varying frailty COX model that allows the baseline hazard to have different functional forms across degree programmes.

References

1. O. Aljohani. A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher education studies*, 6(2):1–18, 2016.
2. S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 903–912, 2016.
3. ANVUR. *Rapporto biennale sullo stato del sistema universitario e della ricerca*, 2018. https://www.anvur.it/rapporto-biennale/rapporto-biennale-2018.
4. F. Belloc, A. Maruotti, and L. Petrella. How individual characteristics affect university students drop-out: a semiparametric mixed-effects model for an italian case study. *Journal of applied Statistics*, 38(10):2225–2239, 2011.
5. M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni. Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, pages 1–22, 2021. DOI: https://doi.org/10.1080/03075079.2021.2018415.
6. D. G. Kleinbaum, M. Klein, et al. *Survival analysis: a self-learning text*, volume 3. Springer, 2012.
7. T. M. Therneau and P. M. Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.
8. M. C. Voelkle and N. Sander. University dropout: A structural equation approach to discrete-time survival analysis. *Journal of Individual Differences*, 29(3):134–147, 2008.

Family history in survival and disease development

Familiarità nella sopravvivenza e nello sviluppo di malattie

Maria Veronica Vinattieri and Marco Bonetti

Abstract We are interested in a specific aspect related to modeling survival, i.e. the study of family-specific risk. We focus on the genetic "from birth" risk component as opposed to the environmental component. Indeed, the true genetic family risk may be assumed to be latent from birth, and it can be either discrete (with two or more ordered risk levels) or continuous. We explore some latent multiplicative frailty multivariate time-to-event models, clustered at the family level. In particular, we focus on: (i) a binary mixture model with two family frailty levels (high vs. other); (ii) a continuous frailty model. Our contribution is about the prediction of the estimated family-specific risk of a subject from the observed family history (birth and time of event or censoring).

Abstract *Siamo interessati ad un aspetto specifico riguardo ai modelli di sopravvivenza, cioè lo studio del rischio familiare. Ci concentriamo sulla componente genetica "dalla nascita" del rischio, opposta alla componenete ambientale. Infatti, il vero rischio familiare genetico si può assumere come latente dalla nascita, e può essere discreto (con due o più livelli di rischio ordinati) o continuo. Studiamo dei modelli latenti moltiplicativi frailty multivariati per la sopravvivenza, raggruppati per famiglia. In particolare, ci focalizziamo su: (i) un modello mistura binario con due livelli di frailty familiare (alto o altro); (ii) un modello di frailty continuo. Il nostro contributo riguarda la previsione del rischio specifico familiare stimato per un soggetto dalla storia familiare osservata (nascite e tempi all'evento o di censura).*

Key words: survival analysis, censored data, family history, frailty models.

Maria Veronica Vinattieri
Bocconi University, via Roentgen 1, Milano (MI) Italy, e-mail: maria.vinattieri@phd.unibocconi.it

Marco Bonetti
Bocconi University, via Roentgen 1, Milano (MI) Italy, e-mail: marco.bonetti@unibocconi.it

1 Introduction

We are interested in studying the effect of family-specific risk on survival. The definition of family history refers to the collection of the survival experience of other family members. We assume that families can be defined as clusters, within which individuals share the same risk. The interest is in classifying a subject's family to one of a set of risk groups or, more generally, to estimate the family-specific risk from the available data.

First consider the effect of family risk on longevity, that is defined as the difference between the age at death and the expected age of death as based on temporal and environmental factors. For the jth individual, we let t_j be the longevity, and $s_j \in \{\text{male, female}\}$ the sex. Below we use the subscripts m (mother), f (father) and p (generic parent) to identify the family members. Two models explain the heritability of longevity [5]:

$$t_j = \alpha_0 \frac{(t_{m_j} + t_{f_j})}{2} + \beta$$
$$t_j = \beta_0 + \alpha_0 t_{p_j} + \alpha_1 \mathbb{I}(s_j, s_{p_j}) + \alpha_2 (t_{p_j} \times \mathbb{I}(s_j, s_{p_j}))$$

where $\mathbb{I}(s_j, s_{p_j}) = 1 \iff s_j = s_{p_j}$. Estimation of such models can be performed by least square method [5] from data consisting of individual's medical and biological information in population-scale family trees.

We investigate an alternative model that can be used to study the family risk from a fully multivariate perspective. Differently from above, the quantity of interest t_j is the time-to-event from birth to death. We assume that families are split into groups with different hazard and survival functions. We define a mixture of survival models, where the latent family risk is the mixing variable. For simplicity, we let all family members share the same hazard function.

Mixtures of survival functions are indeed studied through frailty models, where the frailty is a random effect that captures the unobserved heterogeneity among groups (families), so that the distribution of the frailty random variable is the mixing distribution. With a multiplicative frailty variable R, we have that the R-specific hazard functions satisfy the proportional hazard (PH) assumption.

As a special case, if families are split into two groups, say a low and a high risk group, the latent family risk takes values R=low/high, then the hazard function in the two groups can be defined as $\lambda_0(t)$ and $\lambda_1(\text{t}) = \beta\lambda_0(\text{t})$ for R=0 and 1, respectively. Under such PH structure, the survival functions in the two groups are thus:

$$S_0(t) = e^{-\int_0^t \lambda_0(u)\mathrm{d}u};\ S_1(t) = e^{-\int_0^t \lambda_1(u)\mathrm{d}u} = e^{-\int_0^t \beta\lambda_0(u)\mathrm{d}u} = [S_0(t)]^\beta .$$

Moving beyond the two-group case (or the similar k-group case), one can define the multiplicative frailty model with hazard function $\lambda(t|R) = R \cdot \lambda_0(t)$, where $\lambda_0(t)$ is the reference hazard function corresponding to $R = 1$, and $f_R(r)$ is the corresponding frailty density function. If all family members have the same frailty, this describes a clustered shared frailty multivariate time-to-event model with families

representing the clusters (see e.g., [8]). Parameter estimation and family-specific prediction for R can then be based on the available (typically, right-censored) survival data of all relatives in all families.

We now describe estimation in such models and classification into a high risk group.

2 Classification from a clustered multiplicative frailty multivariate time-to-event model

As in Section 1, we call R the continuous frailty variable that follows a parametric distribution characterized by θ. Within such framework, we use i to identify the family (out of G) and j to identify its n_i members. Following [4] we develop the complete likelihood $L(\underline{R};\underline{\mathbf{Z}})$ for the problem, where $\underline{\mathbf{Z}} = \{\underline{\mathbf{z}}_i, i = 1,\ldots,n\}$, and $\underline{\mathbf{z}}_i = (z_{ij}, \delta_{ij})^T$, $z_{ij} = \min(t_{ij}, c_{ij})$, $\delta_{ij} = \mathbb{I}(t_{ij} \leq c_{ij})$ following the usual notation, that has t indicate the survival time and c indicate the (independent) censoring time. X_{ij} indicates the baseline covariate vector for subject j in family i. The complete likelihood $L(\underline{R};\underline{\mathbf{Z}})$ can be written in terms of the frailty parameter θ and the survival parameters, i.e. the vector coefficient β for the covariate effects and the baseline hazard function λ_0. So, following the shared frailty hazards structure, we have $\lambda_{ij}(t|x_{ij}, R_i) = R_i \cdot \lambda_{0ij}(t|x_{ij})$, and $\lambda_{0ij}(t|x_{ij}) = \lambda_0(t)\exp(x'_{ij}\beta)$ for family i. The full likelihood is composed by two quantities: $\mathscr{L}(\beta, \lambda_0, \theta) = \mathscr{L}_1(\theta)\mathscr{L}_2(\beta, \lambda_0)$. The estimation procedure follows the approach from [4] with the notation from [9]. The frailty R can be taken to be distributed as a Gamma with shape θ and rate $1/\theta$:

$$\mathscr{L}_1(\theta) = \prod_i \frac{1}{\Gamma(1/\theta)\theta^\theta} R_i^{\theta-1} e^{-R_i/\theta},$$

$$L_1 = \log\mathscr{L}_1(\theta) = \sum_i \left[-\log(\Gamma(\theta)) - \theta\log(\theta) + (\theta-1)\log(R_i) - \frac{R_i}{\theta}\right].$$

We compute also the survival component of the likelihood:

$$\mathscr{L}_2(\beta, \lambda_0) = \prod_{i=1}^{G}\prod_{j=1}^{n_i} \lambda_{ij}(z_{ij})^{\delta_{ij}} S_{ij}(z_{ij}) = \prod_{i=1}^{G}\prod_{j=1}^{n_i} \left(R_i \cdot \lambda_{0ij}(z_{ij}|x_{ij})\right)^{\delta_{ij}} \exp(-R_i \cdot \Lambda_{0ij}(z_{ij}|x_{ij})),$$

$$L_2 = \log\mathscr{L}_2(\beta, \lambda_0) = \sum_{i=1}^{G}\sum_{j=1}^{n_i} \delta_{ij}\log(R_i \cdot \lambda_{0ij}(z_{ij}|x_{ij})) - R_i \cdot \Lambda_{0ij}(z_{ij}|x_{ij}),$$

where $\Lambda(t)$ indicates the cumulative hazard function. The full log-likelihood is then $L(\theta, \beta, \lambda_0) = L_1(\theta) + L_2(\beta, \lambda_0)$.

In order to estimate the model parameters, we may specify the form of the baseline hazard function. Indeed, the baseline hazard can assume a parametric form or it can be left unspecified (this corresponds to the semi-parametric case) [3].

For example, for the parametric specification a common model for the time-to-event variable is the Weibull distribution $T \sim \text{Weibull}(\text{shape}=\gamma, \text{scale}=\mu)$ with the corresponding hazard functions. Given the multiplicative frailty structure, one can re-parametrize the conditional (on R) survival distribution as $T \sim \text{Weibull}(\text{shape} = \delta, \text{scale} = \mu/R^{1/\delta})$. Parameter estimation is then achieved by maximizing the log-likelihood function [7] through the Expectation Maximization (EM) algorithm [3], [2]. In both parametric and semi-parametric cases, all parameters can be estimated and used to perform classification.

We move now to the classification procedure. In the parametric approach, the prediction follows a method based on the expectation step of the EM algorithm [7]. Our contribution refers to the semi-parametric approach instead. We suggest to perform risk prediction by fixing a grid of values $\{u_1, \ldots, u_K\}$ for the frailty quantity and implementing the following steps:

(i) obtain $\widehat{S}_0(z_{ij})$ and $\widehat{\lambda}_0(z_{ij})$ from the Breslow estimator;
(ii) compute $\widehat{f}(\underline{\mathbf{z}}_i|u_k) = \prod_{j=1}^{n_i} \left[(\widehat{\lambda}_0(z_{ij})u_k)^{\delta_{ij}} [\widehat{S}_0(z_{ij})]^{u_k} \right]$;
(iii) compute $\widehat{f}(\underline{\mathbf{z}}_i, u_k; \widehat{\theta}) = \prod_{j=1}^{n_i} \left[(\widehat{\lambda}_0(z_{ij})u_k)^{\delta_{ij}} [\widehat{S}_0(z_{ij})]^{u_k} \right] f(u_k|\widehat{\theta})$, $\forall u_k$ in the grid;
(iv) compute the integral $\widehat{f}(\underline{\mathbf{z}}_i; \widehat{\theta}) = \int_{\mathbb{R}^+} f(\underline{\mathbf{z}}_i|u_i) f(u_i; \widehat{\theta}) \mathrm{d}u_i = \sum_k \Delta(u_k) \widehat{f}(\underline{\mathbf{z}}_i, u_k; \widehat{\theta})$, where $\Delta(u_k) = u_{k+1} - u_k$;
(v) compute $\widehat{f}(u_k|\underline{\mathbf{z}}_i; \widehat{\theta}) = \widehat{f}(\underline{\mathbf{z}}_i, u_k; \widehat{\theta}) / \sum_k \Delta(u_k) \widehat{f}(\underline{\mathbf{z}}_i, u_k; \widehat{\theta})$.

The predicted continuous shared frailty value $\widehat{R}_i$ for each family i is computed with the rule below, i.e. it takes the value corresponding to summing up the estimated (as above) density function on the grid values until a desired threshold quantile $q \in [0,1]$:

$$\widehat{R}_i = u_k : \sum_{j:u_j \le u_k} \widehat{f}(u_j|\underline{\mathbf{z}}_i; \widehat{\theta}) \Delta(u_j) \le q. \tag{1}$$

Indeed we choose the posterior percentile so that it minimizes the misclassification rate.

In order to explore the discrete splitting of families into, say, two risk groups, we may transform the continuous frailty into a binary 2-group variable. We can then carry out the (actionable) classification according to the rule:

$$\widehat{RB}_i = \begin{cases} \text{low} & P(u_i < \widehat{\eta}_u|\underline{\mathbf{z}}_i; \widehat{\theta}) \ge q \\ \text{high} & P(u_i < \widehat{\eta}_u|\underline{\mathbf{z}}_i; \widehat{\theta}) < q \end{cases} \Leftrightarrow \begin{cases} \text{low} & \text{Quant}(u_i|\underline{\mathbf{z}}_i; \widehat{\theta}) \le \widehat{\eta}_u \\ \text{high} & \text{Quant}(u_i|\underline{\mathbf{z}}_i; \widehat{\theta}) > \widehat{\eta}_u \end{cases} \tag{2}$$

where $\widehat{\eta}_u = \text{Quantile}\left(\text{Gamma}(\widehat{\theta}, 1/\widehat{\theta})\right) \in [0,1]$, $\text{Quant}(u_k|\underline{\mathbf{z}}_i; \widehat{\theta}) = u_k : P(u_k|\underline{\mathbf{z}}_i; \widehat{\theta}) \le q$ with $q \in [0,1]$ as above. And, $P(u_k|\underline{\mathbf{z}}_i; \widehat{\theta}) = \sum_{j:u_j \le u_k} \widehat{f}(u_j|\underline{\mathbf{z}}_i; \widehat{\theta}) \Delta(u_j)$.

We now describe a simulation scenario that can be used to implement the classification procedure introduced above.

3 A simulation scenario

One can generate some family structures and survival times, and implement the 2-group classification above. We firstly fix the number of families in our dataset. We count for each woman a mother and a grandmother. Plus, we assume a varying number of sisters and aunts that can be distributed as a Poisson(λ_s) and a Poisson(λ_a) respectively. We fix $\lambda_s = 1,\ \lambda_a = 0.5$, so that the varying family size is

$$\text{family size} = 3 + \text{Poisson}(1) + \text{Poisson}(0.5).$$

The expected value of the sample size is therefore $\mathbb{E}(N) = 3+1+0.5 = 4.5$. We generate the time-to-event T from the frailty Weibull distribution $T \sim \text{Weibull}(\text{shape} = \gamma, \text{scale} = \mu/R^{1/\gamma})$, with $R \sim \text{Gamma}(1,1)$, $\mu = 1,\ \gamma = 5$. The censoring times are generated from a Uniform distribution $C \sim \text{U}(0,12)$, and for each subject we generate $Z = \min(T,C)$ and $\delta = \mathbb{I}(Z = T)$.

One can then implement the parametric and semi-parametric estimation with setting described in Section 2, and apply some diagnostic tools in order to analyse the goodness of the classification, such as the scatter-plot of R versus $\widehat{R}$ (see 1) to obtain a visual analysis of the classification accuracy, and the confusion matrix between the median-based risk group $RB = \mathbb{I}(R \leq \text{Median}(R))$ and the estimated risk group $\widehat{RB}$ obtained as in 2 for $q = 0.5$. We also extend to $q = 0.25$ so that we keep low (and realistic) the posterior high risk families proportion. We can use the agreement index Cohen's kappa [6], in order to have a summary of the binary classification results.

$$\text{Cohen's kappa } k = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP+FP)\cdot(TN+FP) + (TP+FN)\cdot(TN+FN)},$$

with TP, TN, FP, FN indicate the true positive, true negative, false positive and false negative proportions, where positive (negative) stands for $\widehat{RB}$=high risk (low risk). The classical accuracy[1] can be also computed:

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \geq 0.$$

Both indices vary in the range $[0,1]$, where zero means no agreement and one means perfect agreement. Sensitivity and specificity can also be used as classification accuracy measures.

Thus, we sample 3000 families and we explore the classification accuracy in three different scenarios: (i) parametric hazard and binary classification with median as threshold; (ii) semi-parametric case with $q = 0.5$; (iii) semi-parametric case with $q = 0.25$. We carry out this analysis for family size and overall. The results for family size are not reported because irrelevant, while the summary of the overall results is in table [1]. The posterior high risk families proportion in the semi-parametric case, varying q, is: 0.21 ($q = 0.25$); 0.27 ($q = 0.5$).

	Cohen's k	Accuracy	Sensitivity	Specificity
(i)	0.73	0.91	89.89	87.34
(ii)	0.89	0.95	97.97	93.51
(iii)	0.92	0.96	94.65	96.92

Table 1: classification accuracy indices for the three scenarios (see text).

4 Discussion

Preliminary results suggest the absence of important differences in classification accuracy across different family sizes (results not shown). We notice a substantial difference in classification accuracy between the parametric and the semi-parametric setting, up to the use of the median as classification threshold. Moreover, the use of the first quantile seems to be preferable in terms of classification accuracy and posterior proportion of high risk families. This was confirmed also through further checks (results not shown), as the scatter-plots (see text). We plan to extend this study to the case of development of a disease (e.g. breast cancer), where risk classification can help implement tailored, more intensive screening schedules for subjects with higher risk of developing the disease, as estimated from frailty member's history of breast cancer.

References

1. Accuracy and precision. Wikipedia: https://en.wikipedia.org/wiki/Accuracy_and_precision
2. Balan, T.A., Putter, H.: frailtyEM: An R package for estimating semiparametric shared frailty models. Journal of Statistical Software, **90**, 7, 1–29, (2019)
3. Duchateau, L.: The frailty model. Springer Science & Business Media, (2007)
4. Hougaard, P.: Analysis of multivariate survival data. Springer Science & Business Media, (2012)
5. Kaplanis, J., Gordon, A., Shor, T.: Quantitative analysis of population-scale family trees with millions of relatives. Science, **360**, 6385, 171–175, (2018)
6. Cohen's kappa. Wikipedia. https://en.wikipedia.org/wiki/Cohen%27s_kappa
7. Munda, M., Rotolo, F., Legrand, C.: parfm: Parametric frailty models in R. Journal of Statistical Software, **51**, 11, 1–20, (2012)
8. Rodrıguez, G.: Multivariate survival models, (2010)
9. Yu, B.: Estimation of shared gamma frailty models by a modified EM algorithm. Computational statistics & data analysis, **50**, 2, 463–474, (2006)

Text mining

Topics & Metaverse: an explorative analysis

Tematiche e Metaverso: un'analisi esplorativa

Emma Zavarrone, Alessia Forciniti, Emanuele Parisi and Maria Gabriella Grassia[2]

Abstract In October 2021 Zuckerberg drew the attention of media, businesses and common people to a virtual reality platform called Metaverse, so that its search has become a trend. In this paper, we explored the people's opinion in Italy on Twitter by analysing the hashtags. Through the *Latent Dirichlet Allocation*, we detected latent topics of hashtags and projected them in an inter-topic distances map which clustered similar topics in two-dimensional space. To study the inter-topic distances, we used the measures of saliency and distinctiveness. Despites the saliency presents similarities between the groups, the distinctiveness denotes that one topic is generalist, based on platforms interconnection, and the other one is specialistic and focused on virtual technologies and virtual economies.

Abstract *Nell'ottobre 2021 Zuckerberg ha attirato l'attenzione di media, imprese e persone su una piattaforma di realtà virtuale chiamata Metaverso, così che la sua ricerca è diventata un trend. In questo paper, abbiamo esplorato quello che dicono le persone in Italia su Twitter analizzando gli hashtag. Con la Latent Dirichlet Allocation, abbiamo rilevato gruppi latenti di hashtag e li abbiamo proiettati in una mappa di distanze inter-topic che raggruppa i topic simili in uno spazio bidimensionale. Per studiare le distanze inter-topic, abbiamo usato misure di salienza e distintività. Nonostante la salienza presenti similarità tra i gruppi, la distintività denota un topic più generalista, basato sull'interconnessione tra le piattaforme, l'altro più specialistico focalizzato sulle tecnologie e economie virtuali.*

Key words: metaverse, virtual world, LDAvis, intertopic-distance, hashtag

[1] Emma Zavarrone, Università IULM; emma.zavarrone@iulm.it
Alessia Forciniti, Università IULM; alessia.forciniti@iulm.it
Emanuele Parisi; Università IULM; emanuele.parisi@iulm.it
[2] Maria Gabriella Grassia, Università di Napoli, "Federico II"; mg.grassia@unina.it

1 Introduction

In October 2021 Mark Zuckerberg announced Facebook's name changed to Meta shifting everyone's attention towards a new virtual reality called *Metaverse*. It is a new platform of encountering between the physical and digital world (Nesbo, 2021) which through a deep user-experience has the potential to contribute to the access of new creative, social and economic opportunities for consumers and enterprises. The platform represents a 3D virtual world where the activities are carried out by means of augmented and virtual reality equipment, allowing to collaborate, meet, buy, sell, play games, and socialize in 3D spaces. The economic aspect is based on cryptocurrencies as medium exchange and on Non-Fungible Tokens (NFTs), digital items bought and sold using the blockchain technology. These systems - focused on virtual social connection and virtual economy - became popular during the pandemic of Covid-19 for remote working and online activities of leisure or shopping.

The metaverse is not a new concept since the term appeared for the first time during 1992 in a science fiction novel by Neal Stephenson called Snow Crash. In the novel, the metaverse was presented as the integration between the virtual and physical spaces. In fact, the concept has never been more popular as today to indicate the combination between virtual reality, augmented reality, and virtual economies (Newton, 2021). The concept has turned into a trend topic for journalists, researchers, investors and common people. For some, the aim is to emphasize development progress of technologies (Neiger, 2021), for others, to discuss criticism and concerns connected to feasibility, information privacy (Lee *et al*., 2021), user safety (Chalmers, 2022), social issues (Huddleston, 2022).

Since Zuckerberg presented this platform and revealed to the media that he had made significant investments in it, the frequency of the query 'metaverse' on the search engines around the world was 1,339 from 10th October 2021 to 25th February 2022, and in Italy equal to 1,147. At the same time, many companies such as Microsoft, Samsung, Nike and others are engaging to take their place in this market.

The need to search what metaverse is and to take part in it has become a popular trend, but few or none know exactly what metaverse is.

In this paper, we explored what people say in Italy about the metaverse on social media and on Twitter recognized as 'collective brain' (Ceron *et al*., 2014) of the society. The focus of analysis was on hashtags as distinctive elements to refer to themes, contents, and events. By means of techniques of textual analytics and more properly through the topic modelling of Latent Dirichlet Allocation, we detected the main topics as distribution of hashtags to find meaningful latent topics and the relationships between topics and hashtags. In this way, we can better understand the people's opinion on metaverse. We used a web-based interactive visualization which allowed to estimate the hashtag relevance as topic-specific frequency of each hashtag and to project an inter-topic distances map which via multidimensional scaling has faithfully clustered similar topics in two-dimensional space.

To follow, section 2 presents data; section 3 describes the methodology; section 4 shows the results and discussion.

2 Data

To investigate topics of hashtags, we extracted the tweets posted by users from 2nd to 13rd February 2022. Data extraction was performed by means of the *application programming interface* (API) of Twitter and the library *rtweet* (Kearney, 2019) by using the keywords such as 'metaverso, nft, blockchain'. With respect to monitoring period, we extracted 36,449 tweets of which 7,731 in Italian language. Before proceeding to the pre-processing and analysis of the corpus steps, we improved extraction by filtering the tweets according to user's account. Precisely, we used machine learning methods proposed by Kearney in 2018 to classify Twitter accounts as bots or non-bots. The model based on a gradient boosted system uses both users-level (as bio, location, number of followers and friends, *etc*) and tweets-level (as number of hashtags, mentions, etc, in a user's most recent 100 tweets) data to estimate the probability of users being bots. However, Twitter's REST API rate limits the filtering to only 180 estimates per every 15 minutes, even if by default the model shows accuracy equal to 93.53% to classify bots and of 95.32% to classify non-bots. The model allowed to estimate 179 bot accounts, whose tweets have been selected and deleted from the dataset. By cleaning the corpus from bots, we obtained 2,359 non-bot tweets of which only 1,396 included the word 'metaverse'.

3 Methodology

The methodological approach was based on *Natural language Processing* (*NLP*) and specifically textual analysis techniques. At first, data pre-processing was performed on 1,396 texts parsed and tokenized by means of the canonical steps of normalization, lemmatization, removal of punctuation, numbers, stop words, usernames, retweets and elements of noise like URLs, link, emoji, special characters. We obtained a vocabulary composed of 740 types by filtering nouns, adjectives and verbs and 186 hashtags. We focused our attention on topic extraction of the hashtags to obtain a thematic knowledge of the phenomenon. In this case, our aim was not based on bag-of-word coding but it was oriented to study the context in which the hashtags were inserted in the discourse. To reach this goal, we adopted the unsupervised topic modelling method known as *Latent Dirichlet Allocation* (*LDA*) (Blei *et al.,* 2003). This approach represents a fully generative model for identifying sets of latent arguments (Ostrowski, 2015) for each of which it associates a sequence of words in terms of multinomial distribution. We used this approach to work on

short text (e.g.; Chen *et al*., 2016) and detect topics of hashtags. To detect the meaning of each topic, the prevalence of topics and their relationships, we based our work on a web-based interactive visualization called *LDAvis* (Sievert & Shirley, 2014) useful to make the results easily interpretable. It allows to examine the topic-hashtag relationships, the corpus-wide frequency of a given hashtag as well as the topic-specific frequency of the hashtag as proposed by Chuang *et al*., 2012 (in Sievert & Shirley, 2014). The texts visualization shows the inter-topic differences by means of the topic representation in a two-dimensional space given by multidimensional scaling. The inter-topic distances are computed using Jensen-Shannon divergence also known as information radius or total divergence to the average, which represents a method to measure the similarity between two probability distributions.

In addition, to better understand the topic model, the authors have reproduced the Chuang *et al.*'s tool (2012) based on two main measures: distinctiveness and saliency. By computing the Kullback-Liebler divergence, we can obtain the distinctiveness between the distribution of topics given the hashtag and the marginal distribution of topics, and the saliency as hashtag's overall frequency. A tuning parameter λ in the computation of relevance was instead developed by Sievert & Shirley (2014) where $0 \leq \lambda \leq 1$: $\lambda=1$, the parameter ranks hashtags in decreasing order of their topic-specific probability and $\lambda=0$ the parameter ranks hashtags based on the ratio between the probability of generic hashtag, belonged to the topic k and marginal probability distribution of the same hashtag in the corpus.

4 Results and discussion

LDAvis allows a global view of the topics, both their prevalence and similarity to each other in a compact space split in two pieces.
The left panel shows the identified hashtag groups in a two-dimensional space, using an inter-topic distance map via multidimensional scaling. The topics are visualized by the principal components (PC) axis distribution. Each topic is represented by a circle and its overall prevalence is encoded using the areas of the circles, which are proportional to the relative prevalence of the topics in the corpus. As shown in the *Figure 1* the first three topics have a greater marginal topic distribution than others. The areas of first and second circle describe a marginal topic distribution of 10% of the corpus; while the third one describes it at 5%. In fact, in the 15-topic model presented, the first three topics comprise respectively 23%, 19.3% and 7.2% of the corpus. The circle's centers are defined by the computed distance between topics.
The inter-topic distance on the map of the PC shows groups of hashtags. Some of these groups are overlapped, denoting a strong positive correlation. We can observe how the map splits in the positive correlated thematic areas. Specifically, there is a positive correlation between the groups 1 and 6; the groups 8, 15, 12, 13; the groups

2, 9,14,10,11; the groups 3, 4 and 5. Between the first two groups which present the greater relative prevalence of the topics in the corpus, we note that the thematic group 2 is more away from the origin and then better represented on the factorial map. The right panel consists of a horizontal bar chart whose bars represent the individual hashtag. This section is effective to interpret the topic groups, depending on the selected topic. The hashtags are ranked according to the relevance in decreasing order. The relevance is important to understand how much the hashtags are discriminant for each specific topic and is measured by the parameter of topic-specific probability (λ), whose "optimal" value is 0.6 (Sievert & Shirley, 2014). In the *Figure 1* on the right, we see the most useful hashtags for interpreting the topic 1.

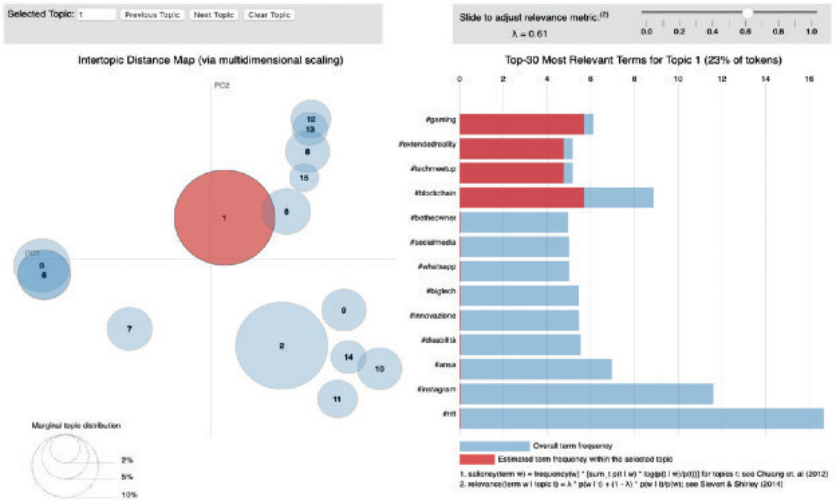


Figure 1*: Inter-topic distance map for all topics and relevance of hashtags for the Topic 1*

According to the abovementioned literature in the methodological paragraph, we can describe the results of the bar chart according to two measures: saliency and distinctiveness at relevance parameter λ = 0.6. The saliency - as hashtag's overall frequency - indicates most frequent #nft, #instagram, #blockchain, #ansa, but the distinctiveness - as estimated hashtag frequency within the topic 1 - brings out the importance of the emphasizing of the user-experience. The characteristic element is the extended reality as umbrella term that covers technologies which allow the human-to-PC screen interface in an immersing virtual environment, including the playful aspect of the virtual through gaming or communities, and that of buying and selling through blockchain technology. The *Figure 2* presents instead the most important hashtags to interpret the topic 2. The bar charts of the *Figure 1* and *2* allow to observe that they contain common hashtags. The right panel in Figure 2 indicates such as very salient #nft, #blockchain, #instagram and so on; but, the metric of distinctiveness suggests as discriminants #instagram, #socialmedia, #whatsapp. Thus, despite the saliency between the topic 1 and 2 presents similarities, the distinctiveness denotes that the topic 2 is generalist and based on

platforms interconnection, while the topic 1 is focused on the specificity of technology, virtual reality, virtual interactions, and virtual economies. This explains the reason why on the inter-topic distances map the two topics are not close. *LDAvis* showed a deep inspection of topic-hashtag relationships and of topic-topic similarity by LDA model. The graphic visualization and the metrics allowed to explore that, in some cases, people describe a phenomenon by the same word or same hashtag but it is the topic grouping to bring out a different semantic knowledge. Specifically, the weight of each hashtag in each topic makes two levels of knowledge on metaverse emerge: one approximative and the other aware of peculiarities, by leading to observe the distance between the topics despite using the same hashtag.

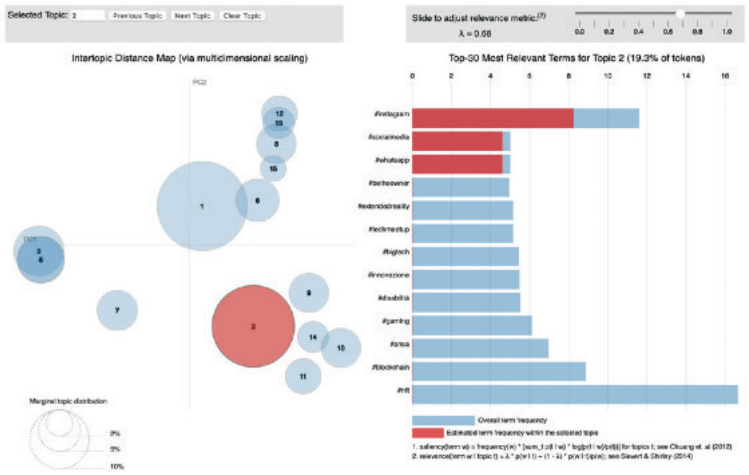


Figure 2*: Inter-topic distance map for all topics and relevance of hashtags for the Topic 2*

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (4-5), 993--1022 (2003).
2. Ceron A., Curini L., Iacus S.M.: Social media e Sentiment analysis. L'evoluzione dei fenomeni sociali attraverso la Rete. Springer per l'Innovazione (2014).
3. Chalmers, D.: What Should Be Considered a Crime in the Metaverse?. Wired. ISSN 1059-1028 (2021).
4. Chen, Q., Yao, L., Yang, L.: Short text classification based on LDA topic model. 2016 International Conference on Audio, Language and Image Processing (ICALIP),749--753 (2016).
5. Huddleston, J.T.:This is creating more loneliness': The metaverse could be a serious problem for kids, experts say. CNBC. (2021).
6. Kearney, M.: tweetbotornot (2018). Available via: https://github.com/mkearney/tweetbotornot . Last access: February 2022.
7. Lee, L.H., Braud, T., Zhou, P. *et al*.: All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda. https://arxiv.org/pdf/2110.05352.pdf (2021).
8. Neiger, C.: Virtual reality is too expensive for most people - but that's about to change. Business Insider (2021).

9. Newton, C.: Mark Zuckerberg is betting Facebook's future on the metaverse. The Verge. Archived from the original (2021). Available via: https://web.archive.org/web/20211025042942/https://www.theverge.com/22588022/mark-zuckerberg-facebook-ceo-metaverse-interview
10. Ostrowski, D.A.: Using latent dirichlet allocation for topic modelling in twitter. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), 493--497 (2015).
11. Sievert; C. & Shirley, K.E.: LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63--70. Association for Computational Linguistics (2014).

Applying Topic Models to bibliographic search: some results in basketball domain.

Applicazione dei Topic Models alla ricerca bibliografica: alcuni risultati nel dominio del basketball

Manlio Migliorati, Eugenio Brentari

Abstract Bibliographic search finalized to identify relevant material about the specific problem to be faced is normally the first phase to be accomplished when starting a new research activity. Such literature analysis can be heavily time consuming, and often only a little part of the analysed material is actually useful. In this short paper, results of using Topic Models approach as support tool for bibliographic search are reported and compared to results obtained analysing via the classic manual approach on a corpus of 495 papers about basketball domain.

Abstract *La ricerca bibliografica finalizzata a identificare il materiale relativo allo specifico problema oggetto di studio è, di norma, la prima attività da svolgersi quando si inizia un progetto di ricerca. Questa attività può risultare onerosa in termini di tempo, e spesso solo una piccola parte del materiale consultato si rivela effettivamente utile. In questo lavoro si riportano i risultati derivanti dall'utilizzo dei Topic Models come strumento di supporto per l'analisi bibliografica, e tali risultati sono confrontati con quelli ottenuti analizzando con la classica ricerca manuale un corpus di 495 documenti relativi al basketball*

Key words: Bibliographic search, topic models, text mining

1 Introduction

Bibliographic search is the first step to be accomplished in facing a new research activity, to identify the relevant material for the domain under analysis. To this purpose, in the last decades several web based services (for instance Web of Science (WoS) from Clarivate Analytics or Scopus from Elsevier) are born, grouping an impressive quantity of scientific material and offering several tools to select and investigate it. Nevertheless, the huge number of scientific sources to be considered,

University of Brescia, Department of Economics and Management, Contrada S. Chiara 50
e-mail: manlio.migliorati@unibs.it

constantly increasing, constitutes an actual challenge for researchers, particularly in approaching a new research field.
The typical approach to identify relevant material is based on querying the on-line repositories to make a preliminary selection, then filtering the results using the provided ad hoc tools and at last applying bibliometric tools as Bibliometrix [1] or manually analysing documents on the base of abstracts, keywords and other classification criteria. This process can actually be time consuming, and often only a little part of the analysed material is actually useful.
In this manuscript, results produced using the Topic Models approach as a support tool for bibliographic search are compared to the classical manual bibliographic analysis output, taking into account a dataset of 495 documents, related to data science applied to basketball domain, produced by on-line repositories querying tool. All implementations of this work have been done in R.
The short paper is structured as follows: section 2 describes the dataset of documents used in this work, section 3 contains a summary of the manual analysis results, section 4 informally summarizes the Topic Models approach, section 5 reports the output we obtained and section 6 presents some conclusions and directions for future steps.

2 The dataset

The specific bibliographic domain to be investigated for our research is related to the usage of data science in general, and machine learning techniques in particular, to the field of basketball, trying to address factors leading to team's victory. The dataset has been produced by a simple (trying to avoid to cut off elements) bibliographic search executed in WoS Core Collection using the query:
("data science" OR "machine learning" OR "business intelligence" OR "data mining" OR analytics OR statistics) AND basketball.
That query, designed to take into account publications years from 1999 to 2019 and to search in title, abstract, author keywords, and Keywords Plus (a set of keywords defined by WoS itself), has been executed on 10th of June 2019, producing a dataset containing 495 documents.

3 Manual analysis of the documents dataset

The original dataset has been manually filtered addressing our specific goal: in particular we are interested in application of statistics about factors addressing team victory. Application of tools made available by WoS, in terms of title analysis, categories and sources, reduced the dataset to 175 elements. Then, this intermediate dataset has been further filtered on the basis of the manual reading of abstracts (and,

in several cases, of the complete papers), to arrive to the final dataset composed by 117 elements, identified as relevant to our research purposes.

4 Using Topic Models

The first dataset produced by WoS query, counting 495 documents, has been classified via Topic Models [2] approach, a statistical modeling widely used in Natural Language Processing (NLP), to clusterize in an unsupervised way a corpus of documents on the basis of latent topics they refer to. In particular, in this work Latent Dirichlet Allocation (LDA) [3] has been used.
In this approach topics are seen as a mixture of words, and the β probability that a term being generated from a topic is calculated. Moreover, LDA models each document as a mixture of topics, and the γ probability that a document belongs to a topic is estimated. Two remarks about Topic Models approach:

1. documents are processed using NLP tools (see for instance [4]):
   - each document in the dataset is tokenized (i.e. words are identified)
   - the set of tokens is cleaned with respect to stop words (i.e. not relevant terms are erased: articles, prepositions, etc)
   - each remaining token must be rooted (i.e. normalized to a base form); in this work we describe the results of stemming, i. e. root form of a word obtained removing the last few characters, using the `SnowBallC` R package [5]).
   - finally, a terms-documents matrix is calculated, constituting the input of the LDA algorithm.

   In our analysis on the original dataset containing 495 documents, we found 55849 tokens, reduced to 4935 stems via the process described above.
2. the number of clusters to be used in classifying the documents of a corpus must be defined in advance. This is an important point, because a low number of clusters does not enable to identify the most interesting subsets, and viceversa a high number of clusters makes it hard to identify differences among them.

In this work we used the `topicmodels` R Package [6] for fitting Topic Models, and the `ldatuning` R Package [7] as a support tool to define the number of clusters. Application of `ldatuning` R Package to our data is represented in Figure 1, where a value (scaled between 0 and 1) is produced for different number of clusters; such a value must be minimed for metrics named *CaoJuan2009* [8] and *Arun2010* [10], and maximized for the metric named *Deveaud2014* [9].
We decided to consider 5 clusters, a number small enough to let us understand the differences among clusters, and corresponding to a local minimum for the metric *CaoJuan2009* and to a local maximum for the metric named *Deveaud2014* (in this case no useful information is provided by the metric *Arun2010*, where the value is monotonically decreasing with the increasing of the number of clusters).

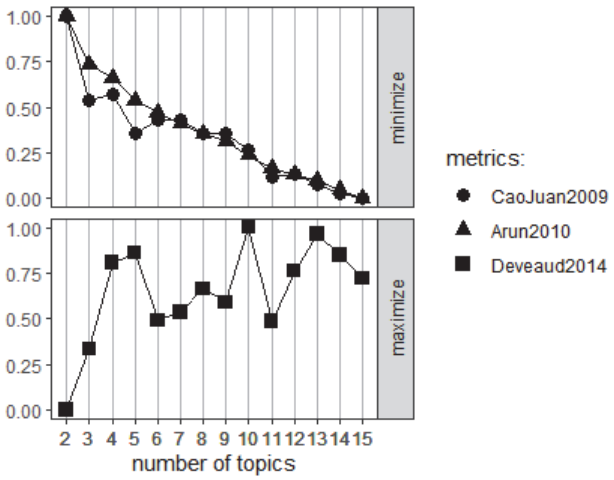


Fig. 1 Trends of some normalized metrics as function of the number of topics

5 First results of the analysis

Figure 2 shows the 10 most representative stems (highest β values) for each cluster:

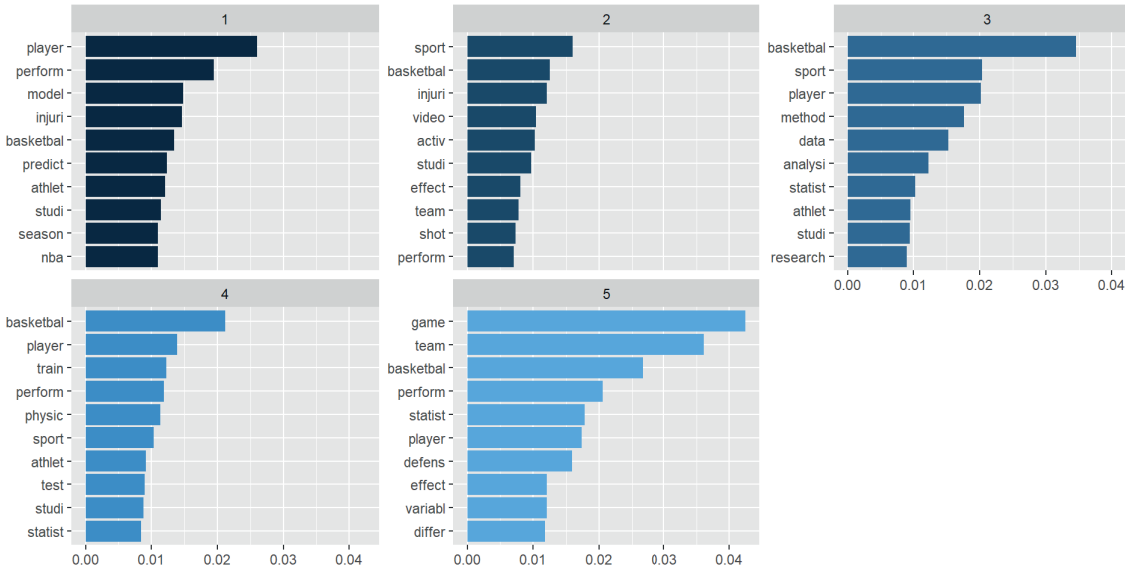


Fig. 2 Top β words for each of the 5 clusters.

Looking to the top beta words of Figure 2, we start to identify the most interesting clusters: cluster 1 seems more focused on single players (*player*, *injuri*, *athlet*); cluster 2 does not seem to address victory drivers, cluster 3 maybe could be interesting for our purposes (*data*, *analysi*, *statist*), cluster 4 seems focused on training and at last cluster 5 seems the most interesting one (*game*, *team*, *perform*, *statist*).
Moreover, for each document the γ probability of being classified in one of the 5 clusters is calculated, and a criterion can be defined to attribute a document to a cluster on the base of γ value. In this work a document is attributed to the cluster for which it has the highestγ, some examples are reported in Table 1.

Table 1 Examples of γ analysis: documents are attributed to a cluster considering the highest γ).

Doc ID	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Highest
1	0.00026	0.00026	0.91826	0.00026	0.08095	3
100	0.00061	0.00061	0.00061	0.70853	0.28963	4
216	0.00036	0.13561	0.34914	0.16657	0.34832	3

For each cluster, the 3 papers with the highest γ are then summarized (in terms of title, source, abstract, keywords), to help in confirming the actual interest for the cluster. For instance, Table 2 reports information[1] about the paper 459 (having the highest gamma among documents attributed to cluster 5), and on the basis of these information we can evaluate the relevance of the cluster.

Table 2 Summary of information about the top γ paper for cluster 5, used to evaluate if the relevance of the cluster.

paper	459
title	discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues
source	European Journal of Sport Science
author keywords	basketball; discriminant analysis; player position;professional;league;game statistics

Quality of clustering has been evaluated on the basis of the distribution among the 5 clusters of the 117 manually adopting a specific criterion of attribution[2], in this work attributing the cluster with the highest γ, and results are reported in Table 3 (where the dataset is composed by 493 documents because of 2 documents with an empty abstract that have not been considered). From data in Table 3 it is possible to verify how about 48% of the 140 papers attribute to cluster 5, seeming to be the most interesting on the basis of β analysis, has been manually selected, and how it contains 58% of manually selected papers.
Considering together clusters 3 and 5, we have a set of 265 documents, containing more than 63% of manually selected papers, and including about 74% of all papers

[1] abstract is not reported to save space

[2] classification can be different adopting other criteria, e.g. attributing a document to a cluster only if γ probability is higher than a fixed threshold greater than 0.5

Table 3 Some Statistics about Topic Models application; 2 documents have not been considered, and all the 117 manually selected papers have been classified

Statistics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
num docs per cluster	96	64	125	68	140	493
num manually selected docs per cluster	22	7	19	1	68	117
% manually selected docs VS docs in cluster	22.92	10.94	15.20	1.47	48.57	
% manually selected docs	18.80	5.98	16.24	0.85	58.12	

manually classified as relevant. This is an interesting result for an unsupervised classification mechanism as Topic Models is.

6 Conclusions and futures directions

In this paper we described the application of Topic Models approach to bibliographic search, and compared its results with the output of the classical manual selection process. First results seem to be interesting: in a 5 clusters model, the more promising cluster (number 5) includes 140 papers and about 48% of them is relevant, covering about the 58% of all manually selected papers. The union of clusters 3 and 5 includes 265 elements: more than 63% of them are manually selected papers, and that set includes about 74% of all paper manually classified as relevant. Future evolutions of this work will address the refinement of Topic Models application and the usage of different criteria in attributing documents to clusters.

References

1. Aria, M., Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics **11(4)**, 959-975, (2017)
2. Blei, D.: Probabilistic Topic Models. Communications of the ACM **55(4)**, 77–84,(2012)
3. Blei, D., Andrew Y., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 993–1022, (2003)
4. Silge, J., Robinson, D.: Text Mining with R: A Tidy Approach. O'Reilly Media, Inc., (2017)
5. Bouchet-Valat, M.: SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. R package version 0.7.0. https://CRAN.R-project.org/package=SnowballC, (2020)
6. Grun, B., Hornik, K.: topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software **40(13)**, 1–30, (2011)
7. Murzintcev, N.: ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. R package version 1.0.2. https://CRAN.R-project.org/package=ldatuning, (2020)
8. Cao J., Xia T., Li J., Zhang Y., Tang S.: A density-based method for adaptive lda model selection. Neurocomputing, **72(7)**, 1775–1781, (2009)
9. Deveaud, R., SanJuan, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, Lavoisier, 61–84, (2014)
10. Arun R., Suresh V., Veni Madhavan C. E., Narasimha Murthy M. N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In Advances in knowledge discovery and data mining, Springer Berlin Heidelberg, 391–402, (2010)

Exploiting Text Mining and Network Analysis for future scenarios development: an application on remote working

Text Mining e Network Analysis per lo sviluppo di scenari futuri: una applicazione sul telelavoro

Yuri Calleo, Simone Di Zio, Vanessa Russo

Abstract In Future Studies, the scenario planning process represents a valuable approach to build images of plausible futures, to set better policy actions in the present. However, the whole process requires great efforts due to the time consuming in selecting the key drivers and issues on the selection of a panel of experts. With the proliferation of web-based quantitative and qualitative data, more information can be obtained, starting from bibliometric data. We propose a methodology to speed up the Scanning phase, by adopting a combination of text mining and network analysis in a chain of steps in which the two approaches integrate to each other. We applied the approach to create useful materials for building future scenarios for remote working.

Abstract *Nei Futures Studies, il processo di pianificazione di scenari rappresenta un approccio utile a costruire immagini di futuri plausibili, per definire migliori azioni politiche nel presente. Tuttavia, l'intero processo richiede sforzi specifici a causa del dispendio di tempo nell'identificazione dei fattori chiave e nella costruzione del panel di esperti. Con il proliferare di dati quantitativi e qualitativi su internet, si possono ottenere maggiori informazioni a partire dai dati bibliometrici. Proponiamo qui una metodologia per accelerare la fase di Scanning, combinando text mining e analisi delle reti in una catena di passaggi in cui i due approcci si integrano tra loro. Abbiamo applicato l'approccio per generare materiale utile alla costruzione di scenari futuri nel contesto del lavoro a distanza.*

Key words: future scenarios, topic analysis, network analysis, remote working

[1] Yuri Calleo, University College Dublin, Ireland; email: yuri.calleo@ucdconnect.ie
Simone Di Zio, University "G. d'Annunzio", Chieti-Pescara; email: simone.dizio@unich.it
Vanessa Russo, University "G. d'Annunzio", Chieti-Pescara; email: v.russo@unich.it

1 Introduction and theoretical framework

Remote working or distance working is a flexible work arrangement that allows an employee to work from a remote location outside the usual offices of the company. This mode of working has existed for many years but with the COVID-19 pandemic we have seen an explosion that has brought out and accentuated both its positive and negative aspects [1]. The future of remote working will be faced with many opportunities as well as challenges, and it is urgent to regulate the sector for the years to come, to safeguard both the well-being and rights of the worker and the needs of the employer. In a highly transitory time, it is practically impossible to predict what will happen, even in just 5 years, therefore we propose a foresight approach. In this paper we try to lay the foundations for the construction of future plausible scenarios on remote working by proposing a new methodological approach. In Futures Studies (FS) the idea of predicting the future has been abandoned decades ago and the research moved on to the issues of mapping alternative futures for better decision making [2]. Among the many methods of the Futures Research, the Delphi and the scenario method are the most widespread. The Delphi is a widely used method for gathering data from a panel of experts, [3] and future scenarios can be defined as descriptions of possible future situations together with the paths of development leading to those futures [4]. Within the FS the Delphi is often used inside the scenario development process and the most popular approach is known as Delphi-based scenarios (DBS) [5]. There are various approaches for the construction of a DBS and here we follow one of the more accredited, based on six steps: Framing, Scanning, Forecasting, Visioning, Planning, Acting [6]. Apart from the initial step of Framing, common in all approaches, Scanning is a crucial step. During this phase, all relevant information on the system under study (trends and key drivers) must be collected as comprehensively as possible. However, the Scanning phase remains a challenging task since the development of the Delphi's projections and the construction of the panel are complex and time-consuming. Commonly, a mix of desk research (review of the relevant literature and specialised reports) and experts' consultation (workshops, interviews, focus groups or Delphi) are performed, but the risk of ignoring important drivers remains. Humans are always subject to cognitive biases and personal views, which can limit and/or distort the search for relevant projections as well as the involvement of notable experts.

In this paper, we propose a new methodology for the Scanning phase through the automatic scanning of the relevant literature (by text mining) and the analysis of networks of experts (by network analysis) which has the following advantages: a) speed up and automatise the formulation of a draft list of Delphi's projections; b) extension of the spectrum of information taken into consideration; c) work around the cognitive biases; d) optimise and objectify the search and recruitment of the experts through the scanning of large networks of scholars and random sampling. Although the final list of projections must be drawn up manually, and the panel of experts must be refined by the research team, the outputs of the automatic analysis here proposed are valuable supporting materials for the Scanning phase of any DBS development.

2 Materials and methods

As known, in social sciences the mixed methods are important but in the Future Research the integration qualitative-quantitative is considered even indispensable [7]. In line with this, the approach proposed here takes advantage of an appropriate combination of qualitative and quantitative methods, which both speeds up the construction of the DBS and makes the development of the scenarios more objective and reproducible. The approach consists of the following tasks: 1) Text Mining for the extraction of the relevant topics on the matter of study. The topic modelling (unsupervised) is performed on bibliometric data (article's abstract) with semantic data from keyword's analysis. We propose a Latent Dirichlet Allocation [8] and a topic coherence metric (Normalised Pointwise Mutual Information) for setting the number of topics to extract. 2) Bibliometric Network analysis, exploiting the collaborations among authors, to create a network of potential experts on the research topic. This produces an undirected graph $G = (V, E)$ where V is the set of scholars (potential experts), E is the set of edges based on collaboration and $N = |V|$ is the number of scholars. Network analysis helps in identifying the most influential people as well as groups of scholars (subgraphs) sharing similar research lines. 3) The graph resulting from step 2) can be complete or disconnected. Let denote with c the number of cliques (i.e., induced subgraphs of G that are complete). Therefore, every clique $g_l = (v_l, e_l)$ - $v_l \subseteq V$; $e_l \subseteq E$; $l = 1, \dots, c$ - contains a group of scholars who collaborate with each other. The c subgraphs will be used both for sampling the expert and for the extraction of subgroups of articles on different research topics. If c were too large, a data driven choice will be made to select a congruous and manageable number of subgraphs (e.g., choosing the larger cliques) and, consequently, N will be restricted to the selected subgraphs. 4) Ranking of the papers within each of the c subgraphs according to the number of citations and subsequent extraction of an appropriate number (say k_l) from each subgraph (k_l can be constant or variable). Then, from each paper we extract authors, title, abstract and keywords, so to create K "cards" ($K = \sum_{l=1}^{c} k_l$). 5) Assigning to each of the N experts a measure of expertise. The problem of measuring the degree of expertise is huge and complex [9], however, here we propose an approach based on a weighted sum of three indicators: a) number of published papers on the topic (P_i); b) number of years referred to the first publication on the topic (Y_i); c) Scopus H-index (H_i). The measure assigned to each expert is $m_i = w_1 P_i + w_2 Y_i + w_3 H_i$ ($i = 1, \dots, N$; $\sum_{j=1}^{3} w_j = 1$) where w_j are weights to modulate the importance of the indicators. 6) Since N could be very large, to build a Delphi panel with an adequate number of experts (say n), it is important to select among the N potential experts according to an objective criterion. We propose a stratified random sampling exploiting the c subgraphs, based on variable probabilities π_i. According to a vertex sampling logic, from each stratum/subgraph we extract a subset of n_i vertices/scholars ($i = 1, \dots, c$; $\sum_{i=1}^{c} n_i = n$) where n_i can be constant or variable and $\pi_i \propto m_i$. This stratified sampling strategy is useful because it ensures the variability of the competences in the panel of experts and the variable probabilities allow scholars with greater expertise to be more likely to enter the panel. 7) Construction of a document-term matrix (dtm) consisting of titles and keywords to

subsequently create a co-occurrence matrix. A semantic co-occurrence network analysis will be developed to highlight specific semantic areas (in number S) using the Modularity algorithm [10]. This widens and strengthens the outputs of step 1).

The topics, the terms with their probabilities, the coherence of terms, the c subgraphs, the c lists of papers, the K cards and the S semantic areas are precious materials for the construction of the Delphi's projections. Even if the final list of projections cannot be automated, the previous results provide, in a very short time and objectively, valid and reliable qualitative/quantitative materials. Furthermore, sampling the scholars provides a fast and objective prototype panel for the Delphi.

3 Application on remote working: results and discussion

We extracted bibliometric data from Scopus with the R Bibliometrix package [11] by considering the following keywords: "smart working", "home working", "distance work", "remote working", "long-distance worker", "teleworking", "telework". Accordingly, we obtained 1031 contributions in a time span of 1 year (from 2020 to 2021). To determine the optimal number of topics we used Python applying coherence metrics; we referred to a coherence score with two validation sets, where $\alpha = 0.1$ and $\beta = 0.1$. We considered all the topics previous to a significant drop, choosing the topics from 6 to 10.

Table 1: *top ten terms of 5 topics sorted by the relevance scores*

Topic	***Keywords (score)***
Work	student (0.022), remote (0.020), work (0.016), employee (0.015), laboratory (0.015), teach (0.015), surgical (0.013), instructor (0.013), method (0.011), result (0.009)
Pandemic	nurse (0.038), bedside (0.019), review (0.015), retirement (0.015), covid (0.012), patient (0.012), clinical (0.012), telecommute (0.006), nursing (0.005), work (0.004)
Remote work and family	work (0.029), system (0.014), model (0.013), smart (0.013), home (0.012), family (0.010), learning (0.010), remote (0.008), individual (0.006), couple (0.005)
Environment	tax (0.030), city (0.025), air (0.015), climate (0.015), transport (0.013), spaces (0.010), energy (0.009), business (0.008), travel (0.007), emissions (0.006)
Anti-contagion measures	contact (0.038), public (0.029), virus (0.019), safe (0.010), masks (0.008), spread (0.006), guidance (0.006), mask (0.005), person (0.005)

We identified the topics semantically coherent with the purpose of our study, highlighting 5 topics with 10 relevant keywords (Tab. 1). Overall, the topics focused on three principal themes: a) the workplace environment and the possible complications due to the remote working (topic 1: "surgical", "teach", and

"instructor"), and the problems of remote working in the family (topic 3: "family", "learning", and "couple"); b) the environment and climate threats (topic 4: "climate", "emissions", "transport"; c) the health theme (topic 2: "nurse", "retirement", "patient") and the anti-contagion measures (topic 5: "contact", "masks", "measures" and "spread").

The academic collaboration matrix identifies an undirected graph of $|V| = 4008$ nodes (potential experts) and $|E| = 41226$ links. The graph is composed of $c = 899$ subgraphs, which is an unmanageable number for our purposes. Since the size of the cliques varies from $|v_l| = 2$ to $|v_l| = 104$, following a qualitative analysis of the structure of the graph and the distribution of $|v_l|$, the subgraphs with $|v_l| \geq 25$ were selected, leading to a reduced number of $c = 7$ larger connected components (Fig.1).

Clique	*# papers (sampled)*	*# Scholars (sampled)*
1	8 (5)	87 (10)
2	1 (1)	37 (10)
3	4 (4)	104 (10)
4	2 (2)	30 (10)
5	4 (4)	64 (10)
6	7 (5)	36 (10)
7	1 (1)	26 (10)
Totals	**27 (22)**	**384 (70)**

Figure 1 - Network of collaboration and data on the $c = 7$ cliques.

The number of papers in each of the 7 subgraphs varies from 1 to 8 (Fig. 1). Therefore, we decided to take all the papers in the subgraphs containing 5 or less papers and $k_l = 5$ papers in the subgraph with more than 5 articles (by taking the most cited). It resulted $K = 22$ selected papers. From each of them we extracted authors, title, abstract and keywords, to form 22 cards, representing a synthesis of the most quoted papers dealing with different aspects of the Remote Working, useful for the development of Delphi's projections. The 7 selected cliques contain $N = 384$ researchers (Fig. 1). Then we extracted a random sample of $n_i = 10$ potential experts from each subgraph. In this first practical trial, for each expert we extracted only the number of published papers (P_i), therefore $w_1 = 1$, $w_2 = w_3 = 0$, $m_i = P_i$ and $\pi_i = m_i / \sum m_i$. The group of $n = 70$ scholars is therefore a prototype panel of experts, representative of 7 areas of research around the theme of remote working. Considering the semantic content of Titles and Keywords, a Semantic Network analysis on co-occurrence links was developed. Through Network analysis, using the Modularity algorithm [12], it was possible to identify $S = 7$ specific clusters of words (i.e., semantic areas). The different themes emerging from the network analysis can be grouped into the following research areas: Social Inequality (Cluster 0; Cluster 4), Psychological Disease (Cluster 3, Cluster 4), Work (Cluster 5), Technological Progress (Cluster 6, Cluster 7; Cluster 1).

4 Concluding remarks

The approach proposed here provides new quantitative advancements for the development of DBS, speeding up the preliminary review of the literature and improving the experts' selection task, by adopting a combination of text-mining and network analysis. For future works, we recommend implementing this contribution adopting different strategies to expand the data scanning. In fact, it could be useful to analyse much larger datasets, referring to different keywords used for the extraction of bibliometric data. Different textual data (e.g., from newspaper articles, web pages of projects, blogs etc.) could be analysed in order to have a more general perspective. For the identification of the experts, we limited the analysis to the network of scholars but the networks from which to draw the experts of the Delphi panel must certainly be many more (e.g., organisations, web pages, journalists, professional associations, networks of enterprises, trade unions, etc.), to increase the variability of competences.

The indicators and the weights to measure the expertise can be of various types. We have proposed three indicators and a simple weighted arithmetic mean, but for future work it could be interesting to study other indicators as well as other aggregation formulas. Finally, it is worth noting that a measure of the expertise is useful not only to sample the experts, but also in the subsequent phase of analysing the results of the Delphi, to weigh the answers given by the participants.

References

1. Smite, D., Tkalich, A., Moe, N.B., Papatheocharous, E., Klotins, E., & Buvik, M.P. Changes in perceived productivity of software engineers during COVID-19 pandemic: The voice of evidence. Journal of Systems and Software, **186** (2022)
2. Inayatullah, Sohail. (ed.). "The Views of Futurists." The Knowledge Base of Futures Studies 4. CD-ROM. Melbourne: Futures Study Centre 2000.
3. Linstone H.A., Turoff M. The Delphi method: Techniques and applications. Addison-Wesley Publishing Company, Reading, Massachusetts (1975)
4. Kosow H., Gaßner R., Methods of future and scenario analysis: overview, assessment, and selection criteria, Bonn, Deutsches Institut für Entwicklungspolitik, (2008).
5. Di Zio S., Bolzan M., Marozzi M. Classification of Delphi outputs through robust ranking and fuzzy clustering for Delphi-based scenarios, Technological Forecasting & Social Change, **173**, 121140 (2021) doi: https://doi.org/10.1016/j.techfore.2021.121140
6. Bishop P., Hines A., Collins T. The current state of scenario development: An overview of techniques, Foresight, 9(1), 5–25 doi: https://doi.org/10.1108/14636680710727516
7. Haegeman K., Marinelli E., Scapolo F., Ricci A., Sokolov A., Quantitative and qualitative approaches in Future-oriented, Technology Analysis (FTA): From combination to integration? Technological Forecasting & Social Change, **80**, 386–397 (2003) doi: 10.1016/j.techfore.2012.10.002
8. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. The Journal of machine Learning research, **3**, 993-1022 (2003).
9. Sossa, J. W. Z., Halal, W., & Zarta, R. H. Delphi method: analysis of rounds, stakeholder and statistical indicators. Foresight **5** 525-544 (2019) doi: https://doi.org/10.1108/FS-11-2018-0095
10. Lambiotte R., Delvenne J.-C., Barahona M. Laplacian Dynamics and Multiscale Modular Structure in Networks (2009) eprint arXiv:0812.1770
11. Aria, M., Cuccurullo, C. Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics, **11**(4), 959-975 (2017) doi: https://doi.org/10.1016/j.joi.2017.08.007
12. Lambiotte R., Delvenne J.-C., Barahona M. Laplacian Dynamics and Multiscale Modular Structure in Networks (2009) eprint arXiv:0812.1770

Emotion recognition in Italian political language to predict positionings and crises government

Il riconoscimento delle emozioni nel linguaggio politico italiano per predire posizionamenti e crisi di governo

Alessia Forciniti and Emma Zavarrone

Abstract The paper aims to analyze the political language adopted on Twitter by the main Italian parties' leaders during the first two waves of Covid-19 pandemic.
A two-step model based on sentiment emotion recognition (ER) and Correspondence analysis detected which emotions characterized the political language and which changes happened between the two waves. The results showed the use of a language with a strong emotional weight for some political actors as opposed to others who used a neutral register of political language in both waves. The comparison between two waves denoted a shift from anger to sadness and fear for Meloni and a moving away Salvini by predicting through ER the rift of the right-wing.

Abstract *L'articolo mira ad analizzare il linguaggio politico adottato su Twitter dai principali leader di partito italiani durante le prime due ondate della pandemia di Covid-19. Un modello di analisi a 2-step basato su sentiment emotion recognition (ER) e Analisi delle Corrispondenze, ha rilevato quali emozioni hanno caratterizzato il linguaggio politico e quali cambiamenti sono avvenuti tra le due ondate. I risultati hanno mostrato l'uso di un linguaggio con un forte peso emozionale per alcuni attori politici in opposizione ad altri che hanno utilizzato un registro del linguaggio politico neutro. Il confronto tra le due ondate ha denotato il passaggio dalla rabbia alla tristezza e paura per la Meloni e un distacco da Salvini preannunciando con l'ER la spaccatura del centro-destra.*

Key words: emotion recognition, sentiment analysis, political leadership, correspondence Analysis

[1] Alessia Forciniti, IULM University; alessia.forciniti@iulm.it
Emma Zavarrone, IULM University; emma.zavarrone@iulm.it

1 Introduction

The health emergency of Covid-19 represented a breaker point in terms of stability both in private life of individuals and in public domain, creating vulnerability in social, economic and political field. In this phase marked by a general perception of disorientation, the communication has assumed a decisive role in order to inform the citizens about the evolution of the pandemic and to explain the governments' containment measures aimed at addressing the behaviors of national communities. In Italy, communication has been broadcast both through traditional media and by means of social media, which have become the main channels for disseminating information and sharing opinions. In fact, during the lockdown period, in Italy the use of social media grown by 30% (Comscore, 2020). Institutional communication and specifically political parties' leaders have also followed this trend, concentrating their activities of communication on social media which, above all on Twitter, have found a large space of narration and representation (Paolillo & Forciniti, 2021). The language of politics represents the language of power (Lasswell, 1979) as exercise of persuasion and with the progressive affirmation of *media logic* has generated the definition of two registers of political language: one didactic - argumentative and the other polemical (Cepernich & Novelli, 2018) based on a very strong *affective* dimension. The register of political language used by institutional actors contributed to affect the so-called *connected public culture* (Boccia Artieri, 2012) through the fragmentation of the public sphere.

In this paper, we propose the study of the political language adopted on Twitter by the main Italian parties' leaders during the first two waves of the first phase of Covid-19 by means of a 2-step analysis model which combines methods for semantic orientation and emotion classification with the multivariate approach of Correspondence Analysis. Our contribution considers the difference existing between sentiment and emotion, the need to go beyond the classification of positive - negative polarity to better understand the semantic aspect enclosed in a text. More specifically, emotions have a central role in political discourse (Huguet Cabot *et al.*, 2020). Therefore, the research aims to classify the political language adopted by Italian leaders through the emotion recognition and the factorial projection to assess if the emotional approach is useful to predict changes in the political positioning and to anticipate crises of government.

The paper explores four main research questions:

RQ1) Which emotions characterized the Italian politicians' language during the first and second waves of pandemic on Twitter?

RQ2) The emotions detected are useful to characterize the use of register didactic - argumentative or polemical?

RQ3) Were there changes in the emotions adopted by the political leaders between the first and second wave?

RQ4) Can the recognition of emotions about political language predicts the political positioning and crises of government?

To follow, section 2 presents data; section 3 describes the methodology used to reach the research questions; section 4 shows results and discussion.

2 Data

We extracted the text content released by each political actor on own official Twitter account to investigate the four research questions. Data extraction was performed using the open-source programming interface Twitter *API* and the *library rtweet* (Kearney, 2019). The official accounts of seven political figures whose hierarchy and popularity dominated the national scene during the first two pandemic waves: Silvio Berlusconi, Giuseppe Conte, Luigi Di Maio, Giorgia Meloni, Matteo Renzi, Matteo Salvini and Nicola Zingaretti have been followed and downloaded. Two temporal slots have been considered: the first wave, from first February to 30 May 2020, and the second one, from first September to 30 November 2020. The monitoring period represents the main media hype about the pandemic, where both citizens and politicians are living uncertainty and bewilderment. Data extraction in two separate periods determined that two different corpora were obtained, thus, each corpus represents a different case of analysis. During the first wave 32,551 tweets have been obtained and the greater number, about it 69.16% (*Table 1*), comes from Matteo Salvini. With the second wave, we extracted 47,094 tweets, recording also in this case the Salvini's leadership in terms of posting on Twitter (74.10%).

Table 1: Percentage of tweets produced by leaders during the first and second wave

	Percentage of tweets	
Leader	*First wave*	*Second wave*
Silvio Berlusconi	7.47%	5.31%
Giuseppe Conte	3.39%	2.19%
Luigi Di Maio	5.64%	2.99%
Giorgia Meloni	9.02%	9.97%
Matteo Renzi	1.57%	2.28%
Matteo Salvini	69.16%	74.10%
Nicola Zingaretti	3.75%	3.16%

3 Methodology

To study the political language used by party leaders during the first and second pandemic wave and to determine if there were changes between the two waves, we adopted a 2-step analysis model.

At first step, we used emotion classification algorithms for Italian language. The literature on sentiment analysis is greater developed in English language than others which instead suffer the lack of resources, but Italian represents an interest evolving field (e.g.; Nozza *et al.,* 2020; Bianchi *et al*., 2021). In fact, after we cleaned the texts from link, URLs, emoji and other special characters, we used UmBERTo, a recent Italian Language Model trained with Whole Word Masking trained on

Commoncrawl ITA, a large number of Italian corpora implemented by means of Facebook Research Artificial Intelligence codes (Ott *et al*., 2019) and annotated with four basic emotions (Ekman, 1992): anger, fear, joy, sadness. The datasets consist in corpora of Italian tweets in a broad topic and domain coverage - from football matches to politics - for emotion and sentiment classification able to predict sentiments and emotions in text by training prediction models. UmBERTo uses two innovative approaches: the first, called Sentence Piece Model (SPM) is a language-independent sub-word tokenizer which creates sub-word units specifically to the size of the vocabulary through neural-based text processing; the second, denominated Whole Word Masking (WWM) applies a mask to an entire word, if at least one of all tokens was originally chosen as mask, so to maintain the sub-words. The task of emotion classification used is fine-tune called UmBERTo-FT, which obtains the best results in terms of the overall performance metrics (e.g.; accuracy, precision, recall) (Bianchi *et al.,* 2021).

At the second step, we coded information coming from the emotion recognition (ER) in a contingency table $\mathbf{T}_{lxe}$ where l indicates the leaders and e denotes the categories of emotions. The generic element t_{le} presents the total value of emotion detected into tweets of each leader for each emotion category.
Finally, the table $\mathbf{T}_{lxe}$ has been analyzed by the Correspondence Analysis (CA) (Benzécri, 1973; Greenacre, 1984). The performing of CA on $\mathbf{T}_{lxe}$ allowed to observe the association between leaders and categories of emotions in order to describe which emotions better characterized their political language at the time of Covid-19. At the same time, this allowed to detect similarities among leader according to emotions of political language adopted on Twitter.
This 2-step analysis has been performed separately both for the first wave corpus and for that the second wave in order to intercept changes in political language between the two periods examined.

4 Results and discussion

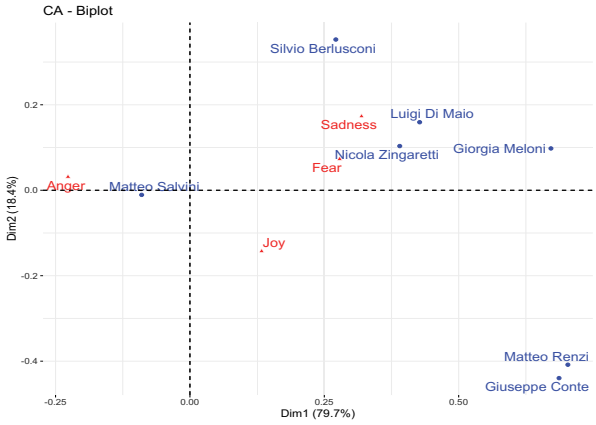
The CA map (*Figure 1*) shows the association between the leaders and the four categories of emotions (refers to Ekman, 1992) for the first pandemic wave. The factorial projection explains 77.3% of variance through the first dimension and 21.7% by means of the second dimension. We can observe the plane how split in four sections, each of which is characterized by a category of emotion ascribable to each leader. The emotional dimension can be interpreted by looking at the map from left to right. This interpretation makes it possible to determine a political language characterized by the shifting from anger to fear. More specifically, the exploratory analysis suggests that Silvio Berlusconi's political language was better represented by sadness; that adopted by Luigi di Maio and Nicola Zingaretti was instead imbued with fear; finally, Giorgia Meloni and Matteo Salvini's language were both associated to anger (RQ1).

These three association areas between political actors and emotions allowed to identify how the use of linguistic expressions with a strong emotional weight have a

central role in their political discourses, hence we can attribute to these leaders the use of register of political language very affective. However, we cannot ascribe any category of emotions to Matteo Renzi and Giuseppe Conte, thus we can find the use of a political language register that was not based on affective dimension during the first pandemic wave (RQ2).

Figure 1: Correspondence Analysis Map "leader by emotion" for the first wave

Figure 2: Correspondence Analysis Map "leader by emotion" for the second wave



The CA map referred to second pandemic wave (*Figure 2*) explains 79.7% of variance by means of the first dimension and 18.4% through the second dimension. On the horizontal axis, we affirm the transition from a political language based on anger to a language based on fear. In this case, we can intercept on the plane three sections which describe the emotions and political actors. More precisely, we observe the association between Luigi Di Maio, Nicola Zingaretti and Giorgia Meloni with emotions as fear and sadness; while Matteo Salvini's language remained characterized by anger (RQ1). No emotional category can be associated

with Matteo Renzi and Giuseppe Conte (RQ1), confirming a non-affective political language different from the other leaders mentioned above (RQ2).

Through the comparison between the CA maps of the first and second wave (*Figure 1* and *2*), it emerges a change in the political language, since in the *Figure 1* the language of Di Maio and Zingaretti was best characterized by fear, while in *Figure 2* we can see also sadness; as well as Meloni's language shifted from anger to sadness and fear. In fact, Meloni approached Di Maio and Zingaretti's language and moved away from Salvini. No change for Salvini, who remained characterized by anger, as well as Conte and Renzi that they are not associated with any category of emotions between the first and second wave (RQ3). ER in the political language adopted on Twitter between Meloni and Salvini in the second wave allowed to predict the rift and disaggregation of the right-wing coalition before the Draghi's government established in February 2021 (RQ4). An interesting aspect is the proximity between Conte and Renzi detected in the second wave, given that this last determined the Conte's government crisis. The first impression can indicate closeness in political strategic positioning as in the first wave. However, the map may show a different interpretation: the tactical outsider represented by Renzi, who based his language on emotionless register in the opposite to Meloni, Salvini, Zingaretti, to obtain consensus without polemics. Thus, in the second period, Renzi and Conte continued to use a similar neutral language but for different strategies.

References

1. Benzécri, J. P.: L'Analyse des données. Tome 2: l'analyse des correspondances. Dunod, Paris (1973).
2. Bianchi,F., Nozza, D., Hovy, D.: FEEL-IT: Emotion and Sentiment Classification for the Italian Language. Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 76--83 (2021).
3. Boccia Artieri, G.: Stati di connessione. Pubblici, cittadini e consumatori nella (Social) Network Society, Franco Angeli, Milano (2012).
4. Cepernich, C., Novelli, E.: Sfumature del razionale. La comunicazione politica emozionale nell'ecosistema ibrido dei media. Comunicazione politica (**19**)(1), 13--30 (2018).
5. Comscore: Coronavirus e nuovi comportamenti online: aggiornamento del 30 marzo 2020 (2020). Available via https://www.comscore.com/ita/Public-Relations/Blog/Coronavirus-e-nuovi-comportamenti-online-aggiornamento-del-30-marzo-2020. Last access: August 2021.
6. Ekman, P.: An argument for basic emotions. Cognition and Emotion, **6**(3-4):169--200 (1992).
7. Greenacre, M.: Theory and Applications of Correspondence Analysis. Academic Press (1984).
8. Kearney, M. W.: Rtweet: Collecting and analyzing Twitter data", Journal of Open Source Software, **4**(42), 18--29 (2019).
9. Lasswell, H.D.: Il linguaggio della politica: studi di semantica quantitativa, Nuova Eri, Torino (1979).
10. Nozza, D., Bianchi, F., & Hovy, D.: What the [MASK]? Making Sense of Language-Specific BERT Models. arXiv preprint Available via: https://arxiv.org/pdf/2003.02912.pdf (2020).
11. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, A.: fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Proceedings of NAACL-HLT 2019: Demonstrations (2019).
12. Paolillo, M., Forciniti, A.: L'impatto del Covid-19 sull'opinione pubblica: una strategia di analisi per lo studio della comunicazione su Twitter. In: in Favretto A.R., Maturo A., Tomelleri S. (eds.), L'impatto sociale del Covid-19, Franco Angeli, Milano (2021).

What does your self-description reveal about you?

A pipeline to analyse StockTwits users

Cosa rivela di te la tua autodescrizione?

Una procedura per analizzare gli utenti di StockTwits

Riccardo Ricciardi

Abstract In this study, we propose a pipeline to analyse heterogeneity in self-descriptions for the users of StockTwits, a microblogging platform about the stock market. By combining word and document embeddings with clustering methods, we aim to study what textual self-descriptions can reveal about users' self-declared trading information. Preliminary results show that some patterns in texts reflect trading characteristics and encourage to explore the relation between textual and non textual self-description.

Abstract *In questo studio, si mostra una procedura per l'analisi dell'eterogeneità presente nelle autodescrizioni degli utenti di StockTwits, una piattaforma di microblogging sui mercati azionari. Combinando tecniche di word e document embedding con metodi di clustering, si propone di studiare il contenuto informativo delle descrizioni testuali riguardante le caratteristiche di trading autodichiarate dagli utenti. I risultati preliminari mostrano che talune informazioni sono estraibili dai documenti di testo, incoraggiando l'esplorazione della relazione tra autodescrizione testuale e non testuale.*

Key words: Natural Language Processing, Clustering, StockTwits, Self-Description

1 Introduction

Since the 2000s, the capacity of computer and information systems, and the proliferation of both social media and commercial applications have driven the development of *Natural Language Processing* (NLP). It is the field of study aiming at processing data created by *natural language*, that is language naturally developed by individuals belonging to the same social group [8].

Riccardo Ricciardi
Department of Economics and Management, University of Brescia, Via S. Faustino 74/b-25122 Brescia, Italy, e-mail: r.ricciardi@unibs.it

Since the coexistence of different groups of users on those platforms, recent studies have explored heterogeneity in language use and have showed how language understanding has to vary based on speaker's and writer's characteristics [1] [5].

Compared with the multitude of studies on other social media, such as Facebook, Twitter, and Reddit, few studies have applied NLP and statistical techniques to study the activity on the StockTwits (ST) social media platform.

ST was founded in 2009, to let investors, traders, and finance enthusiasts share their ideas about the stock market. ST is a microblogging platform: people's activity consists of posting messages, images, videos and similar content, also called *tweets*, and interacting with other users' content. Since the stock market is the central topic in ST, users usually tag their tweets with *bullish* or *bearish*, depending on their confidence in stock price rising. Thus, literature about ST have focused on exploiting the large volume of tweets on ST to predict the stock market behavior and for *sentiment analysis* [10] [14].

For the best of our knowledge, there is not any study leveraging self-descriptions and the other users' self-declared information to find patterns in groups of users.

In this respect, when people join ST, they are asked to complete their profile with personal information and information about their trading strategy. *Inter alia*, users may fill their profile with a *bio*, a short self-description with a maximum length of 250 characters, and the clarification of which kind of traders they are, according to their experience and their primary holding period.

This paper starts to explore users' bios and proposes a pipeline to capture what self-descriptions can reveal about users' trading characteristics.

For processing texts as data it is crucial the decision about the way of representing text documents. Before the spread of neural networks in NLP, the conventional approach was based on word frequencies: a document was represented as a vector composed of frequencies of words in that document [15]. Since these vectors are based on all the words in the available text corpus, this approach yields a high-dimensional and largely sparse document matrix, unable to capture the semantic relationships between words [12].

To solve these drawbacks, new methods were introduced. Based on the *distributional hypothesis* according to which "*a word is characterized by the company it keeps*" [4], researchers proposed to represent words with low-dimensional vectors, also known as *embeddings*, such that words appearing in similar contexts have similar vector representations [9] [11]. In particular, [9] proposed a method that leverages a simple and efficient neural network to obtain word vector representations as a by-product of a predictive task, and in a similar way, [7] introduced an approach to represent text documents in the same vector space. The latter technique is used in this study to train word and document embeddings.

Document embeddings can be used as input of clustering, and thus documents can be grouped through a similarity measure [6]. Several studies analysed clustering efficiency by mixing different document vectorization techniques and clustering algorithms [13] [3]. In particular, [13] highlighted that clustering algorithms benefit from low-dimensional representations of documents, whereas representations based on word frequencies suffer from the *curse of dimensionality*. In the pipeline

of the analysis proposed in this paper, we considered to compare the well-known *K-Means*, *K-Medoids* clustering and some their adaptations to find groups among documents [2] [6].

2 Data and methods

In this section, we describe the process of data extraction, filtering, and sampling, and moreover, the methodologies proposed to study the users' self-descriptions.

2.1 The data pipeline

Through the StockTwits's API[1], we accessed all the tweets from 2010 to 2021. In addition to the particular content and the interactions' records connected to it, each tweet keeps the information about the user; from it we extracted text documents and trading features for each user. First, we considered only users writing an English bio with a minimum length of 6 words. In this way, we obtained a filtered users database composed of about 25,000 people, i.e. 25,000 text documents.

Since we were interested in finding relationships between semantic groups and groups of users based on trading information, we considered users' information about trading to form groups of users. On the one hand, based on their experience, users define themselves as *novice*, *intermediate*, or *professional* traders. On the other hand, based on his/her primary holding period, i.e. the amount of time he/she usually holds stocks in portfolio, a user define themselves as *day trader*, *swing trader*, *position trader*, or *long term investor*. Finally, before fitting an NLP model on bios, we sampled 2,400 users from the database by balancing for trading experience and primary holding period.

2.2 Methods

After sampling, bios were preprocessed, by removing email addresses, URLs, symbols, and stopwords. The latter removal was obtained by using the well-know *Snowball* list of stopwords for English from the `stopwords` *R* package [2].

In this study we trained word and document vectors by using the *paragraph vector* model [7]. This model uses a feedforward neural network with one hidden layer to fully parameterise the word and document vectors. Thus, at the end of the learning process, each word w is described as a vector of the form:

[1] https://firestream-portal.stocktwits.com/

[2] https://cran.r-project.org/web/packages/stopwords/index.html

$$v(w) = (x_{w1}, x_{w2}, ..., x_{wp}) \tag{1}$$

and each document d as a vector of the form:

$$v(d) = (x_{d1}, x_{d2}, ..., x_{dp}) \tag{2}$$

where p is the dimension of the embedding.

To train these parameters, the network initialises them randomly, and then it adjusts them in the learning process whose objective is to maximise the conditional probability:

$$p(w_t | w_{t-\frac{s}{2}}, w_{t-\frac{s}{2}+1}, w_{t-\frac{s}{2}+2}, ..., w_{t+\frac{s}{2}-2}, w_{t+\frac{s}{2}-1}, w_{t+\frac{s}{2}}, d) \tag{3}$$

where w_t is the target word, $w_{t-\frac{s}{2}}, w_{t-\frac{s}{2}+1}, ..., w_{t+\frac{s}{2}-1}, w_{t+\frac{s}{2}}$ are the context words the occur around w_t with a window size of s, and d is the document in which all these words appear.

To find groups in document vectors we propose to compare the performance of *K-Means* and *K-Medoids* clustering, by using Euclidean distance and cosine dissimilarity, respectively, as measures of dissimilarity between documents; moreover, we propose *Spherical K-Means*, that lets us adapt *K-Means* to use the cosine dissimilarity, and *DBSCAN* clustering, in order to consider arbitrary shapes for clusters and the presence of outliers when partitioning objects.

Thus, in addition to internal validity measures for clustering, we propose to adopt an external validation on the entire pipeline, by using resampling methods to obtain new balanced samples.

3 Results and discussion

Here we describe the results of this exploratory analysis. In particular, we show here preliminary results about similarities among words and groups of documents, whereas in the near future we will show the results about document clustering, whose elaboration is in progress.

In Table 1 with respect to five different words, we illustrate the composition of the 10% most similar users by both trading experience and primary holding period, and, for the same terms, in Table 2, we illustrate the most similar words. Cosine similarity was used for this purpose.

First, for the terms *professional*, *intermediate*, and *novice*, among the 10% most similar bios, the majority is composed of users defining themselves as professional, intermediate, and novice traders, respectively. Moreover, we can see that moving from *professional* to *novice*, most similar words get far from the idea of ”success” and ”motivation”. In particular, *novice* is semantically close to contexts quite different from trading, like sports and hobbies.

It’s interesting to note that one of the most similar words for *professional* is *join*. Indeed, in many bios users invite people to join them either by following them on

ST or by clicking on their website and subscribing to their newsletter. It's quite expected that professionals spam their profile in their bios. Notice that users defining themselves as professionals can be either true professionals or not true professionals, but also bots.

In reference to the term *swing*, it is interesting to note how the most similar bios mostly come from day traders and swing traders, i.e. the kind of traders that usually holds stocks for short periods, from one day to one month. This kind of traders has to be always *active alert* on the *liquid* market, where there are large volume of operations, many agents, and low transaction costs.

Eventually, *planning* is mostly similar to bios written by long-term traders, and the most semantically similar word is *mortgage*, which is a long-term investment.

These preliminary results show that some patterns in bios can reflect group memberships, and so they encourage to explore the relation between language use in self-description and trading information. Future steps will cover both preprocessing steps and cluster analysis: on the one hand, words composed of more than one term, i.e. *bigram* and *trigram*, will be included; on the other hand, clustering results will be explored and resampling methods will be considered to validate the entire analysis pipeline.

Table 1 Trading groups' composition (%) of the 10% most similar bios with respect to five words in the vocabulary.

word	**10% most similar users**			
professional	**Intermediate**	**Novice**	**Professional**	
	12.9	24.6	62.5	
	Day Trader	**Long Term Trader**	**Position Trader**	**Swing Trader**
	23.3	22.1	12.5	42.1
intermediate	**Intermediate**	**Novice**	**Professional**	
	45.4	27.1	27.5	
	Day Trader	**Long Term Trader**	**Position Trader**	**Swing Trader**
	34.2	15.8	23.8	26.2
novice	**Intermediate**	**Novice**	**Professional**	
	32.1	54.6	13.3	
	Day Trader	**Long Term Trader**	**Position Trader**	**Swing Trader**
	24.2	25.8	28.3	21.7
swing	**Intermediate**	**Novice**	**Professional**	
	39.2	26.2	34.6	
	Day Trader	**Long Term Trader**	**Position Trader**	**Swing Trader**
	32.5	19.2	16.2	32.1
planning	**Intermediate**	**Novice**	**Professional**	
	37.9	36.7	25.4	
	Day Trader	**Long Term Trader**	**Position Trader**	**Swing Trader**
	20.8	36.7	27.1	15.4

Table 2 Most similar single terms and maximum similarity with respect to five different words in the vocabulary

word	Most similar words	Max similarity
professional	*join, succeed, recommended, advanced, executive, complex, improvement, seminars*	0.49
intermediate	*follower, constantly, basic, methodology, gap, mental, direction*	0.68
novice	*second, junkie, football, travel, hockey, reads, dogs, gamer, cooking*	0.52
swing	*wide, active, alert, day, liquid, profitable*	0.57
planning	*mortgage, concentration, junkie, enthusiast, pilot*	0.64

References

1. Bamman, D., Dyer, C., Smith, N.A.: Distributed Representations of Geographically Situated Language. In: ACL. DOI 10.3115/v1/P14-2134
2. Bock, H.H.: Clustering methods: A history of k-means algorithms pp. 161–172
3. Curiskis, S.A., Drake, B.L., Osborn, T., Kennedy, P.J.: An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit DOI 10.1016/J.IPM.2019.04.002
4. Firth, J.R.: A synopsis of linguistic theory
5. Hovy, D.: Demographic Factors Improve Classification Performance. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 752–762. Association for Computational Linguistics. DOI 10.3115/v1/P15-1073
6. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons
7. Le, Q.V., Mikolov, T.: Distributed Representations of Sentences and Documents
8. Liu, B.: Sentiment Analysis and Opinion Mining **5**(1), 1–167. DOI 10.2200/S00416ED1V01Y201204HLT016
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space
10. Oliveira, N., Cortez, P., Areal, N.: On the predictability of stock market behavior using stocktwits sentiment and posting volume. In: Portuguese Conference on Artificial Intelligence, pp. 355–365. Springer
11. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics. DOI 10.3115/v1/D14-1162
12. Pilehvar, M.T., Camacho-Collados, J.: Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning **13**(4), 1–175. DOI 10.2200/S01057ED1V01Y202009HLT047
13. Radu, R.G., Iulia-Maria, R., Truică, C.O., Apostol, E.S., Mocanu, M.: Clustering Documents Using the Document to Vector Model for Dimensionality Reduction. DOI 10.1109/AQTR49680.2020.9129967
14. Renault, T.: Intraday online investor sentiment and return patterns in the US stock market **84**, 25–40
15. Turney, P.D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics **37**, 141–188. DOI 10.1613/jair.2934

Variable selection and complete matrix approaches

A Statistical Approach for the Completion of Input-Output Tables

Un approccio statistico per il completamento delle matrici Input-Output

Rodolfo Metulini, Giorgio Gnecco, Francesco Biancalani and Massimo Riccaboni

Abstract The methodologies adopted by national statistical offices to collect data of within- and cross-country Input-Output (I/O) economic relations raises the issue of obtaining reliable data in a timely fashion, making the reconstruction of I/O tables of particular interest. In this work, we propose a method combining Hierarchical Clustering and Matrix Completion to impute missing entries of a partially unknown I/O matrix. Through a validation study based on World Input-Output Database as well as on synthetic matrices, we show the effectiveness of the proposed method to impute missing values from both previous years data and current data related to countries similar to the one for which current data are missing.

Abstract *Le metodologie adottate dagli uffici statistici nazionali per raccogliere i dati sulle relazioni economiche di Input-Output (I/O) pongono il problema di ottenere dati affidabili in modo tempestivo. La ricostruzione delle tabelle I/O risulta perciò di particolare interesse. Proponiamo un approccio statistico che combina clustering gerarchica e Matrix Completion allo scopo di imputare i dati mancanti di una matrice I/O. Attraverso uno studio di validazione basato su World Input-Output Database e matrici sintetiche, mostriamo l'efficacia del metodo proposto per imputare i valori mancanti utilizzando sia dati dello stesso paese negli anni precedenti sia dati dello stesso periodo relativi a paesi simili a quello che presenta dati mancanti.*

Key words: Matrix Completion, Hierarchical Clustering, missing data, LASSO-like Nuclear Norm Penalty.

Rodolfo Metulini
Department of Economics and Statistics (DISES), University of Salerno, Via Giovanni Paolo II, 132, 84084, Fisciano, SA, Italy, e-mail: rmetulini@unisa.it

Giorgio Gnecco · Francesco Biancalani · Massimo Riccaboni
Laboratory for the Analysis of CompleX Economics Systems (AXES), IMT School for Advanced Studies, Piazza S. Francesco, 19 - 55100 Lucca, Italy, e-mail: giorgio.gnecco@imtlucca.it; francesco.biancalani@alumni.imtlucca.it; massimo.riccaboni@imtlucca.it

1 Introduction

I/O tables portray flows of what is produced by some economic agents and consumed by other economic agents (either as intermediate or final consumption). Literature refers to the study of these flows with the term "Input-Output (I/O) analysis". Our work has to do with the completion of I/O matrices. The interest arises if one takes into account the aspects related to the methodology of data collection adopted by national statistical offices. Direct methods (e.g., surveys) lead to sampling errors. Moreover, to collect data in timely fashion, historical data are typically used to approximate the current ones [4]. Direct compilation of I/O tables also rely on administrative registers and sectorial reports. It follows that many different statistical sources are contrasted. This is a huge resources consuming operation, and the time gap between the publication of the data and its reference year is one of the reasons why also indirect methods exist [7]. In this work we propose to apply Matrix Completion (MC, [1]) to reconstruct part of an I/O table. Some studies highlighted the emergence of a clustering pattern among either countries or sectors [8]. Due to this reason, the I/O matrix might be well-approximated by a low-rank one. This low-rank property suggests, among other possible statistical techniques, the adoption of MC, which permits, through a choice of the regularization parameter, to select the number of non-zero singular values to be kept in the reconstructed matrix. Nevertheless, the application of MC to a full I/O table is not straightforward, as we found a complex structure characterized by sparsity and by a clear separation between i) large-to-large countries' values and small-to-small countries' ones, ii) within-country values and cross-country ones. This suggests performing a pre-processing step to restrict the application of MC to suitable groups of similar countries. We do so via a Hierarchical Clustering (HC, [5]) approach along with the use of Average Absolute Correlation Distance (AACD), highly related to the functioning of MC. The performance of our approach is validated in terms of the Root Mean Square Error (RMSE) - on a 5-year panel of I/O submatrices, where a known part of the matrix associated with a specific year has been artificially obscured. Results based on an application to real I/O tables, show the effectiveness of the proposed method to predict missing values in the current I/O matrix from both previous years data and current data related to similar countries. In contrast, the effectiveness reduces if similar countries are replaced by the ones belonging to quite different clusters.

2 Methods

MC is a statistical method used to predict unobserved entries of a matrix in terms of the set of the remaining observed entries. Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, MC works by finding a suitable low-rank approximation of $\mathbf{M}$, by assuming the following model:

$$\mathbf{M} = \mathbf{C}\mathbf{G}^T + \mathbf{E}, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{m\times r}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, whereas $\mathbf{E} \in \mathbb{R}^{m\times n}$ is a matrix of errors. We find the rank-r approximating matrix $\mathbf{C}\mathbf{G}^T$ by solving the following MC optimization problem based on a LASSO (Least Absolute Shrinkage and Selection Operator)-like nuclear norm penalty [2, 3]:

$$\underset{\hat{\mathbf{M}}\in\mathbb{R}^{m\times n}}{\text{minimize}} \left(\frac{1}{2} \sum_{(i,j)\in\Omega^{\text{tr}}} \left(M_{i,j} - \hat{M}_{i,j}\right)^2 + \lambda \|\hat{\mathbf{M}}\|_* \right), \tag{2}$$

where Ω^{tr} is a subset of pairs of indices (i, j) corresponding to positions of known entries of $\mathbf{M}$, $\hat{\mathbf{M}}$ is the matrix to be optimized, $\lambda \geq 0$ is a regularization constant, and $\|\hat{\mathbf{M}}\|_*$ is the nuclear norm of the matrix $\hat{\mathbf{M}}$. The regularization constant λ controls the trade-off between fitting the known entries of the matrix $\mathbf{M}$ and achieving a small nuclear norm. We solve equation (2) using the state-of-the-art "Soft Impute" algorithm [2]. We select λ by first randomly dividing the set of positions of unobserved entries of the matrix $\mathbf{M}$ into a validation set Ω^{val} (about 25% of the positions) and a test set Ω^{test} (the positions of the remaining entries)[1]. Then, the optimization problem is solved for several choices $\lambda_k \sim 2^{k/2-10}$, for $k = 1,\ldots,40$. Given

$$RMSE^{\text{s}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\text{s}}|} \sum_{(i,j)\in\Omega^{\text{s}}} \left(M_{i,j} - \hat{M}_{\lambda_k,i,j}\right)^2}, s = val, test, tr, \tag{3}$$

for each λ_k, RMSE of MC on the training set is computed, then the λ_k° that minimizes $RMSE^{\text{val}}_{\lambda_k}$ is found. The RMSE of MC on the test (and validation) set is then computed in correspondence of λ_k°. Clustering is used to find the groups of similar countries, for which the associated blocks of I/O tables are going to form matrix $\mathbf{M}$. We adopt a HC approach because alternative methods (e.g., k-means) require assumptions on the homoschedasticity, the spherical variance of the variables and clusters' dimension. HC performs well even when those assumptions are not satisfied. We use the Average Absolute Correlation Distance (AACD) because it is related to the formulation of the MC optimization problem, since it quantifies the average linear dependence of corresponding columns of blocks associated with different countries[2]. Given n and l to be the number of intermediate and final sectors,

$$AACD_{c_1,c_2} = 1 - \frac{\sum_{i=1}^{n+l} |corr(\mathbf{b}_i^{c_1}, \mathbf{b}_i^{c_2})|}{n+l} \tag{4}$$

is the complementary to 1 of the absolute correlation between the i-th columns $\mathbf{b}_i^{c_1}$ and $\mathbf{b}_i^{c_2}$ of $n \times (n+l)$ blocks associated with countries c_1 and c_2, averaged with respect to the columns. In our clustering step we compare any two countries c_1 and c_2 (except country c_h) in terms of what their $n+l$ sectors use from the n intermediate sectors of country c_h.

[1] The union of the validation and test sets corresponds to the positions of the artificially obscured entries (but available as a ground truth), the training set to the positions of the remaining entries.

[2] Moreover AACD, differently from the l_1 and l_2 norms, is not affected by country's dimension.

3 Application of clustering and matrix completion

We present an application of the proposed method based on World Input-Output Database (WIOD) release 2016 [6], considering 43 world countries and 61 sectors (56 intermediate, 5 finals) over the period 2000–2014. The main diagonal blocks (also called domestic blocks) are excluded, since they are characterized by much larger values than the other blocks and may cause troubles in the application of MC. To obtain groups of similar countries for the application of MC, simulations based on synthetic I/O tables have been performed to determine the proper number of clusters. Specifically, we present the case in which countries are grouped with respect to the similarity in terms of inputs from Italy. In doing so, we consider a 56×2562 matrix, where 56 is the number of italian industries and 2562 is the product of 61 industries and 42 countries (all except Italy), reporting all sectors' inputs from Italy to all other countries[3]. For the application of MC we consider groups of similar countries according to the dendrogram in Figure 1, in turn based on the HC performed on WIOD data and using the number of clusters as from the simulations. We start from WIOD tables relative to years 2010–2014 and we keep the information associated with a subset of countries (keeping only the off-main diagonal blocks). Then, for one of these pairs of countries, the information about the last year is obscured, and one tries to reconstruct it by MC. We consider a group of 4 similar countries (belonging to the same cluster, Austria, Belgium, Germany, and the Netherlands) and a group of 4 dissimilar ones (belonging to 4 different clusters, Australia, Belgium, Japan, and Malta). The MC algorithm is first applied

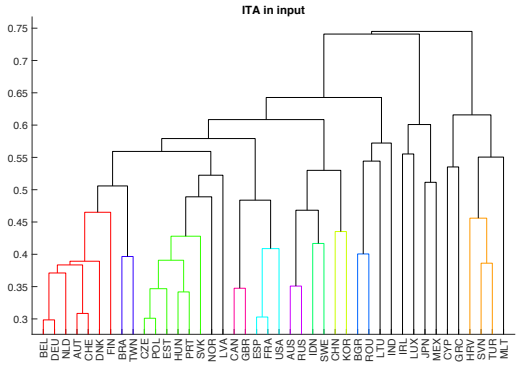


Fig. 1 Dendrogram of countries in terms of input from Italy, WIOD tables (years 2010–2013). HC with AACD (y-axis) and complete linkage. 21 desired groups. Countries in the same cluster are depicted with the same color. Countries in singleton clusters are highlighted in black.

[3] We perform HC with AACD and complete linkage on $N = 1000$ synthetic I/O matrices generated by adding normally distributed random terms ε to the WIOD submatrix above defined (where each element of the matrix $\varepsilon_{i,j} \sim \mathcal{N}(0, sd_\varepsilon), \forall i, \forall j$, sd_ε being a Gamma(1, 1) to obtain different levels of variability. We choose the optimal number of clusters to be the minimum number such that $\frac{WSS}{TSS} < 0.5$, WSS is the "Within-cluster sum of squares" and TSS is the "Total sum of squares".

to the submatrix of similar countries (top of Table 1), where the elements related to a specific output country in year 2014 (in bold) are obscured. To show a comparison of MC performance when dissimilar countries are considered, we also apply MC to the submatrix at the bottom of Table 1 (in bold the obscured part). As expected, the results show a better performance of the MC algorithm, measured in terms of the reduction of the RMSE on the validation set from the base case of $\lambda \simeq 0$ to the optimal choice of λ, in the case of similar countries.

Table 1 Structure of the WIOD submatrix. Similar countries (top); dissimilar countries (bottom).

I/O, year			
ITA/AUT, 2010	ITA/BEL, 2010	ITA/DEU, 2010	ITA/NDL, 2010
ITA/AUT, 2011	ITA/BEL, 2011	ITA/DEU, 2011	ITA/NDL, 2011
ITA/AUT, 2012	ITA/BEL, 2012	ITA/DEU, 2012	ITA/NDL, 2012
ITA/AUT, 2013	ITA/BEL, 2013	ITA/DEU, 2013	ITA/NDL, 2013
ITA/AUT, 2014	ITA/BEL, 2014	ITA/DEU, 2014	ITA/NDL, 2014

I/O, year			
ITA/AUS, 2010	ITA/BEL, 2010	ITA/JPN, 2010	ITA/MLT, 2010
ITA/AUS, 2011	ITA/BEL, 2011	ITA/JPN, 2011	ITA/MLT, 2011
ITA/AUS, 2012	ITA/BEL, 2012	ITA/JPN, 2012	ITA/MLT, 2012
ITA/AUS, 2013	ITA/BEL, 2013	ITA/JPN, 2013	ITA/MLT, 2013
ITA/AUS, 2014	ITA/BEL, 2014	**ITA/JPN, 2014**	ITA/MLT, 2014

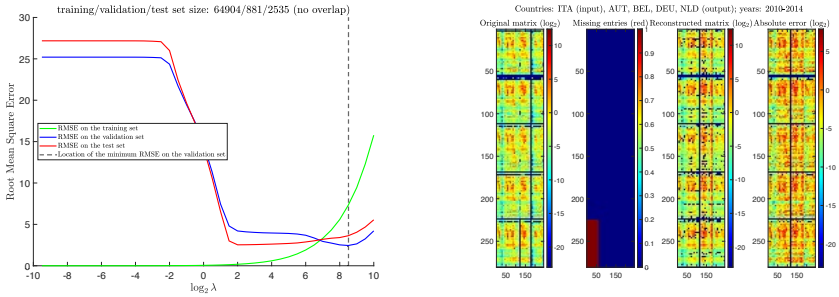


Fig. 2 RMSE results of MC for the WIOD submatrix in Table 1 (left); visualization of the original elements, positions of the missing entries, reconstruction obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (right).

4 Concluding remarks

This work represents the first attempt to adopt matrix completion combined with hierarchical clustering to forecast missing entries in I/O tables. The effectiveness

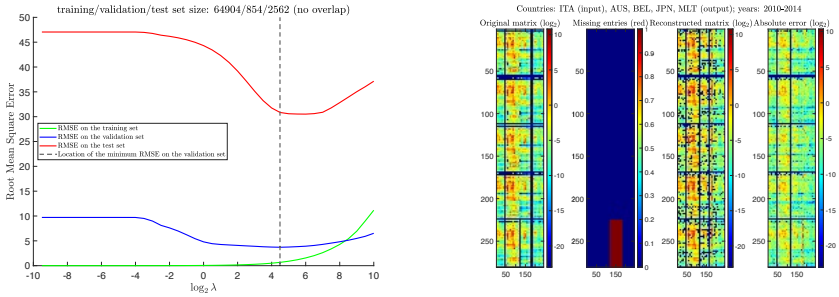


Fig. 3 RMSE results of MC for the WIOD submatrix in Table 1 (left); visualization of the original elements, positions of the missing entries, reconstruction obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (right).

of the proposed method according to historical data available from previous years has been demonstrated for I/O subtables obtained by selecting similar countries. A first possible extension of the analysis concerns comparing matrix reconstruction of I/O tables based on the repeated application of matrix completion to several I/O subtables, instead of a single more computationally expensive and (presumably) less effective application to the whole table. Our approach could be compared with other imputation methods for missing entries in panel data models, which is left for future research.

References

1. Hastie, T., Tibshirani, R., & Wainwright, M., Statistical learning with sparsity: the Lasso and generalizations. CRC Press, New York (2015)
2. Mazumder, R., Hastie, T., & Tibshirani, R., Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11, 2287–2322 (2010)
3. Negahban, S., & Wainwright, M. J., Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. Journal of Machine Learning Research, 13(1), 1665–1697 (2012)
4. Percoco, M., Hewings, G., & Senn, L., Structural change decomposition through a global sensitivity analysis of input-output models. Economic Systems Research, 18(2): 115–131 (2006)
5. Revelle, W., Hierarchical cluster analysis and the internal structure of tests. Multivariate Behavioral Research, 14(1), 57–74 (1979)
6. Timmer, M. P., Los, B., Stehrer, R., & De Vries, G. J., An anatomy of the global trade slowdown based on the WIOD 2016 release. GGDC research memorandum number 162, University of Groningen (2016)
7. Valderas-Jaramillo, J. M., Rueda-Cantuche, J. M., Olmedo, E., & Beutel, J., Projecting supply and use tables: new variants and fair comparisons. Economic Systems Research, 31(3), 423–444 (2019)
8. Zhu, Z., Morrison, G., Puliga, M., Chessa, A., & Riccaboni, M., The similarity of global value chains: A network-based measure. Network Science, 6(4), 607–632 (2018)

On multivariate records over sequences of random vectors with Marshall-Olkin dependence of components

Record multivariati su successioni di vettori aleatori con dipendenza di tipo Marshall-Olkin

A. Khorrami Chokami and S. A. Padoan

Abstract Let us consider a sequence of independent and identically distributed random variables on the real line with a joint continuous distribution function F. The n-th random variable is called a record if it is greater than all the preceding ones. The stochastic behavior of the sequence of subsequent records is well known. Falk, Khorrami Chokami and Padoan (2018) considered multivariate records over sequences of random vectors with independent components, providing results on the arrival times process of the records. In this work, we study the case of records over sequences of random vectors where the dependence among their components is of the Marshall-Olkin type.

Abstract *Si consideri una successione di variabili aleatorie indipendenti e identicamente distribuite sulla retta reale, con distribuzione continua* F. *L'* n*-esima variabile aleatoria è detta record se è maggiore delle precedenti. La successione dei record nel caso univariato è stata ampiamente studiata in passato. Un'estensione al caso multivariato si ha in Falk, Khorrami Chokami e Padoan (2018), in cui vengono forniti risultati riguardanti il processo dei tempi d'arrivo di record multivariati su successioni di vettori aleatori indipendenti, con componenti indipendenti. In questo lavoro, invece, gli autori studiano i record multivariati su successioni di vettori aleatori indipendenti, ma con dipendenza fra le loro componenti di tipo Marshall-Olkin.*

Key words: Complete records, Marshall-Olkin dependence.

A. Khorrami Chokami
University of Turin, Corso Unione Sovietica, 218 Bis, 10134, Torino, Italy, e-mail: amir.khorramichokami@unito.it

S. A. Padoan
Department of Decision Sciences, Bocconi University of Milan, via Roentgen 1, 20136 Milano, Italy, e-mail: simone.padoan@unibocconi.it

1 Introduction

Let us consider a sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots$. The rv X_m is called a *record* if $X_m > \max(X_1, \ldots, X_{m-1})$, $m \geq 2$. Clearly, X_1 is a record. Records among a sequence of i.i.d. rv on the real line have been investigated extensively over the past decades, see [5, Sections 6.2 and 6.3] and [1]. It is, for example, well known that in the univariate case the indicator functions $I_m^{\mathrm{R}} = \mathbb{1}(X_m > \max(X_1, \ldots, X_{m-1}))$, $m \in \mathbb{N}$, are independent, see, e.g., [5, Lemma 6.3.3]. This implies in particular for integers $i \neq k$

$$\Pr(I_j^{\mathrm{R}} = 1, I_k^{\mathrm{R}} = 1) = \Pr(I_j^{\mathrm{R}} = 1) \Pr(I_k^{\mathrm{R}} = 1) = j^{-1}k^{-1}. \tag{1}$$

More recently, [2] found new results concerning records over sequences of random variables, in the novel framework of not knowing their position in the sequence of records. This new approach is adopted in [4] to address the problem of computing finite distributions of both single and joint records over stationary Gaussian processes.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. random vectors (rv) in $\mathbb{R}^d$. Put for integers $j < k$

$$\boldsymbol{M}_j := \max_{1 \leq i \leq j} \boldsymbol{X}_i, \quad \boldsymbol{M}_j^k := \max_{j \leq i \leq k} \boldsymbol{X}_i,$$

where the maximum is taken componentwise. All our operations on vectors $\boldsymbol{x} = (x_1, \ldots, x_d)$, $\boldsymbol{y} = (y_1 \ldots, y_d)$, such as $\boldsymbol{x} < \boldsymbol{y}$, are meant componentwise.

Trying to generalize the concept of records to rv $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ in $\mathbb{R}^d$, the question which arises is how to define records in higher dimensions. Various definitions of records have been proposed in the past, see for example [6]. In this work, we consider the case of complete records: a rv $\boldsymbol{X}_m$ is a *complete* record if each component is a record itself, i.e., $\boldsymbol{X}_m > \boldsymbol{M}_{m-1}$.

Multivariate records have not been discussed that extensively, yet they have been approached by e. g. [7] or [1, Chapter 8]. For supplementary material on multivariate and functional records we refer to the thesis by [8].

We denote the event that $\boldsymbol{X}_m$ is a complete record by the indicator function $I_m^{\mathrm{CR}} := \mathbb{1}(\boldsymbol{X}_m > \boldsymbol{M}_{m-1})$. Results concerning complete records over sequences of random vectors with independent components can be found in [3].

This paper investigates, among others, the problem of which results for univariate records *do not* carry over to the multivariate case.

2 Marshall-Olkin

We have independence of record events in the univariate case and in the multivariate case with independent margins. One might conjecture that this is true, if we consider a convex combination of these two. The Marshall-Olkin standard max-stable distributions exactly provide this approach. When the elements of $\boldsymbol{X}$ are completely de-

pendent, then the probability that two joint records take place is equal to (1), while in the case the elements of $\boldsymbol{X}$ are independent we have

$$\Pr(I_j^{\mathrm{CR}} = 1, I_k^{\mathrm{CR}} = 1) = j^{-d}k^{-d}. \tag{2}$$

Therefore, one can conjecture that it also holds true when the elements of $\boldsymbol{X}$ are dependent. We show with a counter-example that this is not true. The relevant aspects of choosing the Marshall-Olkin copula lie in the fact that convex combinations of of complete dependence and independence structures represent a natural extension in exploring the properties of random vectors, and the parameter λ which rules the dependence is a flexible tool to have an idea on the role of the strength of the dependence. Moreover, as remarked in [1], the problem of studying records is the immediate difficulty of the necessary computations as soon as we move from the independence cases to better model the reality. The Marshall-Olkin distribution lets, not without complexities, to obtain exact formulas for complete records' occurrences, thus permitting comparisons with the known cases.

For simplicity assume $d = 2$. Suppose that $\boldsymbol{\eta}$ is a two-dimensional Marshall-Olkin distribution function:

$$\Pr(\boldsymbol{\eta} \leq \boldsymbol{x}) = \exp\left(-\{\lambda \left\|\boldsymbol{x}\right\|_\infty + (1-\lambda)\left\|\boldsymbol{x}\right\|_1\}\right), \qquad \boldsymbol{x} \leq \boldsymbol{0} \in \mathbb{R}^2, \tag{3}$$

where $\lambda \in [0,1]$ is a dependence parameter. The components of $\boldsymbol{\eta}$ are completely dependent when $\lambda = 1$, while they are independent when $\lambda = 0$.

Theorem 1. *Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots$ be i.i.d. rv with distribution function* (3)*. Then,*

$$\Pr(I_n^{CR} = 1) = \frac{1}{n(2-\lambda)}\left(\frac{2(1-\lambda)}{n} + \lambda\right) \tag{4}$$

and

$$\Pr(I_{j,k}^{CR} = 1) = \frac{1}{j^2k(2-\lambda)}\left(\frac{2(1-\lambda)}{k(2-\lambda)}\frac{(2-\lambda)k + (2-\lambda)(1-\lambda)j - \lambda(1-j)k}{k+(1-\lambda)j} + \lambda(2(1-\lambda)+j\lambda)\right). \tag{5}$$

Remark 1. When $\lambda = 0$ equation (5) becomes

$$\Pr(I_{j,k}^{\mathrm{CR}} = 1) = \frac{1}{j^2k^2},$$

while

$$\Pr(I_{j,k}^{\mathrm{CR}} = 1) = \frac{1}{jk},$$

when $\lambda = 1$, as expected.

Now, note that

$$\Pr(I_j^{\mathrm{CR}}=1)\Pr(I_k^{\mathrm{CR}}=1)=\frac{1}{jk(2-\lambda)^2}\left(\frac{2(1-\lambda)}{j}+\lambda\right)\left(\frac{2(1-\lambda)}{k}+\lambda\right)$$

and from this and (5) we have

$$\begin{aligned}&\Pr(I_{j,k}^{\mathrm{CR}}=1)-\Pr(I_j^{\mathrm{CR}}=1)\Pr(I_k^{\mathrm{CR}}=1)\\&=\frac{(\lambda-1)\lambda(k^2(\lambda(j-2)+2)-(\lambda-1)jk(\lambda(j-2)+2)+2(\lambda-1)(j-1)j)}{(\lambda-2)^2j^2k^2((\lambda-1)j-k)}\end{aligned}$$

and therefore the events $(I_j^{\mathrm{CR}}=1)$ and $(I_k^{\mathrm{CR}}=1)$ are not independent.

Proof (Theorem 1).

$$\begin{aligned}\Pr(I_n^{\mathrm{CR}}=1)&=\Pr(\boldsymbol{\eta}_n>\boldsymbol{M}_{n-1})=\int_{(-\infty,0]^2}\Pr(\boldsymbol{\eta}<(n-1)\boldsymbol{x})\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{x})\\&=\int_{-\infty}^{0}\int_{-\infty}^{x_2}(1-\lambda)\,\mathrm{e}^{n(x_1+(1-\lambda)x_2)}\,\mathrm{d}x_1\mathrm{d}x_2\\&+\int_{-\infty}^{0}\int_{x_2}^{0}(1-\lambda)\,\mathrm{e}^{n(x_2+(1-\lambda)x_1)}\,\mathrm{d}x_1\mathrm{d}x_2\\&+\int_{-\infty}^{0}\lambda\,\mathrm{e}^{n(2-\lambda)x}\,\mathrm{d}x\\&=\frac{1-\lambda}{2-\lambda}\frac{1}{n^2}+\frac{1}{n^2}-\frac{1}{2-\lambda}\frac{1}{n^2}+\frac{\lambda}{2-\lambda}\frac{1}{n}.\end{aligned}$$

For what concerns the probability of the joint events we have

$$\begin{aligned}\Pr(I_j^{\mathrm{CR}}=1,I_k^{\mathrm{CR}}=1)&=\Pr(\boldsymbol{\eta}_j>\boldsymbol{M}_{j-1},\boldsymbol{\eta}_k>\boldsymbol{M}_{k-1})\\&=\Pr\left(\boldsymbol{\eta}_j>\frac{\boldsymbol{\eta}_1}{j-1},\boldsymbol{\eta}_k>\max\left(\boldsymbol{\eta}_j,\frac{\boldsymbol{\eta}_2}{k-j-1}\right)\right)\\&=\int_{(-\infty,0]^d}\int_{(-\infty,\boldsymbol{y}]}\Pr\left(\boldsymbol{\eta}_1<(j-1)\boldsymbol{x},\boldsymbol{\eta}_2<(k-j-1)\boldsymbol{y}\right)\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{x})\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{y})\\&=\int_{(-\infty,0]^2}\int_{(-\infty,\boldsymbol{y}]}\mathrm{e}^{(j-1)\|\boldsymbol{x}\|_\lambda+(k-j-1)\|\boldsymbol{y}\|_\lambda}\,\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{x})\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{y})\\&=\int_{-\infty}^{0}\int_{-\infty}^{y_1}\mathrm{e}^{(k-j-1)\|\boldsymbol{y}\|_\lambda}\,I(\boldsymbol{y})\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{y})+\int_{-\infty}^{0}\int_{-\infty}^{y_2}\mathrm{e}^{(k-j-1)\|\boldsymbol{y}\|_\lambda}\,I(\boldsymbol{y})\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{y})\\&+\int_{-\infty}^{0}\mathrm{e}^{(k-j-1)\|\boldsymbol{y}\|_\lambda}\,I(\boldsymbol{y})\mathrm{d}\Pr_{\boldsymbol{\eta}(\boldsymbol{y})}=I_1+I_2+I_3,\end{aligned}$$

where

$$I(\boldsymbol{y})=\int_{(-\infty,\boldsymbol{y}]}\mathrm{e}^{(j-1)\|\boldsymbol{x}\|_\lambda}\,\mathrm{d}\Pr_{\boldsymbol{\eta}}(\boldsymbol{x}).$$

Computation of I_1. In the case $y_1<y_2$, we have

$$I(\boldsymbol{y}) = \int_{-\infty}^{y_1}\int_{x_1}^{y_2}(1-\lambda)\,\mathrm{e}^{j(x_1+(1-\lambda)x_2)}\,\mathrm{d}x_2\mathrm{d}x_1 + \int_{-\infty}^{y_1}\int_{-\infty}^{x_1}(1-\lambda)\,\mathrm{e}^{j(x_2+(1-\lambda)x_1)}\,\mathrm{d}x_2\mathrm{d}x_1$$
$$+ \int_{-\infty}^{y_1}\lambda\,\mathrm{e}^{j(2-\lambda)}\,\mathrm{d}x = I_{1,1}+I_{1,2}+I_{1,3}.$$

Integrals $I_{1,1}$, $I_{1,2}$ and $I_{1,3}$ are equal to

$$I_{1,1} = (1-\lambda)\int_{-\infty}^{y_1}\mathrm{e}^{jx_1}\,\frac{e^{j(1-\lambda)y_2}-\mathrm{e}^{j(1-\lambda)x_1}}{j(1-\lambda)}\mathrm{d}x_1 = \frac{1}{j}\left(\mathrm{e}^{j(1-\lambda)y_2}\,\frac{\mathrm{e}^{jy_1}}{j} - \frac{\mathrm{e}^{j(2-\lambda)y_1}}{j(2-\lambda)}\right),$$
$$I_{1,2} = (1-\lambda)\int_{-\infty}^{y_1}\mathrm{e}^{j(1-\lambda)x_1}\,\frac{\mathrm{e}^{j\lambda x_1}}{j}\mathrm{d}x_1 = \frac{1-\lambda}{j^2}\,\frac{\mathrm{e}^{j(2-\lambda)y_1}}{j(2-\lambda)}$$
$$I_{1,3} = \lambda\,\frac{\mathrm{e}^{j(2-\lambda)y_1}}{j(2-\lambda)}.$$

Therefore,

$$I_1 = \int_{-\infty}^{0}\int_{-\infty}^{y_1}\frac{1}{j^2}\,\mathrm{e}^{j(y_1+(1-\lambda)y_2)} - \frac{1-j}{j^2(2-\lambda)}\lambda\,\mathrm{e}^{j(2-\lambda)y_1}\,\mathrm{e}^{(k-j)(y_1+(1-\lambda)y_2}\,\mathrm{d}y_1\mathrm{d}y_2$$
$$= \frac{1-\lambda}{j^2}\int_{-\infty}^{0}\mathrm{e}^{k(1-\lambda)y_2}\,\frac{\mathrm{e}^{ky_2}}{k} - \frac{\lambda}{2-\lambda}\,\frac{1-j}{k+(1-\lambda)j}\,\mathrm{e}^{(2-\lambda)k}\,\mathrm{d}y_2$$
$$= \frac{1-\lambda}{(2-\lambda)^2}\,\frac{1}{j^2k^2}\,\frac{(2-\lambda)k+(2-\lambda)(1-\lambda)j-\lambda(1-j)k}{k+(1-\lambda)j}.$$

The computation for I_2 is similar to that of I_1 and we obtain $I_2 = I_1$.

Computation of I_3. In the case $y_1 = y_2$, we have

$$I(\boldsymbol{y}) = 2\int_{-\infty}^{y}\int_{-\infty}^{x_2}(1-\lambda)\,\mathrm{e}^{j(x_1+(1-\lambda)x_2)}\,\mathrm{d}x_1\mathrm{d}x_2 + \int_{-\infty}^{y}\lambda\,\mathrm{e}^{j(2-\lambda)}\,\mathrm{d}x$$
$$= 2\frac{1-\lambda}{j^2(2-\lambda)}\,\mathrm{e}^{j(2-\lambda)y} + \lambda\,\frac{\mathrm{e}^{j(2-\lambda)y}}{j(2-\lambda)} = 2\frac{1-(1-j)\lambda}{j^2(2-\lambda)}\,\mathrm{e}^{j(2-\lambda)y}.$$

Therefore,

$$I_3 = \int_{-\infty}^{0}\lambda\left(2\frac{1-\lambda}{j^2(2-\lambda)} + \frac{\lambda}{j(2-\lambda)}\right)\mathrm{e}^{k(2-\lambda)}\,\mathrm{d}x = \lambda\,\frac{2(1-\lambda)+\lambda j}{(2-\lambda)}\,\frac{1}{j^2k}.$$

Finally

$$\Pr(I_j^{\mathrm{CR}} = 1, I_k^{\mathrm{CR}} = 1) = 2I_1 + I_3.$$

It is still an open question how to show this result for a sequence $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_n$, without assuming a specific dependence form.

Note that from the previous formulas we can deduce $\Pr(I_n^{\mathrm{CR}} = 1\,i.o.)$:

- Case $\lambda = 1$. Since the I_n^{R}'s are independent and $\sum_{n=2}^{\infty}\frac{1}{n} = \infty$, then

$$\Pr(I_n^{\mathrm{CR}} = 1\,i.o.) = 1$$

by the second Borel-Cantelli Lemma;

- Case $\lambda = 0$. Since the I_n^{CR}'s are independent and $\sum_{n=2}^{\infty} \frac{1}{n^d} < \infty$ if $d > 1$, then

$$\Pr(I_n^{\mathrm{CR}} = 1\, i.o.) = 0$$

by the first Borel-Cantelli Lemma,

- Case $\lambda \in (0,1)$. Call $N = \sum_{n=2}^{\infty} I_n^{\mathrm{CR}}$, then

$$\mathrm{e}[N] = \sum_{n=2}^{\infty} \mathrm{e}[I_n^{\mathrm{CR}}] = \sum_{n=2}^{\infty} \Pr(I_n^{\mathrm{CR}} = 1) = \sum_{n=0}^{\infty} 2\frac{1-\lambda}{2-\lambda}\frac{1}{n^2} + \frac{\lambda}{2-\lambda}\frac{1}{n} = \infty.$$

These results are consistent with Theorem 5.3 in [6].

3 Conclusions

Despite the detailed study on records over independent sequences of random variables, the literature in the multivariate setting suffers of non extensive discussions. This is mainly due to two aspects: the first one concerns the various definitions of records which can be given over sequences of random vectors, the second one is the extreme complexity of the problem as soon as dependence between components of random vectors is taken into account. The Marshall-Olkin distribution is a handy tool to explore multivariate records where dependence is present, and represents the first step to consider to have an idea of which results are not true anymore in such cases. The main goal would be that of finding closed formulas for both the joint complete records occurrences and distribution without imposing defined dependence structures: this still represents an open question.

References

[1] Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja (1998). *Records*. Wiley Series in Probability and Statistics. New York: Wiley.

[2] Falk, M., A. Khorrami Chokami, and S. A. Padoan (2018a). Some results on joint record events. *Statistics and Probability Letters 135*(11–19).

[3] Falk, M., A. Khorrami Chokami, and S. A. Padoan (2018b). On multivariate records from random vectors with independent components. *Journal of Applied Probability 135*(11–19).

[4] Falk, M., A. Khorrami Chokami, and S. A. Padoan (2020). Records for time-dependent stationary Gaussian sequences. *Journal of Applied Probability 57*(78–96).

[5] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics* (2 ed.). Malabar: Krieger.

[6] Goldie, C. M. and S. I. Resnick (1989). Records in a partially ordered set. *Ann. Probab. 17*(2), 678–699.

[7] Goldie, C. M. and S. I. Resnick (1995). Many multivariate records. *Stochastic Process. Appl. 59*(2), 185–216.

[8] Zott, M. (2016). *Extreme Value Theory in Higher Dimensions. Max-Stable Processes and Multivariate Records*. Ph. D. thesis, University of Würzburg.

The Joint Censored Gaussian Graphical Lasso Model

Inferenza penalizzata del modello grafico Gaussiano congiunto

Gianluca Sottile, Luigi Augugliaro and Veronica Vinciotti

Abstract The Gaussian graphical model is one of the most used tools for inferring genetic networks. Nowadays, the data are often collected from different sources or under different biological conditions, resulting in heterogeneous datasets that exhibit a dependency structure that varies across groups. The complex structure of these data is typically recovered using regularized inferential procedures that use two penalties, one that encourages sparsity within each graph and the other that encourages common structures among the different groups. To this date, these approaches have not been developed for handling the case of censored data. However, these data are often generated by gene expression technologies such as RT-qPCR experiments. In this paper, we fill this gap and propose an extension of joint Gaussian graphical modelling to account for censored, or more generally missing, data.

Abstract *Il modello grafico Gaussiano è uno degli stimatori più utilizzati per fare inferenza sulle reti genetiche. Al giorno d'oggi, i dati raccolti sono spesso generati da diverse fonti o da diverse condizioni biologiche, risultando in dataset eterogenei, la cui struttura complessa viene analizzata utilizzando stimatori con due penalizzazioni per incoraggiare, da un lato, la sparsità all'interno di ciascun grafo e, dall'altro, le strutture comuni tra i grafi. Tuttavia, in diversi campi applicativi i limiti degli strumenti di rilevazione ne rendono teoricamente ingiustificato l'utilizzo, anche quando l'assunzione relativa alla distribuzione normale multivariata è soddisfatta. In questo articolo proponiamo un'estensione ai dati censurati.*

Key words: Gaussian Graphical Models, High-Dimensional Incomplete Data, Graphical Lasso, Heterogeneous Data

---

Gianluca Sottile
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: gianluca.sottile@unipa.it

Luigi Augugliaro
Department of Economics, Business and Statistics, University of Palermo, Italy, e-mail: luigi.augugliaro@unipa.it

Veronica Vinciotti
Department of Mathematics, University of Trento, Italy, e-mail: veronica.vinciotti@unitn.it

1 Introduction

Recently, sparse inference of Gaussian Graphical Models (GGMs) defined in a high-dimensional setting has received intense development. Given the assumption of independent and identically distributed data, a penalized estimator of a GGM is the maximizer of a specific objective function where the log-likelihood function is compensated by a sparsity inducing penalty function that controls the amount of shrinkage on the resulting estimators. The estimator proposed in Yuan and Lin (2007), called graphical lasso (glasso), uses a lasso-type penalty function and has been extensively used in applied research and well studied in the computational as well as theoretical literature (e.g. Friedman et al., 2008). The interested reader is referred to Augugliaro et al. (2016) for an extensive review.

Despite a widespread literature on the glasso estimator, the assumption of independent and identically distributed data has heavily reduced its application to modern datasets, which are often generated by more complex sampling schemes. A typical example of complex data structures is the case of data collected from different sources, such as gene expression data measured on multiple tissues. In this setting, sparse inference is typically carried out by using two specific penalty functions. These are chosen to encourage, on the one hand, sparsity within each graph and, on the other hand, the common structures across the graphs.

In this paper, we extend the estimator proposed in Danaher et al. (2014), called joint glasso (jglasso), to the setting studied in Augugliaro et al. (2020). That is, we consider the case where a part of the data is unobserved due to a known censoring mechanism. Gene expression measured by transcription quantitative polymerase chain reaction (RT-qPCR) is a well-known example of highly dimensional right-censored data.

The remaining part of this paper is structured as follows. Section 2 briefly reviews the inference of GGMs under censoring, while Section 3 proposes an extension of the jglasso estimator to the setting of multiple conditions. Computational aspects are addressed in Section 4 whereas in Section 5 we evaluate the performance of the proposed estimator by a simulation study. Finally, in Section 6 we draw some conclusions.

2 Background on censored Gaussian Graphical Models

Suppose that a p-dimensional random vector, say $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$, is distributed according to a multivariate Gaussian density function denoted by:

$$\phi(\boldsymbol{z}; \boldsymbol{\mu}, \Theta) = (2\pi)^{-\frac{p}{2}} |\Theta|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu})^\top \Theta (\boldsymbol{z} - \boldsymbol{\mu}) \right\},$$

where $E(\boldsymbol{Z}) = \boldsymbol{\mu}$ and $\Theta = (\theta_{ij})$ is the precision matrix, that is, the inverse of the covariance matrix. Now, let $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$ be a realization of $\boldsymbol{Z}$ and assume that a part of this vector is censored. Thus, $\boldsymbol{z}$ can be split into $\boldsymbol{z}^o$ and $\boldsymbol{z}^c$, i.e., the observed

and unobserved part of $\boldsymbol{z}$, respectively. Denoting by l_j and u_j the lower and upper censoring values of Z_j, respectively, it is easy to show that the probability density function can be written as follows (see Augugliaro et al. (2020) for more details):

$$\widetilde{\phi}(\boldsymbol{z}^o;\boldsymbol{\mu},\Theta) = \int_D^c \phi(\boldsymbol{z}^o,\boldsymbol{z}^c;\boldsymbol{\mu},\Theta)\mathrm{d}\boldsymbol{z}^c, \tag{1}$$

where the integral refers only to the variables whose index belongs to the set $c = \{j : z_j \text{ is censored}\}$, whereas the region of integration D is the Cartesian product of $|c|$ intervals, denoted by D_j, whose definition depends on whether z_j is left or right censored. Formally, $D_j = (-\infty, l_j)$ if $z_j \leq l_j$ otherwise we let $D_j = (u_j, +\infty)$.

Using (1), a censored GGM is the set $\{\widetilde{\phi}(\boldsymbol{z}^o;\boldsymbol{\mu},\Theta),\mathscr{G}\}$, where $\mathscr{G} = \{\mathscr{V},\mathscr{E}\}$ is the undirected graph encoding the conditional independences among the p random variables. More specifically, $\mathscr{V}$ is the set of nodes associated with the random variables, and $\mathscr{E}$ is the subset of the Cartesian product $\mathscr{V} \times \mathscr{V}$, such that $(i,j) \notin \mathscr{E}$ iff Z_i and Z_j are stochastically independent given all the remaining variables. As shown in Lauritzen (1996), the topological structure of the graph $\mathscr{G}$ is related to the entries of the precision matrix Θ, i.e., $(i,j) \in \mathscr{E}$ iff $\theta_{ij} \neq 0$, thus the problem of estimating $\mathscr{E}$ is equivalent to the problem of fitting a sparse precision matrix.

3 Sparse estimation of multiple GGMs with censored data

Suppose that n observations are collected from $K \geq 2$ different but related subpopulations, under a censoring mechanism and assuming a GGM for each subpopulation. To simplify our notation, all quantities related to the kth population are indexed by k. For instance, the kth censored GGM is denoted by $\{\widetilde{\phi}(\boldsymbol{z}_k^o;\boldsymbol{\mu}_k,\Theta_k),\mathscr{G}_k\}$. Under the assumption of independent sampling, the average observed log-likelihood function is

$$\bar{\ell}(\{\boldsymbol{\mu}\},\{\Theta\}) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\int_{D_{i,k}}^{c_{i,k}} \phi(\boldsymbol{z}_{i,k}^o,\boldsymbol{z}_{i,k}^c;\boldsymbol{\mu}_k,\Theta_k)\mathrm{d}\boldsymbol{z}_{i,k}^c, \tag{2}$$

where $\{\boldsymbol{\mu}\} = \{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}$ and $\{\Theta\} = \{\Theta_1,\ldots,\Theta_K\}$. Maximum likelihood estimators are found by maximizing equation (2). However, these estimators have a very high variance when the sample sizes n_k are larger but close to p. Thus, to overcome the inferential problems related with the high-dimensional setting, we propose to extend the jglasso estimator of Danaher et al. (2014), by replacing the log-likelihood function with the function (2). The resulting estimator, that we call censored jglasso, is formally defined as follows:

$$(\{\widehat{\boldsymbol{\mu}}\},\{\widehat{\Theta}\}) = \arg\max \bar{\ell}(\{\boldsymbol{\mu}\},\{\Theta\}) - \rho\left(\alpha\sum_{k=1}^{K}\|\Theta_k\| + (1-\alpha)P(\{\Theta\})\right), \tag{3}$$

where $\alpha \in [0,1]$ is a tuning parameter that controls the trade-off between two convex penalty functions, where the first one is a lasso-type penalty function that encourages sparsity within each estimated precision matrix, while the second one $P(\{\Theta\})$ is chosen to encourage some form of similarity across the K graphs. Following Danaher et al. (2014), two possible choices are:

$$P_F(\{\Theta\}) = \sum_{k<k'} |\theta_{ij,k} - \theta_{ij,k'}| \quad \text{or} \quad P_G(\{\Theta\}) = \sum_{i \neq j} \left(\sum_{k=1}^{K} \theta_{ij,k} \right)^{1/2}.$$

The censored fused glasso (cfglasso) estimator is defined by solving problem (3) with $P_F(\{\Theta\})$ as the second penalty function whereas, using $P_G(\{\Theta\})$, we get the censored group glasso (cgglasso) estimator. Like the standard glasso, both estimators result in sparse precision matrices. However while cfglasso encourages a stronger form of similarity between the precision matrices, enforcing some entries of $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_K$ to be identical, $P_G(\{\Theta\})$ encourages only a shared pattern of sparsity. Finally, the positive tuning parameter ρ controls the overall amount of shrinkage on the resulting estimators.

4 Computational Aspects

The estimator (3) can be efficiently solved by combining the penalized EM algorithm proposed in Augugliaro et al. (2020) with the alternating directions method of multipliers (ADMM) algorithm developed in Danaher et al. (2014).

The EM algorithm is based on the idea of repeating two steps until a convergence criterion is met. Since the multivariate Gaussian distribution belongs to the exponential family, the first step, called E-Step, simply requires the computation of the conditional expected values of the sufficient statistics. In our model, the E-Step requires the computation of the K imputed vectors of empirical means, denoted by $\bar{\boldsymbol{x}}_k$, along with the empirical covariance matrices $\boldsymbol{S}_k$ (see Augugliaro et al. (2020) for more details). In the second step of the EM algorithm, the M-Step, we first update the current estimates of the K expected values using $\bar{\boldsymbol{x}}_k$, and then the precision matrices are updated by solving the following maximization problem:

$$\max_{\{\Theta\}} \sum_{k=1}^{K} \frac{f_k}{2} \{\log|\Theta_k| - \operatorname{tr}(S_k\Theta_k)\} - \rho \left(\alpha \sum_{k=1}^{K} \|\Theta_k\| + (1-\alpha) P(\{\Theta\}) \right), \quad (4)$$

where $f_k = n_k/n$. The maximization problem (4) is equivalent to the problem studied in Danaher et al. (2014) and can be easily solved using the ADMM algorithm.

5 A Simulation Study

In this section, we compare our proposed estimator with jglasso (Danaher et al., 2014). For the latter, we use the R package `JGL` after imputing the censored data

with their limit of detection. The estimators are evaluated both in terms of recovery of the true graph (area under the precision-recall curve) and estimation of the true precision matrix (mean squared error). We set the right-censoring value to 40, the number of groups K to 3, and the sample size n_k to 100. Each right-censored response vector $\boldsymbol{z}_k$ is simulated according to sparse concentration matrices as reported in Fig. 1. The diagonal entries are fixed to 1 while the non-zero partial correlation coefficients $\theta_{h(h+j),k}$, with $h = 1, 6, 11, \ldots, p-4$ and $j = 1, \ldots, 4$, are sampled from a uniform distribution on the interval $[0.30, 0.50]$. In terms of graph theory what we have is a set of stars, i.e., complete bipartite graphs with only one internal vertex and four leaves. The values of the expected means are chosen in such a way that M

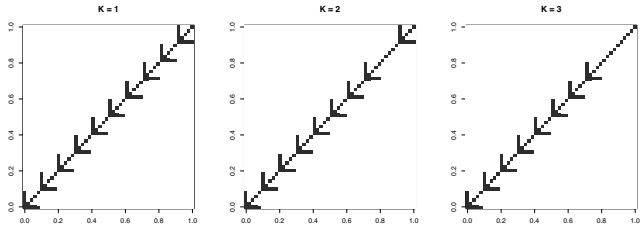


Fig. 1 True graph structure for the $K = 3$ concentration matrices

response variables are right-censored with probability equal to 0.40. The quantities p and M are used to specify the different scenarios used to analyze the behavior of the proposed estimators. In particular, we consider the following cases:

- **Scenario 1:** $p = 50$ and $M = 20$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the estimators when $n > p$.
- **Scenario 2:** $p = 200$ and $M = 80$. This setting is used to evaluate the impact of the high dimensionality on the estimators ($p > n$).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficient path of cglasso and jglasso, respectively, using the group lasso penalty in both cases. The path is computed using an equally spaced sequence of ρ-values keeping the α parameter equal to 0.50. For each Scenario and along the path of ρ values, the precision-recall curves and the area under these curves (AUC) are computed and then averaged between the K groups. The curves report the relationship between precision and recall for any ρ-value, which are defined as:

$$\text{Precision} = \frac{1}{K}\sum_k \left\{ \frac{\text{TP}}{\text{TP} + \text{FP}} \right\}_k, \quad \text{Recall} = \frac{1}{K}\sum_k \left\{ \frac{\text{TP}}{\text{TP} + \text{FN}} \right\}_k,$$

where TP, FP, and FN are given by the number of correctly selected edges, the number of wrongly selected edges and the number of wrongly selected missing edges, respectively. Table 1 shows how cgglasso gives a better estimate of the concentra-

tion matrices in terms of AUC and a lower mean squared error, for any given value of ρ. We report only five evenly spaced values of ρ.

Table 1 The first 5 columns refer to the average mean squared error of the concentration matrices $\boldsymbol{\Theta}$ for five evenly spaced values of ρ under the specification of the two Scenarios. The last column refers to the mean area under the curves across the sequence of ρ-values. Standard errors are reported in brackets.

	$\rho/\rho_{\max}$					AUC
	0.10	0.25	0.50	0.75	1.00	
Scenario 1						
cgglasso	4.65 (0.54)	8.06 (1.37)	12.10 (3.18)	12.92 (3.82)	12.81 (3.83)	0.97 (0.01)
jglasso	10.19 (1.46)	13.67 (0.94)	16.56 (0.93)	16.84 (1.51)	16.76 (1.57)	0.83 (0.02)
Scenario 2						
cgglasso	17.65 (2.37)	34.85 (4.71)	52.84 (10.72)	55.18 (12.38)	54.65 (12.26)	0.96 (0.01)
jglasso	45.32 (10.09)	62.55 (9.16)	75.53 (4.94)	75.50 (2.65)	75.32 (2.56)	0.81 (0.01)

6 Conclusion

In this paper, we have proposed an extension of the joint glasso estimator to multivariate censored data generated under multiple conditions. A simulation study showed that the proposed estimator performs better than the existing estimators both in terms of parameter estimation and of network recovery.

References

L. Augugliaro, A. M. Mineo, and E. C. Wit. ℓ_1-penalized methods in high-dimensional Gaussian Markov random fields. In M. Dehmer, Y. Shi, and F. Emmert-Streib, editors, *Computational Network Analysis with R: Applications in Biology, Medicine, and Chemistry*, chapter 8, pages 201–267. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2016.

L. Augugliaro, A. Abbruzzo, and V. Vinciotti. ℓ_1-penalized censored gaussian graphical model. *Biostatistics*, 21(2):e1–e16, 2020. doi: 10.1093/biostatistics/kxy043.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B*, 76(2):373–397, 2014.

J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Variable selection with unbiased estimation: the CDF penalty

Selezione delle variabili con stima non distorta: la penalizzazione CDF

Daniele Cuntrera, Vito M.R. Muggeo, Luigi Augugliaro

Abstract We propose a new SCAD-type penalty in general regression models. The new penalty can be considered a competitor of the LASSO, SCAD or MCP penalties, as it guarantees sparse variable selection, i.e., null regression coefficient estimates, while attenuating bias for the non-null estimates. In this work, the method is discussed, and some comparisons are presented.

Abstract *Proponiamo una nuova penalizzazione di tipo SCAD per modelli di regressione. La nuova penalizzazione può essere considerata un concorrente delle penalità LASSO, SCAD o MCP in quanto garantisce la selezione di variabili, cioè stime dei coefficienti di regressione esattamente nulle, mentre attenua la distorsione per le stime non nulle. In questo lavoro viene discussa la metodologia e vengono presentati alcuni confronti.*

Key words: Variable selection, L_1-type penalty, LASSO, SCAD, MCP

1 Introduction

Variable selection is an essential issue in regression modelling. Nowadays, data collection can involve several hundred or even thousands of variables. Therefore, spotting noise and significant covariates is a major concern when studying the effect

Daniele Cuntrera
Ph.D Student, Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche,
e-mail: daniele.cuntrera@unipa.it

Vito M.R. Muggeo
Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche,
e-mail: vito.muggeo@unipa.it

Luigi Augugliaro
Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche,
e-mail: luigi.augugliaro@unipa.it

on the response variable of interest. Out of the many existing variable selection techniques, including all-possible-regressions (Garside, 1965) or stepwise selection (Efroymson, 1996), L_1-penalized regression is one of the most widely used. In particular, the LASSO penalty (Tibshirani, 1996) is probably the most prominent and widespread approach in applications (Wu et al., 2009; Li et al., 2011; Lu et al., 2011). Unfortunately, in LASSO regression, the selected coefficients suffer from important bias. Therefore some alternatives have been proposed, notably: adaptive LASSO (Zou, 2006), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010), which enjoy the oracle property (unbiasedness, sparsity and continuity). There are also penalties such as bridge (Frank and Friedman, 1993) and elastic-net (Zou and Hastie, 2005) for purposes other than selecting variables, but we do not pursue them further in this paper.

The work is organised as follow: in section 2 we will talk about the method, then in section 3, we present our proposal. In section 4, a simulation study is presented. Then, the conclusions will follow.

2 Methods

In a classical linear regression context, the dependence of a response variable on one or more covariates is modeled as

$$y_i = x_i^T \beta + \varepsilon_i \qquad i = 1, ..., n,$$

where y_i is the response variable, x_i is the $1 \times p$ vector of explanatory variables, β are the coefficients of the model and ε_i are i.i.d. gaussian noise. The ordinary least squares (OLS) estimates are obtained by minimizing the sum of the model residuals, thus considering only the model's fidelity to the data. This way, we obtain a model that estimates a parameter for each covariate considered without considering that some parameters may be null. This ensures an unbiased estimate having however large variance.

The penalized regression models tend to mediate between the need to have a model as close as possible to the data and the need to select the variables useful in explaining the dependent variable, thus obtaining estimates with less variance. In the case of simple linear regression models, we can define the penalized objective loss function

$$-\mathscr{L}(\beta) + \sum_{j=1}^{p} p(\beta_j, \lambda) \tag{1}$$

$-\mathscr{L}(\beta)$ is the fidelity term, typically $\sum_i (y_i - x_i^T \beta)^2$, and $p(\beta, \lambda)$ is the penalty function that allows for selection of variables, and thus reducing the variance of the estimates. The tuning parameter λ determines the weight that the penalty has in the optimisation of the function: higher λ values result in the estimation of sparser but

potentially more biased models. Usually, the difference between various penalized regression models lies in the different specifications of the second term.

3 The proposal: the cumulative distribution function penalty

We propose to specify the penalty term in (1) via

$$p(\beta_j, \lambda) = \lambda\, p(|\beta_j|) = \lambda\ F(|\beta_j|; \theta) \qquad \beta_j \in \mathbb{R} \tag{2}$$

where $F(\cdot)$ is the cumulative distribution function of any specified random variable, such as Normal, Exponential or Chi-square. The penalty proposed in this work is derived from the best-known probability distribution, the Normal distribution. It is defined as

$$p(\beta_j, \lambda) = \lambda\, p(|\beta_j|) = \lambda\, \Phi_{0,s}(|\beta_j|).$$

It, therefore, consists of the cumulative distribution function (CDF) of β in absolute value. Note that it is considered a Normal distribution with mean 0 and standard deviation s: the central value equal to 0 ensures the singularity of the penalty at the origin (which makes the penalty capable of variable selection). In contrast, different values of s imply how "severe" the penalty is in bringing the parameters estimated by the model to a null value. The first significant difference that can be observed concerning SCAD and MCP penalties lies in the multiplication of the penalty for the tuning parameter λ: the definition of the penalty omits λ, and therefore it is not modified during the selection process of the tuning parameter.

As mentioned by Fan and Li (2001), a good penalty function must have three properties: unbiasedness (estimated non-zero parameters are near unbiased), sparsity (some estimates are equal to zero), and continuity (the estimator is continuous for the data, to avoid instability problems). Based on our evidence, the proposed penalty enjoys these three properties.

Suppose we have a data set (y_i, x_i), $i = 1, ..., n$, where $x_i = (x_{i1}, ..., x_{ip})^T$ are the covariates and y_i is the response. Operating in a simple linear regression framework, we assume independent observations, and that the set of covariates x_i is standardised. Denoted by $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, the estimates of the model with CDF penalty are obtained as

$$\hat{\beta} = arg\,min\left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \Phi_{0,s}(|\beta_j|) \right\}. \tag{3}$$

Problem (3) falls into a classic minimization problem of an objective function consisting of a fidelity component (the sum of the squared errors) plus a penalty component (precisely, the cumulative distribution function of the absolute values of the estimated parameters). A fundamental aspect concerns the choice of the s, which is used to weight the coefficients: there are more or less severe penalties for

different values of s. Based on our experience and simulations, a possible solution is to pre-estimate a model with s equal to 1; we take the higher value of λ that sets all the coefficients to 0 as the s used on the penalization. This ensures better stability on the path of the estimated coefficients.

4 Simulation study

A simulation study is conducted to compare the performance of the proposed penalty with alternatives already established in the literature (LASSO, SCAD and MCP). Models' performance is measured using mean square error (MSE) and false discovery rate (FDR). Both quantities are calculated using estimates corresponding to λ that minimize BIC. For SCAD and MCP, the constant is equal to 3.7.

To carry out the simulation, we generate the response variable as

$$y_i = x_i^T \beta_0 + \sigma_h \varepsilon_i \qquad h = 1,...,12, \qquad i = 1,...,n \tag{4}$$

where $n = 50$. Only the first 5 coefficients β_0 are non-null and equal to 1, then $\beta_0 = (1,1,1,1,1,0,...,0)^T$. The number of covariates is equal to 100. We generate regressor as $x_i \sim N(0,\Sigma)$, with the Toeplitz correlation matrix $\Sigma_{jk} = \frac{1}{2}^{|j-k|}$.

We also considered two different independent and identically distributed types of noise ε_i such that $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = 1$: the first one (a) $\varepsilon_i \sim N(0,1)$ and the second one (b) $\varepsilon_i \sim \frac{t_4}{2}^2$. The σ_h are twelve different equally-spaced values ranging in $[0.25, 3.0]$ of the random noise variance regulating the signal-to-noise ratio.

The results of the two experiments are shown in Figure 1.

In the first setting reported in Figure 1 we observe that, for all the values of σ tested, the value of MSE is always lower than SCAD and MCP, while better results than LASSO are observed only for high values of σ. Considering instead the fraction of coefficients erroneously different from 0, except for the smallest value of σ tested, our proposal is the one that has the lowest FDR compared to the other competitors.

Looking at the second simulation setting reported in Figure 1, the results do not change in a substance: our proposal has the lowest amount of MSE, just as it continues to be the method with the lowest FDR over the path of σ.

5 Conclusion

In this paper we have briefly introduced the problem of variable selection in regression models. We presented a novel proposal that we are working on, namely a penalty based on the cumulative distribution function of the Normal. The goal is to find a penalty that enjoys the stability of LASSO and some important propoprieties typical of SCAD and MCP (such as the oracle property). Results based on some

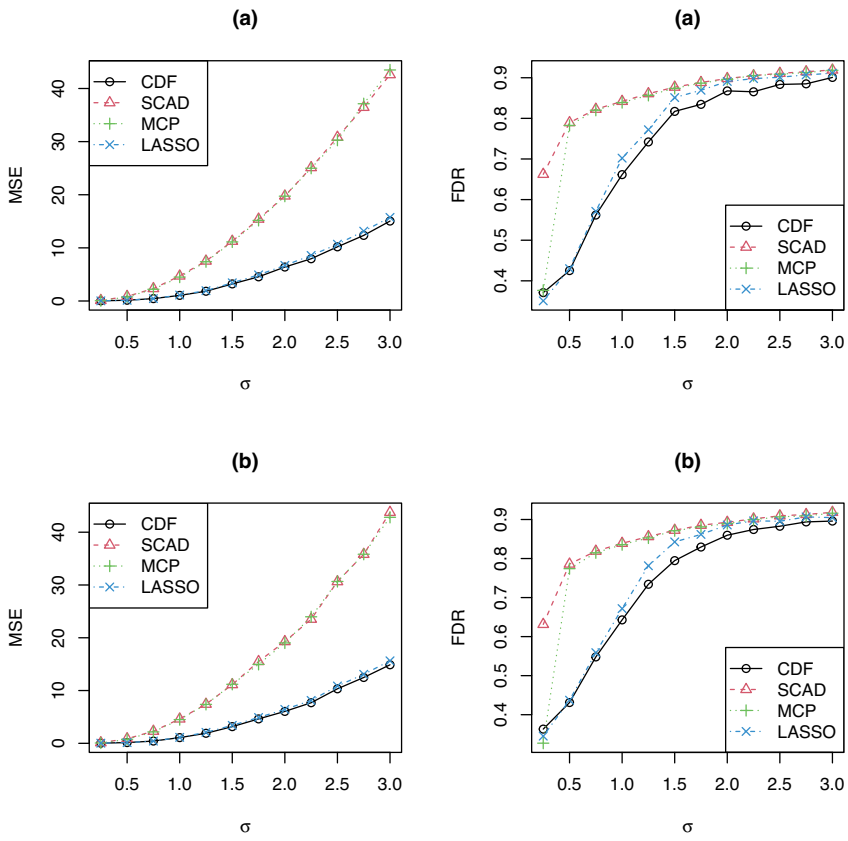


Fig. 1 MSE and average FDR for models estimated using CDF, SCAD, MCP and LASSO penalties, using settings (a) and (b) at different values of σ

simulation experiments showed that our proposal had good performance with respect to the traditional competitors: the CDF penalty attained the lowest value of MSE and FDR.

References

Efroymson, M. (1996). Stepwise regression-a backward and forward look. *Presented at the Eastern Regional Meetings of the Inst. of Math.*

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Garside, M. (1965). The best subset in multiple regression model. *Applied statistics*, 14:196–200.

Li, W., Feng, J., and Jiang, T. (2011). Isolasso: a lasso regression approach to rna-seq based transcriptome assembly. *Journal of Computational Biology*, 18(11):1693–1707.

Lu, Y., Zhou, Y., Qu, W., Deng, M., and Zhang, C. (2011). A lasso regression model for the construction of microrna-target regulatory networks. *Bioinformatics*, 27(17):2406–2413.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Automatic variable selection for MIDAS regressions: an application

Selezione automatica delle variabili per regressioni MIDAS: una applicazione

Consuelo Rubina Nava, Luigi Riso and Maria Grazia Zoia

Abstract Many micro and macroeconomic studies need to estimate and forecast time series using independent variables observed at higher frequencies than the dependent one. Mixed Data Sampling (MIDAS) regression models prove useful to model variables sampled at different frequencies. However, given that many time series are not very long, identification issues can emerge. To overcome this problem, we propose a two step procedure based on: i) the use of the so called "Best Path algorithm" for variable selection; ii) the identification and application of the best MIDAS regression specification. The test bed focuses on the half-yearly born of innovative start-ups in Italy, which are explained by using macroeconomic and financial indicators via a MIDAS regression. The results so obtained how advantages and powerfulness of the proposed approach.

Abstract *Molti studi micro e macroeconomici devono stimare e prevedere serie temporali utilizzando variabili indipendenti osservate a frequenze più elevate rispetto alla variabile dipendente. In questi casi è opportuno utilizzare i modelli di regressione Mixed Data Sampling (MIDAS) di recente sviluppo. Tuttavia, dato che molte serie temporali non sono molto lunghe, possono emergere problemi di identificazione. Al fine di superare questo problema, proponiamo una procedura in due fasi basata: i) sull'uso del cosiddetto "algoritmo Best Path" per la selezione delle variabili; ii) sull'identificazione e applicazione della migliore specifica di regressione MIDAS. Il banco di prova è basato su dati semstrali relativi alla nascita di start-up innovative in Italia, spiegati usando indicatori macroeconomici e finanziari*

Consuelo Rubina Nava
Università degli Studi di Torino, Lungo Dora Siena 100, Torino, e-mail: consuelorubina.nava@unito.it

Luigi Riso
Università Cattolica del Sacro Cuore di Milano, Largo Gemelli 1, Milano e-mail: luigi.riso@unicatt.it

Maria Grazia Zoia
Università Cattolica del Sacro Cuore di Milano, Largo Gemelli 1, Milano e-mail: maria.zoia@unicatt.it

in una regressione MIDAS. I risultati ottenuti mostrano i vantaggi e la potenza dell' approccio proposto.

Key words: MIDAS, U-MIDAS, automatic variable selection, Best Path algorithm, Start-ups

1 Introduction

The use of data sampled at different frequencies to explain and forecast variables that are emblematic in the many (micro and macro) economic sectors has increased in recent years. In fact, the need to estimate and forecast time series, taking advantage of independent variables observed (in may cases) at higher frequencies than the dependent one, emerges. This is one of the reason for which the recently developed Mixed Data Sampling (MIDAS) regression models has become extremely useful in such contexts [1, 2, 3].

However, given that in many empirical applications the available time series may be not very long and there may be a multitude of useful independent variables, identification issues can arise. To overcome this problem, together with the need to avoid weak multicollinearity problems, we propose a two step procedure. i) First an automatic variable selection algorithm, called the "Best Path algorithm" [6], is used to undertake a selection of the most relevant variables among the the ones suggested by the economic theory. Once a High-Dimensional Graph is built for the variables of the data set, the notion of mutual information between the latter is exploited to detect the path of the graph including the variables with the most explanatory power for the one of interest. ii) Then, the best MIDAS regression model, that considers different lags of the independent variables detected at the first step as well as different functional distributed lag polynomials, is built.

The powerfulness and the advantages of the proposed procedure is showed through an empirical application. The latter is based on the estimate of the born of innovative start-ups in Italy and considers a set of multiple macroeconomic and financial indicators as independent variables.

The paper is organized as follow. In Section 2 the methodology is proposed. Section 3 describes the dataset used for the empirical application and provides main results. Section 4 concludes the paper.

2 Methodology

In the following subsections, we describes in details the two steps characterizing the here proposed methodology.

2.1 Best Path Algorithm

We take advantage of a new algorithm [6] for an automatic variable selection procedure, based on high dimensional graphical models, which exploits mutual information. The algorithm first builds a high-dimensional graphical model by using the [5] algorithm and then detect the path of the graph that includes the variables with the most explanatory power for a node of interest by using the mutual information. As well known, mutual information is a measure of dependence or "closeness" between variables. The Best Path Algorithm proves useful in a many application as it allows to deal with mixed variables, either discrete or continuous. In case of linear models, the possibility of establishing a connection between the mutual information and the adjusted coefficients of determination ($\bar{R}^2$) makes the application of the algorithm very simple. Indeed, in this case, the last step of the algorithm turns out to be ultimately based on the analysis of the $\bar{R}^2$ of the models including the variables of the paths of graph. All the methodological details on the Best Path algorithm are reported in [6].

2.2 MIDAS regression

MIDAS regression models [1, 2, 3] can be usefully employed to avoid the ex-post aggregation of high-frequency data, that can discard potentially useful information in regression model devoted to the estimate and forecast of variables observed at lower frequency. MIDAS models allow either the prediction or the nowcast of the low-frequency variables as functions of higher-frequency ones. There are evidence that MIDAS regressions represent a more parsimonious alternative and are less sensitive to specification errors [2] with respect to other approaches such as State Space and mixed frequency VAR models.

In MIDAS regression, in fact, the frequency alignment is achieved by using a parametric functional constraint expressing the impact of the high-frequency variable on the lower-frequency one. More in details, let consider a dependent (low-frequency) variable y_t as function of r regressors, $x_\tau^{(i)}, i = 1,2,\ldots,r$, observed at higher-frequencies m_i, that is observed m_i times at time t, meaning $\tau = m_i t$ for all t, can be specified as follows

$$y_t = \alpha + \sum_{i=0}^{r} \sum_{k=0}^{K_i} \beta_k^{(i)} x_{tm_i-k}^{(i)} + \varepsilon_t, \;\; t = 1,2,\ldots,T \tag{1}$$

where $\sum_{k=0}^{K_i} \beta_k^{(i)} = 1$ for any i. To reduce the number $\sum_{i=0}^{r} K_i$ of the unknown parameters $\beta_0^0,\ldots,\beta_{K_0}^0,\ldots,\beta_0^r,\ldots,\beta_{K_r}^r$ a weighting scheme of the type

$$\beta_k^{(i)} = \frac{\psi(k,\boldsymbol{\theta}_i)}{\sum_{k=0}^{K_i} \psi(k,\boldsymbol{\theta}_i)} L^k \tag{2}$$

is assumed. In Eq. (2), the function $\psi(k, \boldsymbol{\theta}_i)$ depends on a low dimensional parameter vector $\boldsymbol{\theta}_i$, which allows a parsimonious parametrization. In general, the exponential Almon polynomial and the beta polynomial are used for this scope. When parameters in Eq. (1) are not constrained by a given function $\psi(k, \boldsymbol{\theta}_i)$, the resulting specification is called unrestricted MIDAS (U-MIDAS) [4]. The latter is particularly useful in macroeconomic applications when the sampling frequency gap between the variables included in the model is not so high, as in the empirical example undertaken in Section 3.

3 Data and Results

To show the powerfulness of the proposed approach, we have collected the half-yearly number of innovative start-ups from 2013 to 2021. Since 17 December 2012, the Law 221/2012 introduced in Italy this new class of firms defined as young companies, with a high technological contents and strong growth potentials. In particular, companies classified as innovative start-ups must match specific requirements to be inscribed into the special section of the the Italian companies register at the Chamber of Commerce. We have, thus, built a dataset combining the half-yearly number of innovative start-ups together with 28 macroeconomics and financial indicators/variables.

From this database we notice that, at the end of 2021, the number of innovative startups registered was equal to 14,083. They were principally firms providing services with a prevalence of software production and IT consulting and R&D. In fact, their Ateco 2007 classification principally refers to computer manufacturing, software production and R&D. This is confirmed also by Figure 1a which shows a text mining analysis, based on the description of the activity of each innovative start-up born from 2013 to 2021. From Section 2, both MIDAS and U-MIDAS regressions are suitable econometric models to estimate, nowcast and/or forecast the number of innovative start-ups. Following the MIDAS approach, the low-frequency variable consists of the number of innovative start-ups born every six months, while the macroeconomics and financial indicators are observed quarterly or monthly. Figure 1b shows the high dimensional model obtained from our dataset. The complete list of variables and the corresponding label node are summarized in Table 3. According to the results provided by the automatic variable selection algorithm, shown in Table 2, only two variables form the best path. We have, then, run all possible MIDAS (with the Almon specification) and U-MIDAS regressions that do not violate any identification issues. The resulting optimal model is a U-MIDAS regression specified as in Table 1. Finally, Figure 1c shows the forecasting performance of this optimal model given the two-steps procedure here developed.

(a)

(b)

(c)

Fig. 1: (a) Word cloud based on the registered description of the activity of each innovative start-up born from 2013 to 2021. (b) The minimal BIC forest for our dataset. In yellow the node representing the number of innovative start-ups. (c) Prediction performs of the optimal MIDAS model: in red the predicted values.

Parameters	**Estimate**	**S.E.**	**t-value**	**Signif.**
Intercept	$-9.23 \cdot 10^{3}$	$1.68 \cdot 10^{3}$	-5.49	***
Trend	$5.06 \cdot 10^{1}$	9.99	5.07	***
Seasonally adjusted export data	$4.14 \cdot 10^{-2}$	$8.43 \cdot 10^{-3}$	4.91	***
Seasonally adjusted retail sales retail	$7.76 \cdot 10^{1}$	$1.63 \cdot 10^{1}$	4.77	***

Table 1: Optimal MIDAS model (*** indicates a p-value$<$0.001)

Path Steps	$\bar{R}^2$	**MSE**
Path Step 1	0.886	46.1
Path Step 2	0.827	59.2
Path Step 3	0.821	106.3
Path Step 4	0.816	153.3
Path Step 5	0.832	114
Path Step 6	0.823	156.7
Path Step 7	0.823	146.1

Table 2: Output of Best Path Algorithm

Label nodes	Variable names	Freq.	Label nodes	Variable names	Freq.
1	Number of innovative start-ups	h-y	15	Seasonally adjusted industrial production data	m
2	Seasonally adjusted construction data	m	16	Seasonally adjusted retail sales data	m
3	Seasonally adjusted total industrial sales data	m	17	Seasonally adjusted retail sales retail Sales - non food data	m
4	Seasonally adjusted domestic industrial sales data	m	18	Seasonally adjusted retail sales retail Sales - food data	m
5	Seasonally adjusted foreign industrial sales data	m	19	Seasonally adjusted information and communication services data	q
6	Total Business Trust	m	20	Seasonally adjusted total amount of hours data	m
7	Total Manufacturing Trust	m	21	GDP series	m
8	Climate of trust	m	22	Milan Stock Exchange index	m
9	Climate of trust-Employees	m	23	Exchange Euro\Dollar	m
10	Climate of trust-Economics	m	24	Oil price index	m
11	Climate of trust-Construction	m	25	Intesa San Paolo index	m
12	Seasonally adjusted import data	m	26	Unicredit index	m
13	Seasonally adjusted export data	m	27	Eni index	m
14	Value added	m	28	Price of raw materials	m

Table 3: Name and frequency (m=monthly, q= quarterly, h-y=half-yearly) of the considered variables together with the related identification code of used in label nodes

4 Conclusions

The results obtained from the preliminary analysis, presented in Section 3, show that the proposed approach allows to: i) overcome potential identification issues; ii) identify the main relevant variables to be employed in MIDAS regression to estimate and forecast the variable of interest; iii) reduce the redundant information collected in the set of variables identified by the economics of innovation theory to explain the number of innovative start-ups registered, given the use of the "Best Path algorithm" which is grounded on mutual information.

References

1. Ghysels, E. Santa-Clara, P. Valkanov, R.: There is a risk-return trade-off after all. Journal of Financial Economics, **76** (3), 509–548, (2005)
2. Ghysels, E. Santa-Clara, P. Valkanov, R.: Predicting volatility: getting the most out of return data sampled at different frequencies. Journal of Econometrics, **131**, 59–95, (2006)
3. Ghysels, E., Sinko, A., Valkanov, R.: MIDAS regressions: Further results and new directions. Econometric reviews, **26** (1), 53–90, (2007)
4. Foroni, C, Marcellino, M., and Schumacher, Christian Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. Journal of the Royal Statistical Society. Series A (Statistics in Society), 57–82, (2015)
5. Edwards, D., De Abreu, G., Labouriau, R.: Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC bioinformatics, 11(1), 1–13 (2010)
6. Riso, L.: Use of High Dimensional Modeling for automatic variables selection: The Best Path algorithm. arXiv preprint, arXiv:2105.03173, (2021)

Distribution Theory and Estimation

A general framework for unit distributions

Un framework generale per le distribuzioni a supporto unitario

Francesca Condino, Filippo Domma and Bozidar V. Popovic

Abstract In this paper, we present a general framework for probability distributions having bounded support which is based on techniques of random variables transformations. We show that different existing models can be viewed as particular cases of this general formulation, by choosing a proper transformation. Moreover, we propose two type of unit Dagum distribution, belonging to the general family, obtained by considering two different transformation of Dagum random variable.

Abstract *In questo articolo, presentiamo un contesto generale per distribuzioni di probabilità a supporto limitato che si basa su tecniche di trasformazione di variabili casuali. Mostriamo che diversi modelli esistenti in letteratura possono essere visti come casi particolari di questa formulazione generale, scegliendo una opportuna trasformazione. Inoltre, proponiamo due diverse distribuzioni definite sul supporto unitario, appartenenti alla famiglia generale, costruite considerando due diverse trasformazioni della variabile casuale Dagum.*

Key words: transformations, bounded support, flexible shape

Francesca Condino
Department of Economics, Statistics and Finance '*Giovanni Anania*' - University of Calabria, Italy, e-mail: francesca.condino@unical.it

Filippo Domma
Department of Economics, Statistics and Finance '*Giovanni Anania*' - University of Calabria, Italy e-mail: filippo.domma@unical.it

Bozidar V. Popovic
Faculty of Science and Mathematics, Center for Applied Mathematics, University of Montenegro, Podgorica, Montenegro e-mail: bozidarp@ucg.ac.me

1 Introduction

In statistical literature, several proposal are done to develop new and more flexible distributions by using different techniques. Among them, various models are defined on bounded support, to address the need to interpret and model practical phenomena characterized by observations taking values in a limited range. In particular, the so-called unit distributions, having support in the unit interval, are suitable for the case of data derived as proportions, percentages or normalized values. Probably, the most used distribution for modelling continuous variables in the unit interval is the Beta distribution, but many other distributions are widely employed, such as Kumaraswamy distribution [7] and Topp-Leone distribution [11]. A substantial part of models with bounded support is obtained through the use of some transformation of random variables. Examples are the Unit-Burr III [10], Unit-Lindley [8], Unit-Gompertz [9], Unit-Burr XII [5], unit inverse Gaussian [3], Arcsecant-Hyperbolic Normal [6] and Logit-Slash [4], to name a few. In this work, we propose a general framework dealing with the approach based on transformations, that encloses most of the unit distributions already appeared in literature. Moreover, two new models will be proposed by transforming the well-known Dagum distribution.

2 General framework

The unit distributions recently proposed in literature can be described with a single probabilistic structure using simple techniques of transformations of random variables. To this end, suppose we have a random variable Y with probability density function (*pdf*) and distribution function (*df*), respectively, indicated by: $f_Y(y;\theta)$ and $F_Y(y;\theta)$, with $y \in I_Y \subset \Re$, $I_Y = [\underline{I}_Y, \bar{I}_Y]$. We also indicate with $y(p) = F_Y^{-1}(y;\theta)$, the *p-th* quantile, with $p \in (0,1)$, and with $h_Y(y;\theta)$ and $rh_Y(y;\theta)$ the hazard function and reversed hazard function, respectively. Let $k : I_Y \longmapsto J_W$, with $J_W = [\underline{J}_W, \bar{J}_W]$, be the application that defines the transformation of the random variable Y in the random variable W, which could also depends on a vector of parameters $\mathbf{a}$, i.e. $W := k(Y;\mathbf{a})$. To simplify the discussion, in this paper we assume that: (1) $\lim_{y\to\underline{I}_Y} k(y;\mathbf{a}) = \underline{J}_W > -\infty$ and $\lim_{y\to\bar{I}_Y} k(y;\mathbf{a}) = \bar{J}_W < \infty$, (2) $k(y;\mathbf{a}) \in C^1(J_W)$, where $C^1(.)$ is the class of differentiable functions whose first derivative is continuous. If $k'(y;\mathbf{a}) \neq 0$, $\forall y \in I_Y$ then $k'(.)$ is strictly positive (or negative) in I_Y and, therefore, $k(.)$ is a monotone function on I_Y. Consequently, $k(y;\mathbf{a})$ is invertible, its inverse $k^{-1}(.)$ is differentiable on J_W and

$$\left(k^{-1}(w;\mathbf{a})\right)' = \frac{1}{k'\left(k^{-1}(w;\mathbf{a})\right)}. \tag{1}$$

In this context, if $k(.)$ is an increasing monotone function, then the *df* of W is given by

$$F_W(w;\theta,\mathbf{a}) = P(W \leq w) = P\left(Y \leq k^{-1}(w;\mathbf{a})\right) = F_Y\left(k^{-1}(w;\mathbf{a});\theta\right) \tag{2}$$

Using (1), we can write the *pdf* of W as

$$f_W(w;\theta,\mathbf{a}) = \frac{\partial F_Y\left(k^{-1}(w;\mathbf{a});\theta\right)}{\partial k^{-1}(w;\mathbf{a})} \times \frac{\partial k^{-1}(w;\mathbf{a})}{\partial w} = \frac{f_Y\left(k^{-1}(w;\mathbf{a});\theta\right)}{k'\left(k^{-1}(w;\mathbf{a})\right)}. \quad (3)$$

It is immediate to verify that by (2) the *q-th* quantile, with $q \in (0,1)$ is $w(q;\theta,\mathbf{a}) = k\left(y(q;\theta);\mathbf{a}\right)$. and the hazard function (*hf*) and of the reversed hazard function (*rhf*) are given, respectively, by

$$h_W(w;\theta,\mathbf{a}) = \frac{f_W(w;\theta,\mathbf{a})}{1-F_W(w;\theta,\mathbf{a})} = \frac{h_Y(k^{-1}(w;\mathbf{a});\theta)}{k'\left(k^{-1}(w;\mathbf{a})\right)} \quad (4)$$

$$rh_W(w;\theta,\mathbf{a}) = \frac{f_W(w;\theta,\mathbf{a})}{F_W(w;\theta,\mathbf{a})} = \frac{rh_Y(k^{-1}(w;\mathbf{a});\theta)}{k'\left(k^{-1}(w;\mathbf{a})\right)}. \quad (5)$$

Evidently, if $k(.)$ is a decreasing monotone function, with small tricks, we can determine the quantities calculated above; in particular, the *df*, the *pdf*, the quantile of order q of W, respectively, are $F_W(w;\theta,\mathbf{a}) = 1 - F_Y\left(h^{-1}(w;\mathbf{a});\theta\right)$, $f_W(w;\theta,\mathbf{a}) = -\frac{f_Y\left(k^{-1}(w;\mathbf{a});\theta\right)}{k'\left(k^{-1}(w;\mathbf{a})\right)}$ and $w(q;\theta,\mathbf{a}) = k\left(y(1-q;\theta);\mathbf{a}\right)$. The *hf* and *rhf* are calculated accordingly. Most of the new proposal in literature can be thought as a particular case of the general framework above mentioned. For example, in the case of positive random variables, the most common transformations are $W = \frac{Y}{1+Y}$ and $W = e^{-Y}$. Through the first transformation, Unit-Burr III, Unit-Half-Normal, Unit-modified Burr III and Unit-Lindley distributions have been proposed, while with the second transformation Unit-Gompertz, Unit-Birnbaum-Saunders, Bounded Odd Inverse Pareto exponential, Unit-Burr XII, log-Bilal [1] and Unit Inverse Gaussian distributions have been obtained. In the case where Y takes values in the set of real numbers, the most used transformations is $W = \frac{1}{1+e^{-Y}}$, used to obtain Logit-Slash distribution. We point out that the non-monotone transformation $W = \frac{2}{e^{-Y}+e^{Y}}$, for $Y \in \Re$, was used to obtain the Arcsecant-Hyperbolic Normal distribution which, strictly speaking, does not belong to the general framework above described. In this paper, we restrict the use of arcsecant hyperbolic transformation to the case $Y > 0$, to have a monotone function. Finally, in general, to transform the variable Y, we can use any distribution function $G(\cdot)$ and obtain $W = G(Y)$.

3 Unit-Dagum distributions

The main objective of this section is to describe two transformations of the well-known Dagum model [2] that can potentially be more flexible unit-distributions than those present in literature. The Dagum random variable Y has, respectively, *df* and *pdf* $F_{Da}(y;\beta,\lambda,\delta) = \left(1+\lambda y^{-\delta}\right)^{-\beta}$, $f_{Da}(y;\beta,\lambda,\delta) = \beta\lambda\delta y^{-\delta-1}\left(1+\lambda y^{-\delta}\right)^{-\beta-1}$, with positive support $y > 0$, i.e. $I_Y = [0,\infty)$, and $\theta = (\beta,\lambda,\delta)$ with parameters $\beta > 0$,

$\lambda > 0$ and $\delta > 0$, (hereafter, in short $Da(\beta, \delta, \lambda)$). Parameter λ is a scale parameter while β and δ are shape parameters. The Dagum *pdf* has positive asymmetry and it is easy to verify that the mode is $y_m = \lambda^{\frac{1}{\delta}} \left(\frac{\beta\delta - 1}{\delta + 1}\right)^{\frac{1}{\delta}}$ for $\beta\delta > 1$. If $\beta\delta \leq 1$, the *pdf* is zero-modal. The *q-th* quantile of the Dagum distribution is $y(q) = F_{Da}^{-1}(q; \beta, \lambda, \delta) = \lambda^{\frac{1}{\delta}} (q^{-\frac{1}{\beta}} - 1)^{-\frac{1}{\delta}}$, thus, the median has an explicit expression. Moreover, the *r-th* moment is given by $\mu_{Da}^r = E(Y^r; \beta, \lambda, \delta) = \beta\lambda^{\frac{r}{\delta}} B\left(\beta + \frac{r}{\delta}, 1 - \frac{r}{\delta}\right)$ for $\delta > r$, where $B(.,.)$ is the complete Beta function.

3.1 Type I Unit-Dagum distribution

In the case of $Y > 0$, the hyperbolic secant transformation $W := k(Y) = \frac{2e^Y}{1+e^{2Y}}$ is a monotone function. Therefore, considering its inverse, it is simply to verify that the *df* of W is given by:

$$F_{I-UDa}(w; \beta, \lambda, \delta) = 1 - \left\{ 1 + \lambda \left[\log \frac{1 + \sqrt{1 - w^2}}{w} \right]^{-\delta} \right\}^{-\beta} \tag{6}$$

for $w \in (0, 1)$ and $\beta > 0$, $\lambda > 0$ and $\delta > 0$ (hereafter, in short $I - UDa(\beta, \delta, \lambda)$). Using (1), after simple algebra, we have $\left(k^{-1}(w; \mathbf{a})\right)' = \frac{-1}{w\sqrt{1-w^2}}$ and, therefore, the *pdf* of type I Unit-Dagum random variable is:

$$f_{I-UDa}(w; \beta, \lambda, \delta) = \frac{\beta\lambda\delta}{w\sqrt{1 - w^2}} \left[\log(w^*)\right]^{-\delta-1} \left\{ 1 + \lambda \left[\log(w^*)\right]^{-\delta} \right\}^{-\beta-1} \tag{7}$$

where $w^* = \frac{1+\sqrt{1-w^2}}{w}$. It is also possible to obtain the expression for *q-th* quantile, *r-th* moment, *hf* and *rhf* of the I-UDa (not shown, for space reasons). Figure 1 shows the *pdf* of I-UDa for two different set of parameters, while *hrf* of I-UDa for different values of β, λ and δ is shown in Figure 2.

3.2 Type II Unit-Dagum distribution

If we consider the transformation $W := k(Y) = e^{-Y}$ and its inverse, it is simply to verify that the *df* of W is given by

$$F_{II-UDa}(w; \beta, \lambda, \delta) = 1 - \left\{ 1 + \lambda \left[-\log(w)\right]^{-\delta} \right\}^{-\beta} \tag{8}$$

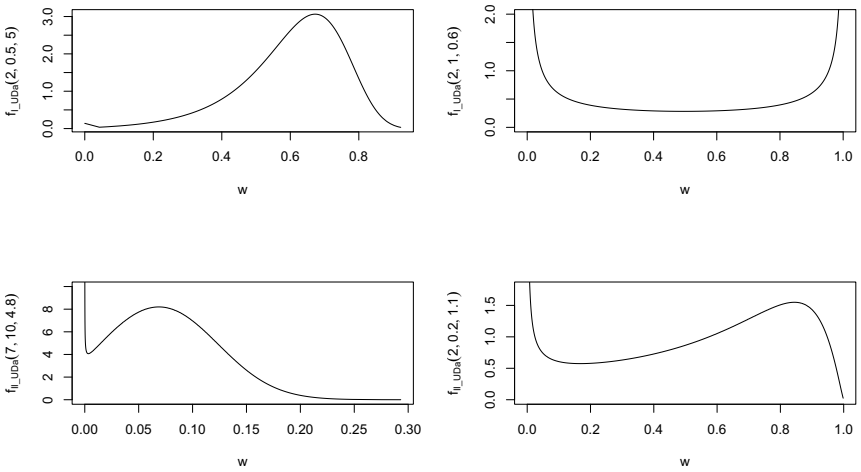


Fig. 1 Pdf of I-UDa (top panels) and II-UDa (bottom panels) for different values of parameters

for $w \in (0,1)$ and $\beta > 0$, $\lambda > 0$ and $\delta > 0$ (hereafter, in short II-UDa). Using (1), after simple algebra, we have $\left(k^{-1}(w;\mathbf{a})\right)' = -\frac{1}{w}$ and, therefore, the *pdf* of type II Unit-Dagum random variable is:

$$f_{II-UDa}(w;\beta,\lambda,\delta) = \frac{\beta\lambda\delta}{w}\left[-\log(w)\right]^{-\delta-1}\left\{1+\lambda\left[-\log(w)\right]^{-\delta}\right\}^{-\beta-1} \qquad (9)$$

Also in this case, the expression for *q-th* quantile, *r-th* moment, *hf* and *rhf* of the II-UDa can be obtained. Finally, Figure 1 shows two curves for the *pdf* of II-UDa for different values of β, λ and δ, while *hrf* is shown in Figure 2.

4 Conclusions and future work

This paper presents a general framework for distributions with bounded support and, consequently, general expressions for different aspects of models, such as quantiles, moments and hazard functions. A great number of distributions proposed in literature are enclosed in this framework. Moreover, it is shown how this approach can be considered to obtain new and very flexible distributions, through two different examples. For the future, others particular transformations could be considered in order to obtain more and more flexible models. Moreover, in the application contexts where specific indicators and/or characteristics of the distribution (such as location indicators, quantiles, variability measures, etc.) are of particular interest, we will try to construct regressive models on them in order to identify possible determinants of the phenomena under consideration.

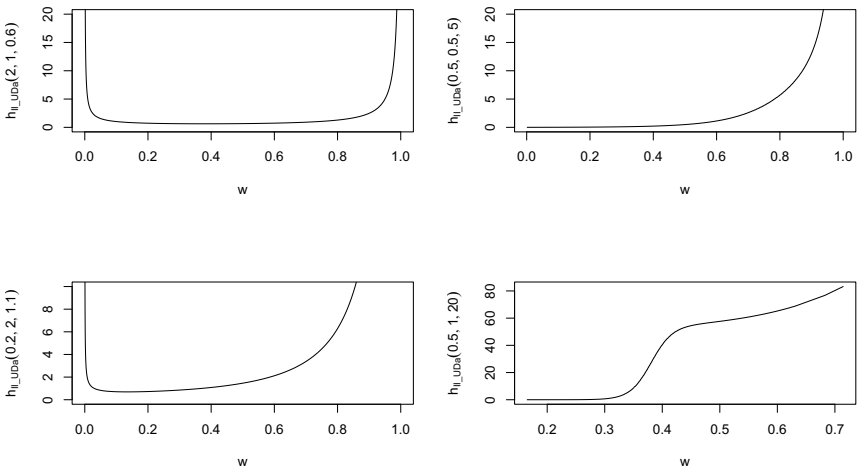


Fig. 2 Hazard rate of of I-UDa (top panels) and II-UDa (bottom panels) for different values of parameters

References

1. Altun, E., El-Morshedy, M., Eliwa, M.S.: A new regression model for bounded response variable: An alternative to the beta and unit-Lindley regression models. PLoS ONE. **16**, e0245627 (2021)
2. Dagum, C.: A new model of personal distribution: specification and estimation. Economie Appliquée. **30**, 413–437 (1977)
3. Ghitany, M., Mazucheli, J., Menezes, A.F.B., Alqallaf, F.: The unit-inverse Gaussian distribution: a new alternative to two-parameter distributions on the unit interval. Communications in Statistics - Theory and Methods. **48**, 3423–3438 (2018)
4. Korkmaz, M.: A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its applications. Journal of Applied Statistics. **47**, 2097–2119 (2019)
5. Korkmaz, M., Chesneau, C.: On the unit Burr-XII distribution with the quantile regresion modeling and applications. Computational and Applied Mathematics. **40**, 1–26 (2021)
6. Korkmaz, M., Chesneau, C., Korkmaz, Z.S.: On the arcsecant hyperbolic normal distribution. Properties, quantile regression modeling and applications. Symmetry. **13**, 2–24 (2021)
7. Kumaraswamy, P.: A generalized probability density function for double-bounded random processes. J. Hydrol. **46**, 79–88 (1980)
8. Mazucheli, J., Menezes, A.F.B., Chakraborty, S.: On the one parameter unit-Lindley distribution and its associated regression model for proportion data. Journal of Applied Statistics. **46**, 700–714 (2019)
9. Mazucheli, J., Menezes, A.F., Dey, S.: Unit-Gompertz distribution with applications. Statistica. **79**, 26–43 (2019)
10. Modi, K., Gill, V.: Unit Burr III distribution with application. Journal of Statistics and Management System. **23**, 579–592 (2019)
11. Topp, C.W., Leone, F.C.: A family of J-shaped frequency functions. J. Am. Stat. Assoc. **50**, 209–219 (1955)

Prediction intervals based on multiplicative model combinations

Intervalli di previsione basati su combinazioni moltiplicative di modelli

Valentina Mameli and Paolo Vidoni

Abstract The aim of this paper is to define prediction intervals based on multiplicative combination of elementary density functions as an useful surrogate of the true unknown predictive model for the interest random phenomenon. The specification of the weights associated to the individual density forecasts is performed by considering the continuous ranked probability score (CRPS) and its weighted extensions. A simple simulation study shows that, using a suitable weighted version of the CRPS, the estimated combined model provides prediction intervals having a coverage probability closed to the target nominal value.

Abstract *Lo scopo di questo lavoro è definire intervalli di previsione basati su combinazioni moltiplicative di modelli come un surrogato del vero modello predittivo per il fenomeno di interesse. La specificazione dei pesi associati ai singoli modelli della combinazione viene effettuata considerando lo score chiamato CRPS e le sue estensioni ponderate. Un semplice studio di simulazione evidenzia che, utilizzando un'opportuna versione ponderata del CRPS, il modello combinato stimato fornisce degli intervalli di previsione la cui probabilità di copertura è vicina a quella nominale.*

Key words: CRPS, coverage probability, estimative quantile, multiplicative model combination, prediction intervals, weigthed CRPS.

1 Introduction

Combining probability distributions is a relevant topic having potential implications for both inference and prediction. This issue, often called model or prediction pooling, is considered quite often in the econometric literature, with particular attention

Valentina Mameli · Paolo Vidoni
Department of Economics and Statistics, University of Udine, via Tomadini, 30/A I-33100 Udine Italy, e-mail: valentina.mameli@uniud.it, paolo.vidoni@uniud.it

to the specification of convenient predictive models (see, for example, [6]). The need of combined distributions may arise because it may not be possible to identify a reasonable model for the interest phenomenon or because the model could be so complex as not to allow the explicit calculation of the joint density function. Beyond that, several models (sometimes called expert opinions) may be available for a quantity of interest and, when these opinions are represented as probability distributions, a suitable (opinion) pooling operator is required.

Among the various pooling approaches proposed in the literature, we focus on multiplicative (logarithmic) model combinations and we face the crucial problem of assigning weights to the component distributions using the continuous ranked probability score (CRPS) as optimality criterion. In particular, we find that, using a suitable weighted version of the CRPS, the estimated combined model provides prediction intervals having a coverage probability closed to the target nominal value. In this respect, a simple simulation study, presented at the end of the paper, confirms this result even if a more in-depth theoretical analysis is required.

2 Multiplicative model combination for prediction

Let $(Y_1,\ldots,Y_n)$, $n \geq 1$, be a continuous random vector, following a suitable joint distribution, where $Y = (Y_1,\ldots,Y_{n-1})$ is observable, while $Z = Y_n$ denotes a future, or yet unobserved, random variable. This is a fairly general formulation but, for simplifying the exposition, we consider the simple case with independent, identically distributed random variables with marginal distribution and density functions given by $G(\cdot)$ and $g(\cdot)$, respectively. The aim here is to make effective prediction statements on the future random variable Z using the observed learning sample $y = (y_1,\ldots,y_{n-1})$ from Y and, in particular, to define prediction intervals or prediction limits having good coverage accuracy. More precisely, we look for α-prediction limits $c_\alpha(Y)$ such that

$$P_{Y,Z}\{Z \leq c_\alpha(Y)\} = E_Y[P_Z\{Z \leq c_\alpha(Y)\}] = E_Y[G\{c_\alpha(Y)\}] = \alpha, \tag{1}$$

for any fixed $\alpha \in (0,1)$. The above probability is called coverage probability and our objective is to define prediction limits, with nominal coverage α, which satisfies equation (1), exactly or approximately.

We consider the challenging situation where a parametric statistical model, including the true density function $g(\cdot)$, is not available and a set of $K \geq 2$ plausible parametric statistical models may be defined instead, giving alternative families of marginal density functions $f_k(\cdot;\theta_k)$ indexed by a d_k-dimensional unknown parameter $\theta_k \in \Theta_k \subseteq \mathbf{R}^{d_k}$, $k = 1,\ldots,K$. The aim is to use the information given by the observed data in order to define a combined model, having marginal density $f_c(\cdot)$, to be considered as a useful surrogate for the true density $g(\cdot)$. In particular, the combined model is expected to provide accurate prediction limits for Z, specified according to the coverage criterion defined above.

In this paper the focus is on multiplicative combinations of densities, called also logarithmic pools, so that

$$f_c(z;\theta,\omega) = \frac{\prod_{k=1}^{K} f_k(z;\theta_k)^{\omega_k}}{c(\omega)}, \quad z \in \mathbf{R}, \tag{2}$$

with $\theta = (\theta_1,\ldots,\theta_K)$, $c(\omega) = \int_{\mathbf{R}} \prod_{k=1}^{K} f_k(z;\theta_k)^{\omega_k} dz$ the normalizing factor and $\omega = (\omega_1,\ldots,\omega_K)$ a vector of non-negative weights, such that $\sum_{k=1}^{K} \omega_k = 1$. Compared to the linear pool, which defines linear mixture densities, the logarithmic one gives densities which are typically uni-modal and less dispersed. Moreover, this combination method is invariant under rescaling and it verifies the property of external Bayesianity [1, 3].

The combined density function (2) can be viewed as a surrogate for the true, unknown density function for Z, provided that a convenient choice for the weights $\widehat{\omega} = (\widehat{\omega}_1,\ldots,\widehat{\omega}_K)$ is made and suitable estimates $\widehat{\theta}$ for the model parameters θ, based on the available data y, are considered. In this case, the associated estimative prediction limit

$$\widehat{z}_\alpha^{\,c} = z_\alpha^c(\widehat{\theta},\widehat{\omega}) = F_c^{-1}(\alpha;\widehat{\theta},\widehat{\omega}) \tag{3}$$

can be used for prediction purposes, instead of the unknown true predictive quantile $z_\alpha = G^{-1}(\alpha)$; here, $F_c(\cdot)$ is the distribution function associated to the multiplicative density $f_c(\cdot)$.

Since the prediction limit (3) is based on a surrogate for the true model and, in addition, a plug-in procedure is considered to deal with the unknown parameters, the coverage probability is expected to differ from the target value α and, in many applications, the coverage error has been found to be substantial. This error can be viewed as the combination of the misspecification error, due to the use of $F_c(\cdot)$ instead of $G(\cdot)$, and an error term which depends on the plug-in approach, since an additional uncertainty is introduced when the unknown parameters ω and θ are substituted by the corresponding estimates $\widehat{\omega}$ and $\widehat{\theta}$.

3 Choosing the weights using a CRPS-based criterion

An effective approach for assessing the goodness of the combined models, and then useful for estimating the weights ω and the model parameters θ, is based on the CRPS. Following [4], the CRPS for the predictive distribution $F_c(\cdot)$, at a specific realization x, is defined as

$$\mathrm{CRPS}(F_c,x) = \int_{\mathbf{R}} \{F_c(u;\theta,\omega) - \mathbf{1}(x \le u)\}^2 du,$$

where $\mathbf{1}(\cdot)$ is the indicator function, which takes value 1 if the argument is true and 0 otherwise. If the quantile function $F_c^{-1}(\cdot;\theta,\omega)$ is uniquely defined, the CRPS can be expressed in the following equivalent form

$$\mathrm{CRPS}(F_c,x) = \int_0^1 2\left[\mathbf{1}\{x \le F_c^{-1}(\beta;\theta,\omega)\} - \beta\right]\{F_c^{-1}(\beta;\theta,\omega) - x\}d\beta.$$

The choice between the two specifications depends on the computational advantage obtained by specifying the predicted distribution as a distribution function or as a quantile function. In what follows, we consider the quantile version, generalized by considering the integrand multiplied by a suitable weight function $\pi(\cdot)$. As emphasized in [5], an appropriate choice for this function may emphasize the center or the tails of the interest variable's range, and this can be useful for specifying good predictions intervals. In particular, we consider

$$\mathrm{wCRPS}(F_c,x) = \int_0^1 2\left[\mathbf{1}\{x \le F_c^{-1}(\beta;\theta,\omega)\} - \beta\right]\{F_c^{-1}(\beta;\theta,\omega) - x\}\pi(\beta)d\beta$$

and, with $\pi(\beta) = \mathbf{1}(\beta \ge r)$ and $\pi(\beta) = \mathbf{1}(\beta \le l)$, where $l, r \in (0,1)$, we focus on the right or on the left tail of the distribution, respectively.

The (weighted) CRPS is a proper scoring rule and it can be used as a criterion for estimating both the model parameters and the system of weights required for defining the logarithmic pool (2). More precisely, we seek suitable values for θ and ω so that the expected score $E_X\{\mathrm{wCRPS}(F_c,X\}$ is minimized, where the random variable X is an independent copy of Y_i, $i = 1,\ldots,n$. Given the observed sample y, the estimates can be derived using, as objective function, the empirical counterpart of the expected wCRPS, that is

$$(\widehat{\theta},\widehat{\omega}) = \arg\min_{(\theta,\omega)} \frac{1}{n-1}\sum_{i=1}^{n-1} \mathrm{wCRPS}(F_c,y_i). \tag{4}$$

Since an explicit solution for this inferential problem is usually not available, the estimates can be obtained using iterative methods. One possibility is to set an initial value $\widehat{\omega}^{(0)}$ for the weights, e.g. $\widehat{\omega}^{(0)} = (1/K,\ldots,1/K)$, and to define $\widehat{\theta}^{(0)}$ as the solution of (4) with respect to θ, with $\omega = \widehat{\omega}^{(0)}$. Then the weights are updated by considering $\widehat{\omega}^{(1)}$, defined as the solution of (4) with respect to ω, with $\theta = \widehat{\theta}^{(0)}$, and the procedure continues sequentially until a suitable stopping criterion is satisfied. If the integral required for computing the wCRPS does not have an explicit solution, a suitable discrete approximation can be employed.

The estimative prediction limit $\widehat{z}_\alpha^c$ defined by (3), with $\widehat{\theta}$ and $\widehat{\omega}$ the wCRPS estimators given by (4), is supposed to have a coverage probability closed to the nominal value on the left or on the right tail of the distribution, according to the corresponding weight function considered in the wCRPS. We find out that, as emphasized in the simple application presented in the following section, choosing the weights and the model parameters using a criterion based on a suitable wCRPS may reduce the coverage error due to misspecification. In particular, when the uncertainty introduced by the plug-in procedure is not so pronounced, the coverage accuracy is closed to that one given by the improved prediction limit obtained using a suitable calibration procedure [2].

4 A simple simulation study

In this section, we perform a simulation study where the true data generating process is a Student t distribution with $\nu = 3$ degrees of freedom, in symbols $t(3)$. We consider the multiplicative combination of two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where $\mu_1, \mu_2 \in \mathbf{R}$, $\sigma_1, \sigma_2 > 0$, are unknown parameters. Then the combined density given in (2) is again normal $N(\mu_c, \sigma_c)$, where $\mu_c = \sigma_c^2 \sum_{k=1}^2 \omega_k \mu_k / \sigma_k^2$ and $\sigma_c = \{\sum_{k=1}^2 \omega_k / \sigma_k^2\}^{-1/2}$. The estimative α-quantile is $\widehat{z}_\alpha^c = \Phi^{-1}(\alpha, \widehat{\mu}_c, \widehat{\sigma}_c)$, where $\Phi^{-1}(\cdot, \mu, \sigma)$ is the quantile function of a normal distribution and $\widehat{\mu}_c, \widehat{\sigma}_c$ are the plug-in estimates of the mean and of the standard deviation of the logarithmic pool, obtained using the estimates for $\omega = (\omega_1, \omega_2)$ and $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ given by (4).

In the following, we consider the estimates based on the unweighted CRPS and on the weighted versions focusing on the left and on the right tails defined in the previous section, and indicated as wCRPS$_l$ and wCRPS$_r$, when $\pi(\beta) = \mathbf{1}(\beta \leq l)$ and $\pi(\beta) = \mathbf{1}(\beta \geq r)$, respectively. In the special case of the normal distribution, both the CRPS and the wCRPS can be easily derived in a closed form (see [7, 4]). Furthermore, the improved quantiles $\widetilde{z}_\alpha^c$ are obtained by adjusting, with the calibration procedure proposed in [2], the corresponding estimative quantiles $\widehat{z}_\alpha^c$ in order to achieve the nominal coverage level α, and for this reason they can be considered as benchmark values.

The simulation study compares the coverage probability of the estimative quantiles $\widehat{z}_\alpha^c$, based on the CRPS, wCRPS$_l$ and wCRPS$_r$, and of the associated improved quantiles $\widetilde{z}_\alpha^c$, for $\alpha = 0.05$ and $\alpha = 0.95$. Table 1 shows the results of a Monte Carlo simulation based on $N = 1{,}000$ replications. Estimated standard errors are smaller than 0.009. Coverage probabilities are obtained by using a bootstrap procedure based on $B = 500$ replications. The sample sizes are $n = 20, 50, 100$ and we set $l = r = 0.05$, for $\alpha = 0.05$, and $l = r = 0.95$, for $\alpha = 0.95$.

As it can be noted in Table 1, the estimative quantiles $\widehat{z}_\alpha^c$ have coverage probability equal or close to the target nominal value, depending on the sample size, whenever we focus on the right tail of the distribution with $r = 0.95$ for $\alpha = 0.95$ or on the left tail with $l = 0.05$ for $\alpha = 0.05$. As the sample size increases the gap between the estimated coverage probability and the target nominal level fades out, as expected. Moreover, the improved predictive quantiles $\widetilde{z}_\alpha^c$ show an accurate coverage probability, since it is almost equals to the nominal value α regardless of the sample size and of the CRPS criterion used. Finally, we can emphasize that estimative predictive intervals, obtained from the multiplicative model, exhibit a good predictive accuracy whenever a suitable choice of the weight function for the CRPS criterion is adopted, and then it is not necessary to improve the predictive performance using an additional, complex calibration procedure. Future work will consider the use of different model pools for prediction and also the possible extension to the time series framework.

Table 1 Coverage probabilities for estimative quantiles $\widehat{z}_\alpha^c$ and improved quantiles $\widetilde{z}_\alpha^c$ with level $\alpha = 0.05, 0.95$. Estimation is based on $1,000$ Monte Carlo replications and the estimated standard errors are smaller than 0.009. The bootstrap procedure is based on 500 samples. The true model is a $t(3)$ distribution, the sample sizes are $n = 20, 50, 100$ and $l = r = 0.05$, for $\alpha = 0.05$, and $l = r = 0.95$, for $\alpha = 0.95$.

		CRPS	wCRPS$_l$	wCRPS$_r$
$n = 20$	$\alpha = 0.05$			
	$\widehat{z}_\alpha^c$	0.083	**0.064**	0.085
	$\widetilde{z}_\alpha^c$	0.056	0.055	0.056
$n = 50$	$\alpha = 0.05$			
	$\widehat{z}_\alpha^c$	0.065	**0.050**	0.069
	$\widetilde{z}_\alpha^c$	0.047	0.042	0.047
$n = 100$	$\alpha = 0.05$			
	$\widehat{z}_\alpha^c$	0.072	**0.048**	0.079
	$\widetilde{z}_\alpha^c$	0.051	0.044	0.050
$n = 20$	$\alpha = 0.95$			
	$\widehat{z}_\alpha^c$	0.920	0.920	**0.946**
	$\widetilde{z}_\alpha^c$	0.946	0.947	0.948
$n = 50$	$\alpha = 0.95$			
	$\widehat{z}_\alpha^c$	0.926	0.924	**0.945**
	$\widetilde{z}_\alpha^c$	0.947	0.948	0.953
$n = 100$	$\alpha = 0.95$			
	$\widehat{z}_\alpha^c$	0.927	0.924	**0.938**
	$\widetilde{z}_\alpha^c$	0.937	0.937	0.937

References

1. Allard, D., Comunian, A., Renard, P.: Probability aggregation methods in geoscience. Math. Geosci. **44**, 545–581 (2012).
2. Fonseca, G., Giummolè, F., Vidoni, P.: Calibrating predictive distributions. J. Stat. Comput. Simul. **84**, 373–383 (2014).
3. Genest, C.: A characterization theorem for externally Bayesian groups. Ann. Statist. **12**, 1100–1105 (1984).
4. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. **102**, 359–378 (2007).
5. Gneiting, T., Ranjan, R.: Comparing density forecasts using threshold- and quantile-weighted scoring rules. J. Bus. Econom. Statist. **29**, 411–422 (2011).
6. Kascha, C., Ravazzolo, F.: Combining inflation density forecasts. J. Forecast. **29**, 231–250 (2010).
7. Taillardat, M., Fougères, A.-L., Naveau, P., de Fondeville, R.: Extreme events evaluation using CRPS distributions. https://arxiv.org/abs/1905.04022 (2019).

Some advances on pairwise likelihood estimation in ordinal data latent variable models

Alcuni avanzamenti per la stima basata sulla verosimiglianza a coppie per modelli a variabili latenti per dati ordinali

Giuseppe Alfonzetti and Ruggero Bellio

Abstract This note compares three estimation methods suitable for GLLVMs with ordinal response data, given by maximum likelihood estimation, pairwise likelihood estimation and a variation of the latter, obtained by a suitable scaling of the estimating equations by means of optimal weighting. The three methods are compared in a simulation study with six items and two correlated latent factors, for different sample sizes. The results suggest that the pairwise likelihood approach entails a limited loss of efficiency with respect to maximum likelihood estimation, that is not significantly reduced by the application of optimal weights.

Abstract *Questa nota confronta tre diversi metodi di stima per un modello GLLVM con risposta ordinale, ovvero la stima di massima verosimiglianza, la stima ottenuta massimizzando la verosimiglianza a coppie, e infine una variazione di quest'ultima ottenuta mediante l'applicazione di opportuni pesi ottimali all'equazione di stima. I tre metodi sono confrontati in uno studio di simulazione per un modello con sei item, due fattori latenti correlati tra loro e varie dimensioni campionarie. I risultati suggeriscono una limitata perdita di efficienza della stima basata sulla verosimiglianza a coppie rispetto alla stima di massima verosimiglianza, con nessun apprezzabile miglioramento dato dall'applicazione dei pesi ottimali.*

Key words: Pairwise likelihood, Optimal weights, Latent variable models

1 Introduction

Whenever the interest of a researcher lies on analyzing some non-observable constructs given a set of observed variables, latent variable models may come to be of

Giuseppe Alfonzetti
University of Padova, e-mail: giuseppe.alfonzetti@phd.unipd.it

Ruggero Bellio
University of Udine, e-mail: ruggero.bellio@uniud.it

help. It is not by chance their popularity in areas like psychometrics and social sciences, where latent traits are often the focus of research. A particular class of models has emerged to extend generalized linear models to incorporate latent variables, given by Generalized Linear Latent Variable Models (GLLVM); see [1] for an extended overview of the models and classical estimation methods for them. GLLVM can be seen as a generalization of normal factor analysis models where manifest data can be any member of the exponential family, allowing at the same time for a non-orthogonal latent space. Such flexibility allows to incorporate in GLLVM many classical models developed in psychometrics under the Item Response Theory (IRT) framework, where interest usually lies in discrete responses and parameters are estimated by the Maximum Likelihood (ML) method.

Here we focus on ordinal data, naturally suitable to model responses like ratings or preferences. An equivalent parameterization of GLLVM with ordinal data is given under the Underlying Response Variable approach (URV), where ordinal data are assumed to be a partial observation of a non-observed normal factor model. Under this framework, a pairwise likelihood (PL) approach has been proposed in [6]. This contribution aims at quantifying the efficiency loss of PL estimates with respect to ML ones by means of a simulation study, which also includes the refinement of PL estimation by means of optimal weighting system, as first proposed by [7]. Some brief remarks conclude the note.

2 Model and estimation methods

Let the observed dataset $y = (y_1^\top, \dots, y_n^\top)$ be the realization of a random matrix $Y \in R^n \times R^p$, with n the number of units and p the number of items. Likewise, the latent variables matrix $u = (u_1^\top, \dots, u_n^\top)$ is the realization of the random matrix $U \in R^n \times R^q$, where $q < p$ is the dimension of the latent space. At the i-th unit level, it follows that $y_i = (y_{i1}, \dots, y_{ip})^\top$ and $u_i = (u_{i1}, \dots, u_{iq})^\top$. Since we are dealing with ordinal data, for each item j, with $j = 1, \dots, p$, there are c_j possible increasing categories. It follows that $y_{ij} \in \{s_0, \dots, s_{c_j - 1}\}$. Furthermore we assume $u_i \overset{iid}{\sim} N_q(0, \Sigma_u)$ for $i = 1, \dots, n$, and Σ_u is constrained to be a correlation matrix. In the following, three different estimation methods are introduced.

2.1 Maximum likelihood estimation

ML estimation is traditionally performed under the IRT model parameterization. For the generic j-th item, $c_j - 1$ intercepts need to be estimated, such that the model intercept vector is defined with $\alpha = ({\alpha^{(1)}}^\top, \dots, {\alpha^{(p)}}^\top)^\top$ where the generic $\alpha^{(j)} = (\alpha_0^{(j)}, \dots, \alpha_{c_j - 2}^{(j)})^\top$ is the $(c_j - 1)$-dimensional intercept vector related to item j. The matrix of coefficients related to each item is defined with $B = (\beta_1^\top, \dots, \beta_p^\top)$, where

$\beta_j = (\beta_{j1}, \ldots, \beta_{jq})^\top$ for $j = 1, \ldots, p$. Let $\psi \in R^d$ be the vector collecting all the free parameters in B, α and Σ_u. The IRT parameterization assumes the items to be locally independent given the latent space, such that the likelihood contribution for the i-th unit is

$$L(B, \alpha; y_i|u_i) = \prod_{j=1}^{p} p(Y_{ij} = s_j|U_i = u_i; \beta_j, \alpha_j)\,. \tag{1}$$

Furthermore it is assumed that the conditional response probability is computed as

$$p(Y_{ij} = s_j|u_i; \beta_j, \alpha_j) = p(Y_{ij} \leq s_j|u_i; \beta_j, \alpha_j) - p(Y_{ij} \leq s_j - 1|u_i; \beta_j, \alpha_j)\,, \tag{2}$$

where the cumulative probability of a response is related to the linear predictor through the logit transformation, namely $\text{logit}\,(p(Y_{ij} \leq s_j|u_i; \beta_j, \alpha_j)) = \alpha_{s_j}^{(j)} - \beta_j^\top u_i$. For the computation of the marginal likelihood for the observed response, the latent variables have to be integrated out, so that the log-likelihood function is

$$\ell(\psi; y) = \sum_{i=1}^{n} \log \int_{R^q} p(u_i; \Sigma_u) \prod_{j=1}^{p} p(Y_{ij} = s_j|U_i = u_i; \alpha_j, \beta_j) du_i\,. \tag{3}$$

Since the q-dimensional integral has no closed-form solution, several estimation methods have been proposed in literature. The most popular approaches rely on the Expectation-Maximization algorithm, as reviewed in [1].

2.2 Pairwise likelihood estimation

Pairwise Likelihood estimation (PL) is a special case of composite likelihood [7], and it has been largely studied in literature to replace intractable likelihoods with the product of bivariate likelihood contributions. The method provides consistent estimates, with only a possible loss in efficiency; see [9] for an overview.

In the context of ordinal factor models, PL has been introduced under the Underlying Response Variable (URV) parameterization in [6]. In such setting, a different d-dimensional set of parameters is estimated, but it possible to re-express the estimates in the IRT parameterization. The URV parameters include the $c_j - 1$ thresholds for each item j, namely we define $\tau = ({\tau^{(1)}}^\top, \ldots, {\tau^{(p)}}^\top)^\top$ with $\tau^{(j)} = (\tau_0^{(j)}, \ldots, \tau_{c_j-2}^{(j)})^\top$, and a loading matrix $\Lambda = (\lambda_1^\top, \ldots, \lambda_p^\top)$ with $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jq})^\top$. Let $\theta \in \mathbb{R}^d$ be the vector collecting all the free parameters in Λ, τ and Σ_u. The data are assumed to be a partial observation of a normally distributed underlying continuous response y^*, such that $Y_j = s_j \iff \tau_{s_j-1}^{(j)} < y_j^* < \tau_{s_j}^{(j)}$, where $y^* = \Lambda u + \delta$ and $\delta \sim N_p(0, \Sigma_\delta)$ with $\Sigma_\delta = I_p - \text{diag}(\Lambda \Sigma_u \Lambda^T)$. In this case, the marginal likelihood is obtained by integration over the underlying response space, hence the integral is p-dimensional rather than q-dimensional, namely

$$\ell(\theta;y) = \sum_{i=1}^{n} \log \int_{\tau_{s_1-1}^{(1)}}^{\tau_{s_1}^{(1)}} \cdots \int_{\tau_{s_p-1}^{(p)}}^{\tau_{s_p}^{(p)}} \phi_p(y_i^*;\Sigma_{y^*}) dy_i^* \,. \tag{4}$$

The PL approach replaces $\ell(\theta;y)$ in (4) with a simpler function, given by the sum of the log-likelihood contributions coming from all the possible $P = p(p-1)/2$ pairs of items. In particular the pairwise log-likelihood computation requires the numerical evaluation of the bivariate normal cumulative density function. To this end, the probability of observing a specific bivariate pattern (s_k, s_l) on the pair of items k and l can be written as

$$\pi_{s_k s_l}^{kl} = p(Y_k = s_k, Y_l = s_l;\theta) = \int_{\tau_{s_k-1}}^{\tau_{s_k}} \int_{\tau_{s_l-1}}^{\tau_{s_l}} \phi_2(y_k^*, y_l^*; \rho_{kl}^{y^*}) dy_l^* dy_k^* \,,$$

where $\rho_{kl}^{y^*} = \lambda_k^T \Sigma_u \lambda_l$ and $\phi_2(x_1, x_2, \rho)$ is the density evaluated at (x_1, x_2) of a bivariate normal distribution with correlation ρ. It follows that the log-likelihood pair contribution can be computed as $\ell_{kl}(\theta;y) = \sum_{s_k=0}^{c_k-1} \sum_{s_l=0}^{c_l-1} n_{s_k s_l}^{kl} \log \pi_{s_k s_l}^{kl}$, where $n_{s_k s_l}^{kl} = \sum_{i=1}^{n} I_i^{s_k s_l}$ and $I_i^{s_k s_l}$ denotes the indicator function of $(Y_{ik}, Y_{il}) = (s_k, s_l)$. The marginal pairwise log-likelihood becomes then

$$\ell^{pl}(\theta) = \sum_{k=2}^{p} \sum_{l=1}^{k-1} \ell_{kl}(\theta) = \sum_{k=2}^{p} \sum_{l=1}^{k-1} \sum_{s_k=0}^{c_k-1} \sum_{s_l=0}^{c_l-1} n_{s_k s_l}^{kl} \log \pi_{s_k s_l}^{kl} \,. \tag{5}$$

2.3 Optimal weights for pairwise likelihood

The expression (5) implicitly assumes that the weight for each pair to be equal, there exists the possibility that a better weighting system for each term can be designed to achieve better statistical efficiency. This was already part of the original proposal of composite likelihood made by [7], who provided a theoretical method for obtaining weights which ensure some optimality properties for the resulting estimator. Such weights are rather complicated to obtain, so that the existing applications are limited to very simple statistical models; see [8] and the references therein. The latter article shows that in the scalar-parameter case one can retrieve a unique weight for each pair, whereas in the d-dimensional parameter case it is often not the case.

In d-dimensional case, the weights take the form of a $d \times dP$ matrix $W(\theta)$. In particular, let consider the dP-dimensional vector of stacked pair scores $\nabla \ell_{kl}(\theta)$ denoted by $g(\theta)$, the $dP \times dP$ covariance matrix of $g(\theta)$ denoted by $V(\theta)$, and finally the $dP \times d$ matrix obtained by stacking variances from $V(\theta)$, denoted by $D(\theta)$. Furthermore, consider a preliminary estimate $\tilde{\theta}$ of the parameter vector θ, which is needed in order to get a consistent estimate of the weights. Therefore, the weights are computed as $W(\tilde{\theta}) = D(\tilde{\theta})^T V(\tilde{\theta})^{-1}$, and the final estimate of θ requires the solution of the d-dimensional nonlinear system of equations

$$W(\tilde{\theta}) g(\theta) = 0 \,. \tag{6}$$

Solving (6) is a challenging task and may often lead to divergent estimates. Furthermore, it is worth stressing that computing the $dP \times dP$ covariance matrix $V(\theta)$ is often a computationally-intense operation.

3 Simulation Study

A comparison between the three methods has been performed on data simulated with the IRT parameterization and logit link. The setting consists of $p = 6$ items, each with 4 categories, $q = 2$ latent variables with correlation $\rho = .7$ and $n \in \{500, 1000, 5000\}$ units. The loading matrix has a simple structure, where each item depends only on one latent factor and each loading is uniformly distributed in $(-1, 1)$. The ML estimates were obtained by the EM method from the `mirt` R package [3], whereas the PL and the optimal weighted pairwise likelihood (OWPL) were obtained with custom `C++` code integrated in R via the `Rcpp` package [4]. In particular, the PL was maximized with the `nlminb` optimizer, whereas the OWPL used the PL estimates as preliminary point to compute the weights, and the package `nleqslv` [5] to solve (6). In order to be comparable to the ML estimates, PL and OWPL estimates have been converted to the IRT parameterization and then scaled by 1.7, as customary in IRT settings for moving between the probit and logit links; see [2] for further comments.

In Table 1 the various methods are compared according to the Mean Squared Error (MSE) and Mean Absolute Bias (MAB) of the estimates. In particular, results are summarized by the median error, classified by the type of parameter considered.

Table 1 Simulation results for ML, PL and OWPL over 100 repetitions for $n \in \{500, 1000, 5000\}$. Median MSE and MAB are reported for parameters in B, α, ρ and the complete vector parameter θ. Median computational time reported in seconds.

		ML		PL		OWPL	
n		MSE	MAB	MSE	MAB	MSE	MAB
500	ρ	0.0085	0.0919	0.0064	0.0798	0.0052	0.0722
	B	0.0390	0.1656	0.1689	0.3295	0.1528	0.3243
	α	0.0164	0.1038	0.0154	0.1007	0.0153	0.0985
	θ	0.0215	0.1180	0.0482	0.1552	0.0456	0.1520
	time (s)	8.1		0.7		13.1	
1000	ρ	0.0025	0.0504	0.0025	0.0504	0.0024	0.0485
	B	0.0381	0.1681	0.1273	0.3190	0.1225	0.3102
	α	0.0085	0.0754	0.0090	0.0767	0.0090	0.0771
	θ	0.0152	0.0952	0.0359	0.1307	0.0344	0.1291
	time (s)	8.0		0.7		12.4	
5000	ρ	0.0004	0.0212	0.0013	0.0359	0.0013	0.0357
	B	0.0360	0.1623	0.1199	0.3145	0.1173	0.3119
	α	0.0019	0.0359	0.0019	0.0348	0.0019	0.0348
	θ	0.0098	0.0646	0.0285	0.0990	0.0283	0.0990
	time (s)	10.7		0.7		12.1	

All in all, the ML provides the best performance. This is most apparent in the results for the complete parameter vector θ, with PL leading to some efficiency loss. The loss may be enlarged by the use of the scaling constant 1.7, which leads to a further approximation of the IRT parameters for PL and OWPL.

The behaviour of OWPL with respect to PL is of some interest. Even though using optimal weights does not lead to a clear improvement for the individual parameters ρ, B and α, the estimates of the vector θ as a whole slightly moves towards the ML estimates, as expected in theory. At any rate, the overall improvement given by the weights is so small that it does not seem to be worth the remarkable computational effort required for computing the weights.

4 Conclusion and ongoing work

In the relatively simple setting considered, the ML estimates turned out to be the most reliable ones. Even though the efficiency loss of the PL is worsened by the approximation required to switch from probit to logit estimates, the loss cannot be neglected in principle. In different settings, with larger p and q, the ML computation may be instead rather challenging to obtain. Further research is needed in order to scale ML methods on higher latent spaces.

On the other hand, PL allows to scale the estimation on larger datasets at the cost of some efficiency loss, which does not seems possible to recover by suitable weighting, as suggested in the simulations proposed here. At any rate, the number of pairs to consider for PL grows with the square of p, which still remains the main challenge in order to scale PL estimation on datasets with a large number of items.

References

1. Bartholomew, D.J., Knott, M., Moustaki, I.: Latent variable models and factor analysis: A unified approach. John Wiley & Sons, Chichester, UK (2011)
2. Camilli, G.: Teacher's corner: origin of the scaling constant $d = 1.7$ in item response theory. Journal of Educational Statistics **19**(3), 293–295 (1994)
3. Chalmers, R.P.: mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software **48**(1), 1–29 (2012)
4. Eddelbuettel, D., François, R.: Rcpp: Seamless R and C++ integration. Journal of Statistical software **40**(8), 1–18 (2011)
5. Hasselman, B.: nleqslv: Solve systems of nonlinear equations (2018). R package version 3.3.2
6. Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., Jöreskog, K.G.: Pairwise likelihood estimation for factor analysis models with ordinal data. Computational Statistics & Data Analysis **56**(12), 4243–4258 (2012)
7. Lindsay, B.G.: Composite likelihood methods. Contemporary mathematics **80**(1), 221–239 (1988)
8. Pace, L., Salvan, A., Sartori, N.: Efficient composite likelihood for a scalar parameter of interest. Stat **8**(1), 1–8 (2019)
9. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Statistica Sinica **21**(1), 5–42 (2011)

Functional Data Analysis

A new functional clustering method: the Functional Clustering and Dimension Reduction model

Un nuovo metodo di clustering funzionale: il modello CDR funzionale

Adelia Evangelista and Stefano Antonio Gattone

Abstract In this work a new general procedure to cluster functional observations in a subspace of reduced dimension is proposed. In particular the method simultaneously performs cluster analysis and dimension reduction of functional data. The Functional Clustering and Dimension Reduction (FCDR) is obtained as the combination of two objective functions: the Functional Principal Component Analysis (FPCA) and the Functional Factorial K-Means (FFKM). The advantage of this approach consists in the optimization of a single global objective function which, by means of the selection of a tuning parameter, incorporates several techniques varying from the FPCA to FFKM including intermediate cases of clustering and dimension reduction.

Abstract In questo lavoro viene proposta una nuova procedura generale per raggruppare dati funzionali in un sottospazio di dimensioni ridotte. In particolare, il metodo esegue simultaneamente la cluster analysis e la riduzione della dimensionalità dei dati funzionali. Il metodo Functional Clustering and Dimension Reduction (FCDR) è ottenuto come la combinazione di due funzioni obiettivo: la Functional Principal Component Analysis (FPCA) ed il Functional Factorial K-Means (FFKM). Il vantaggio di questo approccio consiste nell'ottimizzare una sola funzione obiettivo globale che, attraverso la selezione di un parametro di controllo, include diverse tecniche che vanno dalla FPCA al FFKM includendo casi intermedi di raggruppamento e riduzione dei dati.

Key words: Cluster Analysis, Functional Data, Dimension Reduction

Adelia Evangelista
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: adelia.evangelista@unich.it

Stefano Antonio Gattone
University of Chieti-Pescara, Viale Pindaro, 42, e-mail: gattone@unich.it

1 Introduction

Cluster analysis has been well used in multiple areas [5] and it continues to play a significant role for analysts and researchers. Unsupervised classification (or clustering) methods aim to classify a sample of data into homogeneous groups, without having any previous information about the clustering structure. This work deals with the clustering of functional data which received specific consideration in the last decade, as a result of the relevance of the high frequency data collected with modern technologies ([7]). Several clustering techniques for multivariate data have been suitably modified in order to be applied to functional data ([6]). Jacques and Preda ([9]) made an interesting review of the most important works related to functional clustering. They identify three general approaches: dimension-reduction before clustering, model-based clustering and nonparametric distance-based methods. Several approaches have been introduced for clustering univariate and multivariate functional data. See for example ([2][3][10][17][12]).
This work is based on the Clustering and Dimension Reduction (CDR) technique proposed by Vichi *et al* [15] which simultaneously identifies the optimal subspaces and the optimal clustering for multivariate data. The idea is that dimension reduction can improve the clustering recovery when the data incorporate unimportant information and the clusters lie in a subspace of reduced dimension ([3]). To this end, the proposed general model called Functional Clustering and Dimension Reduction (FCDR) represents an extension to the functional case of the CDR technique. The work is organized as follows. Section 2 describes Functional Principal Component Analysis (FPCA), Functional Factorial K-Means (FFKM)([17]), and Functional Reduced K-Means ([8, 16]) . The objective function of the FCDR model is presented in Section 3. Some concluding remarks are given in Section 4.

2 Background knowledge: FPCA, FFKM and FRKM

2.1 FPCA

A functional random variable X takes values in a Hilbert space of functions defined on a continuous subset of $\mathscr{R}$, say T. Let $\{x_i(t)\}_{i=1}^n$ be the set of observed curves where $t \in$ T. Without loss of generality assume that data are centered, i.e. the mean function has been subtracted and $\sum_{i=1}^n x_i(t) = 0$, for all $t \in T$. Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_J\}$ be a set of basis functions. X can be expanded in terms of basis functions as

$$X(t) = \sum_{j=1}^{J} c_{ig}\phi_j(t) = c_i^T \Phi. \tag{1}$$

FPCA searches an optimal representation of the curves into a functional space of reduced dimension. In other words it tries to minimize the reconstruction error of the

Q-dimensional approximation of the data given by the following objective function:

$$\ell_{FPCA}(\Psi) = \sum_{i=1}^{n} \|x_i(t) - \sum_{q=1}^{Q} a_{iq}\psi_q(t)\|^2 \tag{2}$$

where $a_{iq} = \int_{\mathrm{T}} x_i(t)\psi_q(t)dt$ are the principal component scores, projections of $x_i(t)$ on the orthonormal eigenfunctions ψ_q with $\int_{\mathrm{T}} \psi_q(t)\psi_r(t)dt = 1$ if $q = r$ and 0 otherwise. The eigenfunctions ψ_q belong to the same linear space spanned by the basis Φ, *i.e.* $\psi_q = \sum_{j=1}^{J} b_{jq}\phi_j(t) = b_q^T\Psi$ with $\Psi = \{\psi_1, \psi_2, \ldots, \psi_Q\}$.
Let C be the $n \times J$ matrix of basis coefficients with generic component c_{ij}, W the $J \times J$ matrix with generic component given by $w_{js} = \int_{\mathrm{T}} \phi_j(t)\phi_s(t)dt$, and B the $J \times Q$ matrix with generic column given by b_q. The FPCA loss function can be expressed in matrix form as follows:

$$\ell_{FPCA}(B) = \| \left[C - CWBB^T\right] \Phi\|^2. \tag{3}$$

2.2 FFKM

Cluster analysis and dimension reduction are often performed in tandem ([1]). A standard approach is to first apply a dimension-reduction technique to obtain a fewer number of components than the number of variables, and subsequently using the extracted components for clustering objects (*tandem clustering*). For example, one may apply standard clustering algorithm on the functional principal component scores a_{iq} [11]. A more efficient way to proceed is to implement the clustering simultaneously with dimension reduction. The main subspace functional clustering are the Functional Reduced K-Means (FRKM) ([8, 16]) and the Functional Factorial K-Means (FFKM) ([17]). FFKM applies to functional data the Factorial K-Means technique ([14]) where a clustering and a continuous factorial model are fitted simultaneously to multivariate data.
Let $U = \{u_{ig}\}(i = 1, \ldots, n; g = 1, \ldots, G)$ be the binary matrix defining the cluster parameters, where $u_{ig} = 1$ if $x_i(t)$ belongs to the $g-th$ cluster and 0 otherwise. Let n_g be the size of cluster g, the cluster centroids functions $\bar{x}_g(t) = \frac{1}{n_g}\sum_{i=1}^{n} u_{ig}x_i(t)$ are assumed to lie in a lower dimensional space. FFKM aims at identifying the optimal projection expressing the cluster structure of the data by minimizing the within-cluster deviance of the clusters in the reduced space given by the following objective function:

$$\ell_{FFKM}(U, \Psi) = \sum_{i,g} u_{ig} \| \sum_{q=1}^{Q} \int_{\mathrm{T}} x_i(t)\psi_q(t)dt\,\psi_q(t) - \sum_{q=1}^{Q} \int_{\mathrm{T}} \bar{x}_g(t)\psi_q(t)dt\,\psi_q(t)\|^2 \tag{4}$$

subject to the weight functions $\{\psi_q\}_{q=1}^{Q}$ being orthonormal.

The FFKM loss function can be expressed in matrix form as follows:

$$\ell_{FFKM}(U,B) = \|CWB - H_U CWB\|^2 \tag{5}$$

where $H_U = U(U^TU)^{-1}U^T$ is the projection matrix onto the space spanned by the columns of U.

2.3 FRKM

Functional Reduced K-Means (FRKM) attempts to find an optimal subspace for clustering functional data by maximizing the between cluster deviance in the reduced space.The loss function of FRKM is given by:

$$\ell_{FRKM}(U,\Psi) = \sum_{i=1}^{n}\sum_{g=1}^{G} u_{ig}\|x_i(t) - \sum_{q=1}^{Q}\int_{\mathrm{T}} \bar{x}_g(t)\psi_q(t)dt\,\psi_q(t)\|^2 \tag{6}$$

subject to the weight functions $\{\psi_q\}_{q=1}^{Q}$ being orthonormal.

In matrix form the loss function becomes

$$\ell_{FRKM}(U,B) = \|[C - H_U CWBB^T]\Phi\|^2. \tag{7}$$

3 The Functional CDR

The two methodologies described in Section2 have been shown to perform better than the tandem analysis ([17]). However, one technique maximizes the between cluster deviance (FRKM) and the other minimizes the within cluster deviance (FFKM) in the reduced space. Following Vichi *et al* [15] a more general model can be considered which combines the two methodologies in order to optimize both the within and the between deviance: the clustering and dimension reduction model (CDR). The Functional CDR loss function is developed as the sum of the FPCA loss function given in (3) and the FFKM loss function given in (5):

$$\ell_{FCDR}(U,B) = \alpha\ell_{FPCA}(B) + (1-\alpha)\ell_{FFKM}(U,B). \tag{8}$$

Hence, the minimization of the loss function (8) is a trade-off amid FPCA and FFKM. Clearly, by setting the value of the constant $\alpha = 0$, (8) becomes the FFKM loss function, vice versa fixing $\alpha = 1$ it is equivalent to minimize the loss function of the FPCA. By varying α between 0 and 1 different weights are given to FPCA and FFKM. Interestingly, it can be shown that with $\alpha = 0.5$ the FCDR is equivalent to the FRKM loss function. Therefore, the clustering solution depends on the chosen value of the tuning parameter α, whose selection becomes crucial.

Vichi *et al* [15] proposed to use the pseudo F index in order to automatically select the value of α ([4]). In this work different criterion to select the value of α are

considered and their performance evaluated through a simulation study. A criterion could be clustering stability proposed by [18]. The idea is to evaluate the clustering algorithm on different samples taken from the same population. If the algorithm is good, it should assign the units to the same cluster from one sample to another. An alternative criterion might be the *gap statistic* suggested by [13] for estimating the number of groups in a data set.

4 Future works

A simulation study to evaluate the performance of the FCDR algorithm together with the criterion for the choice of the tuning parameter α is currently under development.
A possible application of the FCDR method could be the identification of homogeneous areas in an environmental setting. Indeed, the FCDR method seems to be an optimal solution when dealing with meteorological variables where correlation and redundant information are usually found in the data.

References

1. Ben-Hur, A., Guyon, I.: Detecting stable clusters using principal components analysis. In: Brownstein, M. J., Khodursky, A.B. (eds.) Functional Genomics, pp.159-182. Human Press (2003)
2. Bongiorno, E.G., Goia, A.: Classification methods for hilbert data based on surrogate density. Comput Stat Data Anal **99(C)**, 204-222 (2016)
3. Bouveyron, C., Jacques, J.: Model-Based Clustering of Time Series in Group-Specific Functional Subspaces. Adv Data Anal Classif **5**, 281–300 (2011)
4. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Commun Stat Theory Methods **3**, 1–27 (1974)
5. Caruso, G., Gattone, S.A., Fortuna, F., Di Battista, T.: Cluster analysis as a decision-making tool: a methodological review. In: Bucciarelli, E., Chen, S.H., Corchado, J. (eds.) Decision Economics: In the Tradition of Herbert A. Simon's Heritage. Advances in Intelligent Systems and Computing, Vol. 618, pp. 48–55. Springer International Publishing, Berlin (2017)
6. Chiou, J. M., Li, P. L.: Functional clustering and identifying substructures of longitudinal data. J R Stat Soc Series B Stat Methodol **69** Issue 4, 679–699 (2007)
7. Hitchcock, D.B., Greenwood, M.C.: Clustering Functional Data. In: Henning, C., Meila, M., Fionn, M., Rocci, R. (eds) Handbook of Cluster Analysis, pp. 265-289. Chapman and Hall/CRC, New York (2015)
8. Gattone, S.A., Rocci, R.: Clustering Curves on a Reduced Subspace. J Comput Graph Stat **21**, 361–379 (2012)
9. Jacques, J., Preda, C.: Functional data clustering: a survey. Adv Data Anal Classif **8**, 231–255 (2014a)
10. Jacques, J., Preda, C.: Model based clustering for multivariate functional data. Comput Stat Data Anal **71**, 92–106 (2014b)
11. Peng J. and Müller H-G. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. The Annals of Applied Statistics, 2(3):1056–1077, 2008

12. Schmutz, A., Jacques, J., Bouveyron, C., Chéze, L., Martin, P.: Clustering multivariate functional data in group-specific functional subspaces. Comput Stat **35**, 1101–1131 (2020)
13. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistics. J R Statist Soc B **63**, Part 2,411-423 (2001)
14. Vichi, M., Kiers, H.A.L.: Factorial K-means analysis of two-way data. Computational Statistics & Data Analysis **37**, 1, 49–64 (2001)
15. Vichi, M., Vicari, D., Kiers, H.A.L.: Clustering and dimension reduction for mixed variables. Behaviormetrika **46**, 243–269 (2019)
16. Yamamoto, M.: Clustering of functional data in a low-dimensional subspace. Adv Data Anal Classif **6**, 219–247 (2012)
17. Yamamoto, M., Terada, Y.: Functional Factorial k-Means Analysis. Comput Stat Data Anal **79**, 133-148 (2014)
18. Wang, J.: consisten selection of the number of clusters via crossvalidation. Biometrika **97**, 4, 893–904 (2010)

Nonparametric functional prediction bands: theory with an application to bike sharing mobility demand in the city of Milan

Bande di previsione non parametriche per dati funzionali: teoria con un'applicazione allo studio della domanda di mobilità bike sharing nella città di Milano

Jacopo Diquigiovanni, Matteo Fontana and Simone Vantini

Abstract The talk will focus on the prediction of a new unobserved functional datum given a set of observed functional data, possibly in presence of covariates, either scalar, categorical, or functional. In particular we will present an approach (i) able to provide prediction regions which could visualized in the form of bands, (ii) guaranteed with exact coverage probability for any sample size, (iii) not relying on parametric assumptions about the specific distribution of the functional data set, and finally (iv) being computational efficient. The method is built on a combination of ideas coming from the recent literature pertaining to functional data analysis (i.e., the statistical analysis of datasets made of functions) and conformal prediction (i.e., a nonparametric predictive approach from Machine Learning). During the talk we will present the general theoretical framework and some simulations enlightening the flexibility of the approach and the effect on the amplitude of prediction bands of different algorithmic choices. Finally, we will apply the method to the joint prediction of bike pick-ups, drop-offs, and unbalance daily curves in the docking station network of the largest bike-sharing provider in the city of Milan (Italy).

Abstract *L'intervento si concentrerà sulla previsione di un nuovo dato funzionale non osservato a partire da un insieme di dati funzionali osservati, eventualmente anche in presenza di covariate scalari, categoriche o funzionali. In particolare presenteremo un approccio (i) in grado di fornire regioni di previsione visualizzabili in forma di bande, (ii) in grado di garantire una copertura esatta per qualsiasi*

Jacopo Diquigiovanni
Dept. of Statistical Sciences, University of Padova, Via Cesare Battisti 241, Padova, 35121, Italy, e-mail: jacopo.diquigiovanni@phd.unipd.it

Matteo Fontana
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. Now at Joint Research Centre - European Commission, Ispra (VA), Italy e-mail: matteo.fontana@ec.europa.eu

Simone Vantini
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy, e-mail: simone.vantini@polimi.it

numerosità campionaria, (iii) che non si appoggia ad ipotesi parametriche sulla distribuzione specifica dei dati funzionali ed infine, (iv) efficiente dal punto di vista computazionale. Il metodo si pone nell'intersezione della recente letteratura relativa all'analisi dei dati funzionali (ovvero l'analisi statistica di set di dati costituiti da funzioni) e quella della predizione conforme (un approccio predittivo non parametrico nato nell'ambioto del Machine Learning). Durante il contributo orale presenteremo il framework teorico generale da noi proposto insieme ad alcune simulazioni mettendo in evidenza la flessibilità dell'approccio e l'effetto sull'ampiezza delle bande di predizione derivanet da diverse scelte algoritmiche. Infine, applicheremo il metodo proposto alla previsione congiunta delle curve giornaliere di ritiro, riconsegna e sbilanciamento delle bici a noleggio nelle stazioni della rete del più grande provider di bike sharing della città di Milano.

Key words: Conformal prediction, Distribution-free prediction set, Exact prediction set, Functional data, Prediction band, Valid prediction band

1 Introduction

Functional Data Analysis ([6] [14]) is a field of Statistics focusing on the statistical analysis of data sets made of functions. Starting from the forerunner work of Jim O. Ramsay [13], several authors have proposed many interesting findings in this field: Functional Boxplots [16], Functional Principal Component Analysis [14] and Functional Linear Regression [14] are only a few examples of works concerning this still very lively area of modern statistics characterized by many still open research problems. Among these, we deal with the issue of creating prediction sets for newly observed functional data. Specifically, given a nominal confidence level $1-\alpha$, the purpose is to develop a method able to output either exact - i.e. guaranteeing a coverage exactly equal to $1-\alpha$ - or valid - i.e. guaranteeing a coverage not smaller than $1-\alpha$ - prediction sets. This crucial and still open challenge in the functional data analysis framework has been faced in two ways. The former approach is based on parametric bootstrapping techniques [3, 2], while the latter one is based on dimensionality reduction techniques [9, 1]. Both approaches carry some drawbacks: first of all, both of them are either based on non-trivial distributional assumptions and/or asymptotic statements. In addition, the former class of techniques is computationally demanding, whereas the latter class is obviously affected by the approximation induced by the dimensionality reduction. In view of this, this contribution will focus on presenting a procedure able to overcome these shortcomings by means of a new approach in the field of Conformal Prediction [18].

2 Short Outline

Our contribution will start with a gentle introduction to Conformal Prediction in the univariate setting [18, 15]. This approach is an innovative nonparametric approach to create prediction sets firstly developed in the field of Machine Learning as a method to construct prediction intervals for Support Vector Machines [8] and already used also in the functional context via the use of a finite-dimension truncated basis expansion [11]. We will show that the core of the approach is the choice of a *nonconformity measure*, namely any measurable function which takes values in $\mathbb{R}^+$ and whose aim is to score the "extremity" of an observation with respect to the other ones. Specifically, we will restrict our focus to the so called Split Conformal approach which allows the computationally efficient construction of finite-sample valid prediction sets under the assumption of exchangeable data by using a "virtual resampling" kind of reasoning.

We will than move to the case of independent and identically distributed functional data, which is our novel original contribution to the existing literature. We will list the properties that a prediction set is expected to satisfy in the functional framework. We will deepen a very relevant aspect of practical interest which concerns the shape of the sets: in particular, we will show that a functional prediction set must be a band [12, 11]. This allows indeed an easy visualization of the prediction set in the natural graphical visualization of functional data [10].

After having introduced Split Conformal prediction for univariate data and the properties required to functional prediction set, we will introduce a new group of nonconformity measures for functional data based on the supremum metric allowing the construction of functional prediction bands. These distribution-free functional bands will be proven to be finite-sample either valid or exact and defined in closed form. Some emphasis will be also placed on the computational cost characterizing the method. The procedure is indeed highly scalable as the computational effort required by the procedure increases only linearly with the sample size n. We will also focus on the the size of the prediction sets returned by the procedure. Different nonconformity measures - belonging to the aforementioned group of nonconformity measures - will be compared in different scenarios through simulation studies.

We will finally extend our proposal to the more general scenario of multivariate functional data with possible scalar, categorical, or functional covariates. We will show that, this extension is theoretically easy and still computationally affordable. We will conclude our contribution presenting an application to the analysis of bike sharing data. The method will be applied to forecast the daily patterns of bike pick-ups, drop-offs, and unbalance in all the different docking stations of the network from the one-day ahead hourly weather forecast of temperature and rain precipitations. In detail, in Fig. 1, the daily drop-off and pick-up rate curves (top left, top right respectively) are reported together with corresponding fitted values (center left, center right) and corresponding residuals (bottom left, bottom right) which are obtained using a concurrent function-on-function regression model with daily temperature and precipitation curves as functional covariates and the day of the week as a categorical covariate, as proposed in [17]. In Fig. 2, an example of conformal

prediction bands for the drop-off rate curve (light blue band) and the pick-up rate curve (red band) of a specific day with joint coverage $1-\alpha = 0.75$ is reported.

A detailed description and a deep theoretical study of our proposal (together with further simulations and details about the application) can be found in [4] and [5]. An introduction to Conformal Prediction can be found instead in the review paper [7].

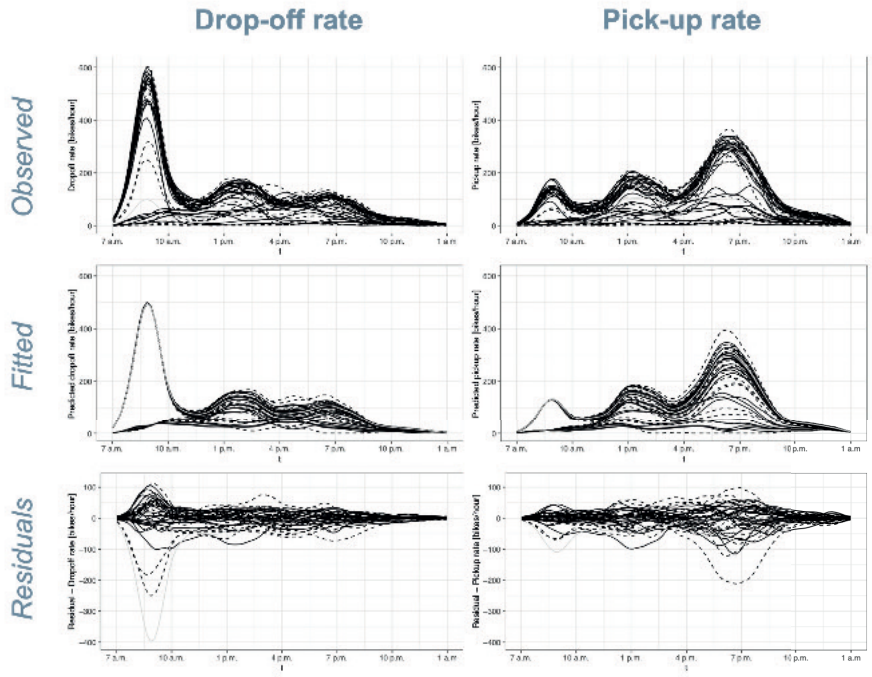


Fig. 1 Daily drop-off and pick-up rate curves (top left, top right respectively), corresponding fitted values (center left, center right), and corresponding residuals (bottom left, bottom right). Continuous curves refer to the observations in the training set, dashed curves to those in the calibration set.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.: A prediction interval for a function-valued forecast model: Application to load forecasting. International Journal of Forecasting. **32**(3), 939-947 (2016)
2. Cao, G., Yang, L., Todem, D.: Simultaneous Inference For The Mean Function Based on Dense Functional Data. Journal of Nonparametric Statistics. **24**(2), 359-377 (2012)
3. Degras, D. A.: Simultaneous confidence bands for nonparametric regression with functional data. Statistica Sinica. **21**(4) (2011)

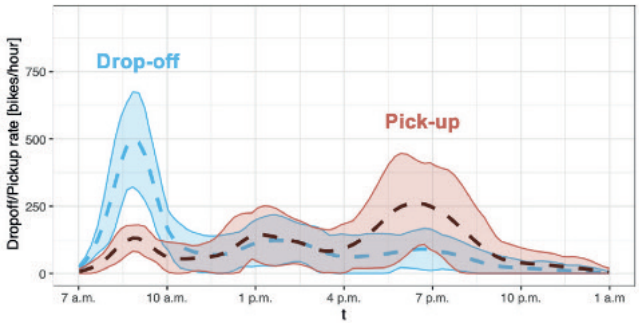


Fig. 2 Example of conformal prediction bands for the drop-off rate curve (light blue band) and the pick-up rate curve (red band) of a specific day with joint coverage $1 - \alpha = 0.75$.

4. Diquigiovanni, J., Fontana, M., Vantini, S.: The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. arXiv:2102.06746 (2021)
5. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal prediction bands for multivariate functional data. Journal of Multivariate Analysis, **189**, art. no. 104879 (2021)
6. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Science & Business Media (2006)
7. Fontana, M., Zeni, G., Vantini, S.: Conformal Prediction: a Unified Review of Theory and New Challenges. Bernoulli, online (2021)
8. Gammerman, A., Vovk, V., Vapnik, V.: Learning by Transduction. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. 148-155 (1998)
9. Hyndman, R. J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis. **51**(10), 4942-4956 (2007)
10. Inselberg, A.: The plane with parallel coordinates. The visual computer. **1**(2), 69-91 (1985)
11. Lei, J., Rinald, A., Wasserman, L.: A conformal prediction approach to explore functional data. Annals of Mathematics and Artificial Intelligence. **74**(1-2), 29-43 (2015)
12. López-Pintado, S., Romo, J.: On the concept of depth for functional data. Journal of the American Statistical Association. **104**(486), 718-734 (2009)
13. Ramsay, J. O.: When the data are functions. Psychometrika. **47**(4), 379–396 (1982)
14. Ramsay, J. O., Silverman, B. W.: Functional data analysis. Springer series in statistics. Second edition (2005)
15. Shafer, G., and Vovk, V.: A Tutorial on Conformal Prediction. Journal of Machine Learning Research. **9**, 371-421 (2008)
16. Sun, Y., Genton, M. G.: Functional Boxplots. Journal of Computational and Graphical Statistics. **20**(2), 316–334 (2011)
17. Torti, A., Pini, A., Vantini, S.: Modelling time-varying mobility flows using function-on-function regression: Analysis of a bike sharing system in the city of Milan. J R Stat Soc Series C.; **70**, 226–247 (2021).
18. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world.Springer Science & Business Media (2005)

An R package for the statistical process monitoring of functional data

Un pacchetto R per il monitoraggio statistico di processo di dati funzionali

Christian Capezza, Fabio Centofanti, Antonio Lepore, Alessandra Menafoglio, Biagio Palumbo, Simone Vantini

Abstract We present the R package `funcharts` for the statistical process monitoring (SPM) of multivariate functional data, which is useful for many industrial applications where data are available as multivariate profiles. Its objective is to build control charts to detect mean shifts in a process. The quality characteristic of interest is a scalar or functional variable and is assumed to be affected by multivariate functional covariates. Functional regression is used to model the relationship between the quality characteristic and the functional covariates in order to improve the ability to detect whether the process is out of control. Moreover, the package allows real-time monitoring of functional data with temporal domain. We show how to use `funcharts` via a real-case study in the shipping industry on the SPM of the CO_2 emissions from navigation of a roll-on/roll-off passenger cruise ship.

Abstract *Quest'articolo presenta il pacchetto R* `funcharts` *per il monitoraggio statistico di dati funzionali multivariati. Il pacchetto è utile per molte applicazioni industriali in cui i dati sono acquisiti come profili multivariati e il cui obiettivo è l'identificazione di derive di processo di una caratteristica di qualità, scalare o funzionale, influenzata da covariate funzionali multivariate. Per modellare la relazione tra la caratteristica di qualità e le covariate viene utlizzata l'analisi di regressione funzionale, in grado di migliorare il potere diagnostico delle classiche carte di controllo per dati funzionali. Viene infine mostrato come il pacchetto* `funcharts` *consente anche il monitoraggio in tempo reale di dati funzionali non completamente osservati. Il pratico utilizzo del pacchetto sviluppato viene mostrato attraverso un caso studio reale su dati raccolti a bordo di una nave da carico e passeggeri.*

Christian Capezza, Fabio Centofanti, Antonio Lepore, Biagio Palumbo
Department of Industrial Engineering, University of Naples Federico II, P.le V. Tecchio 80, 80125 Naples, Italy, e-mail: christian.capezza@unina.it; fabio.centofanti@unina.it; antonio.lepore@unina.it; biagio.palumbo@unina.it

Alessandra Menafoglio, Simone Vantini
MOX - Dept. of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy, e-mail: alessandra.menafoglio@polimi.it; simone.vantini@polimi.it

Key words: Functional data analysis, statistical process monitoring, profile monitoring, R package

1 Introduction

A large amount of operational data about many process variables is nowadays available thanks to the installation of data acquisition technologies with multiple high-frequency sensors installed on board of devices. In this context, there are many applications where the quality characteristics is acquired in the form of multiple profiles, often referred to also as functional data [6, 10]. The increasing amount of available functional data is leading to a growing interest in the literature on profile monitoring [8, 9], whose objective is to check the stability over time of the process through a functional quality characteristic of interest, and to distinguish special causes of variation from normal ones acting on a process.

Most of the current approaches to profile monitoring build control charts to detect out-of-control (OC) conditions, without necessarily considering other variables that may intervene. A few exceptions are the recent works of [3] and [4], which extend the regression control chart [7] in the multivariate case to functional data. In these works, control charts are proposed to improve the effectiveness of the monitoring strategy by considering the influence on the quality characteristic of scalar or functional covariates, in a supervised fashion. In this background, we present the R package `funcharts`, available on CRAN [2], for the Phase II SPM of multivariate functional data.

2 Methodology

Centofanti et al. [4] proposed the functional regression control chart (FRCC) framework based on the function-on-function linear regression model

$$y_i(t) = \beta_0(t) + \sum_{p=1}^{P} \int_{\mathscr{S}} X_{ip}(s)\beta_p(s,t)ds + \varepsilon_i(t), \quad i = 1,\dots,n, \quad t \in \mathscr{T}, \qquad (1)$$

where $y_i \in L^2(\mathscr{T})$ is the functional response variable, $\boldsymbol{X}_i = (X_{i1},\dots,X_{iP})^\top \in \mathbb{H}(\mathscr{S})$ are the functional covariates, $\beta_1,\dots,\beta_P$ are the functional coefficients in $L^2(\mathscr{S} \times \mathscr{T})$, $\beta_0 \in L^2(\mathscr{T})$ is the functional intercept, and $\varepsilon_1,\dots,\varepsilon_n$ are the independent error terms with zero mean. The model can be specialized to a scalar-on-function regression if $y_i(t)$ is constant for all t, as in the work of Capezza et al. [3]. Capezza et al. [3] and Centofanti et al. [4] propose to estimate model (1) by applying multivariate functional principal component analysis to the functional covariates and response. After retaining a subset of the estimated principal components, an approximate multivariate model is obtained, which can be estimated through least squares. The esti-

mated model yields predictions of the response variable, then the prediction error on the response variable can be monitored through the FRCC, which, in the functional response case, is the set of Hotelling's T^2 and the squared prediction error (*SPE*) control charts calculated on the functional residuals. We refer to [3] and [4] for the definition of the two monitoring statistics. In the scalar response case, under normality assumption, the scalar prediction error can be monitored by using control limits based on quantiles of the t distribution. If at least one monitoring statistic is out of the control limits, then an OC alarm is issued. Capezza et al. [3] provide details on how to build contribution plots for fault detection in case an OC is raised.

The R package `funcharts` also implements a real-time monitoring of right-censored functional data where the right edge of the domain progressively moves right over time, as introduced by Capezza et al. [3] for the scalar response case. In fact, hastening the detection of an OC state of the process (before the entire acquisition of the profiles is available) becomes essential in many modern scenarios. We briefly describe below how real-time monitoring is implemented in the `funcharts` package, given a future observation observed in the interval $(0,t)$, where the functional domain is $\mathscr{T} = (0,T)$ and $t < T$. The FRCC is firstly built on the data set obtained by restricting all functional covariates and response in the reference data set used for model estimation to the interval $(0,t)$. Then, the Hotelling's T^2 and *SPE* monitoring statistics are calculated for a future observation to be monitored in real time. By repeating this procedure at several intermediate domain points t between zero and T, real-time evolution of the monitoring statistics is then obtained by means of a reference data set, i.e., a clean data set used to estimate the model and control chart limits.

3 A real-case study

In this section, we aim to demonstrate how the `funcharts` package can be used to build the FRCC in a real-case study on the SPM of CO_2 emissions in the shipping industry. The latter is an important topic because of the dramatic climate change in the last years and urges shipping companies to equip their fleet with multiple devices able to acquire high-frequency operational data about many variables. For each voyage of a ship connecting a departure and an arrival port, these data can often be modeled as profiles.

Data are courtesy of the shipping company Grimaldi Group. For confidentiality reason, the name of the ship, ports and GPS data are omitted and variables are scaled. The data refer to the navigation phase only and all profiles are functions defined on the same domain interval $\mathscr{T} = (0,1)$, which is the fraction of the total distance traveled by the ship at each voyage. The profile of the CO_2 emissions per mile are the quality characteristic of interest and the following four functional covariates are considered as covariates affecting the CO_2 emissions based on the experience of shipping managers, marine engineers and operators, namely the speed over ground,

the trim and the transverse and longitudinal wind components. For more details on the variable description we refer to [1, 5].

In the proposed application, we filter a single route that contains 300 voyages. After an energy efficiency initiative (EEI) performed on the ship, consisting mainly in a silicone foul release coat of the hull, a paint company as described in [5] guarantees a CO_2 emission reduction. In light of this, the reference data set can suitably include profiles related to 159 voyages pertaining to one year before the EEI. The following 141 voyages after the EEI can be then used for the Phase II monitoring.

In the following, we include the main results of the navigation data analysis. First, functional data are obtained from the data frame containing the original discrete data through the function `get_mfd_df`, which returns an object of the class `mfd` for multivariate functional data, provided by the `funcharts` package. After splitting the observations of the `mfd` object into functional response and covariates variables, each in turn split into before and after EEI observations, the function-on-function regression model is fit on the training data (before EEI) using the function `fof_pc`. Finally, based on the output object returned by the regression model and the new data (after EEI), the FRCC is obtained with the function `regr_cc_fof` and is plotted in Figure 1. Note that the monitoring statistics calculated on the before EEI data set are also plotted on the left hand side of the vertical dashed line. The FRCC shows a clear mean shift in the functional residuals since most of the observations collected after EEI are signalled as OC.

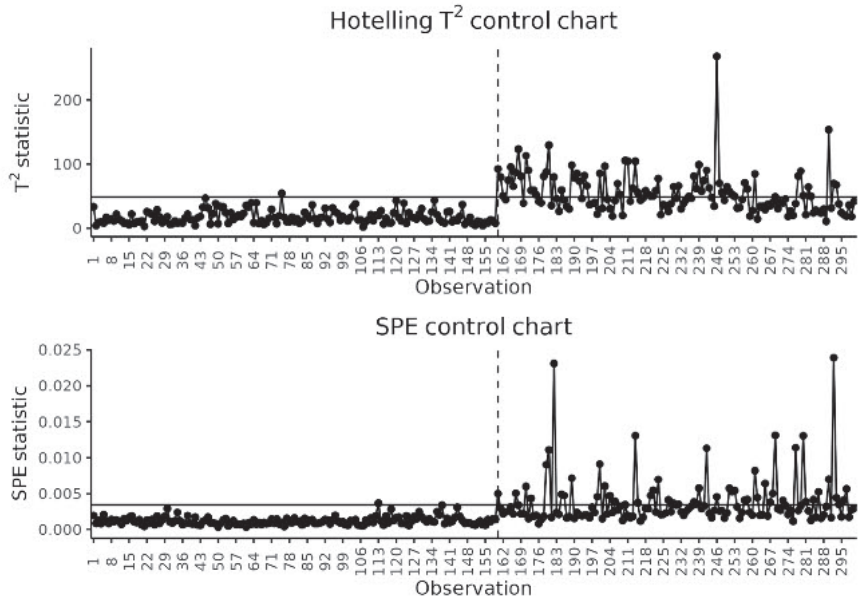


Fig. 1 FRCC used in the real-case study on the navigation data.

Finally, we briefly show how to apply real-time version of the FRCC on the navigation data set. We do not include the code for brevity, however, we report the most relevant result regarding the voyage n. 292, signaled as OC by the FRCC. We can notice that the SPE statistic goes very large in the central part of the domain, being able to anticipate the anomaly long before the end of the voyage.

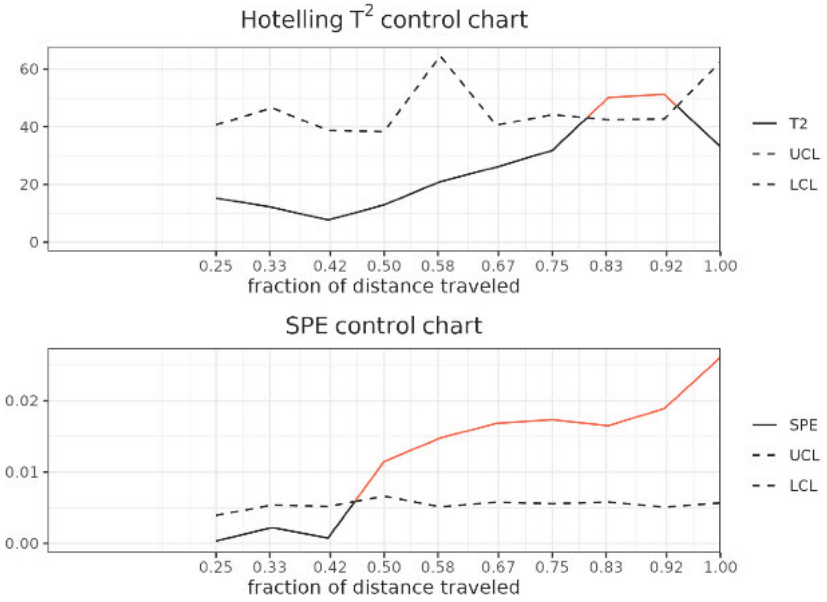


Fig. 2 Real-time FRCC used to monitor the Phase II voyage 292 in the navigation data set.

4 Conclusions

This article introduces the `funcharts` package for the statistical process monitoring (SPM) of multivariate functional data. In particular, it implements through R the recent methods of [3] and [4] and provides tools for building functional data, performing multivariate functional principal component analysis, building control charts for Phase II monitoring of shifts in the functional mean, allowing also the real-time monitoring for functional data with temporal domain. We demonstrate the package with a real-case study in the shipping industry on monitoring CO_2 emissions from navigation. The main added value of this R package is to make available and practically applicabile these novel statistical methodologies for complex data by marine engineers, even though the proposed methodologies are not limited to this

application and can be extended to all applications of SPM in presence of multivariate functional data.

References

1. Bocchetti, D., Lepore, A., Palumbo, B., Vitiello, L.: A statistical approach to ship fuel consumption monitoring. Journal of Ship Research **59**(3), 162–171 (2015)
2. Capezza, C., Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: funcharts: Functional Control Charts (2021). URL https://CRAN.R-project.org/package=funcharts. R package version 1.1.0
3. Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Control charts for monitoring ship operating conditions and co_2 emissions based on scalar-on-function regression. Applied Stochastic Models in Business and Industry **36**(3), 477–500 (2020)
4. Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart. Technometrics **63**(3), 281–294 (2021)
5. Erto, P., Lepore, A., Palumbo, B., Vitiello, L.: A procedure for predicting and controlling the ship fuel consumption: Its implementation and test. Quality and Reliability Engineering International **31**(7), 1177–1184 (2015)
6. Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press (2017)
7. Mandel, B.: The regression control chart. Journal of Quality Technology **1**(1), 1–9 (1969)
8. Menafoglio, A., Grasso, M., Secchi, P., Colosimo, B.M.: Profile monitoring of probability density functions via simplicial functional pca with application to image data. Technometrics **60**(4), 497–510 (2018)
9. Noorossana, R., Saghaei, A., Amiri, A.: Statistical analysis of profile monitoring, vol. 865. John Wiley & Sons (2011)
10. Ramsay, J.O., Silverman, B.W.: Functional data analysis. Springer (2005)

Trend filtering for functional regression

Trend filtering per regressione funzionale

Federico Ferraccioli, Alessandro Casa and Marco Stefanucci

Abstract We propose a penalized approach for scalar on function regression. The method combines ideas from the trend filtering literature and regularization, allowing great flexibility and adaptivity to different degree of smoothness. The method leverages on a proper transformation of the functional predictor based on derivative operators of a certain degree. This allows to define an equivalent generalized lasso problem, that has some similarity to the spline basis approach, but removes the constraints on the locations of the basis. The perfomances of the methods are shown via experimental results.

Abstract *Si propone un approccio di stima penalizzata per regressione funzionale con risposta scalare. Il metodo combina idee provenienti dalla letteratura trend filtering e metodi di regolarizzazione, e permette una grande flessibilità e adattività a diversi gradi di regolarità. Il metodo sfrutta una appropriata trasformazione dei predittori funzionali basata sull'operatore derivata di un dato ordine. Questo permette di definire un problema equivalente in forma di stima lasso, il quale presenta alcune similarità all'approccio spline, senza però richiedere restrizioni sulle locazioni delle basi. Il comportamento del metodo proposto è valutato attraverso uno studio di simulazione.*

Key words: scalar on function, trend filtering, lasso, nonparametric regression.

Federico Ferraccioli
European Commission, Joint Research Centre (JRC), Ispra, Italy,
e-mail: Federico.FERRACCIOLI@ec.europa.eu

Alessandro Casa
Faculty of Economics and Management, Free University of Bozen-Bolzano,
e-mail: alessandro.casa@unibz.it

Marco Stefanucci
Department of Statistical Sciences, University of Rome - La Sapienza,
e-mail: marco.stefanucci@uniroma1.it

1 Introduction

Among the vast class of functional data analysis tools that had risen in popularity in the last decades, functional regression is certainly one of the core building blocks. It found use in spectroscopy, astrophysics and in general when the data are characterized by a temporal dimension. This setting can correspond to either functional predictors or functional responses (see Ramsay and Silverman [2005] for a thorough discussion on this topic).

In this work, we focus on regression models where a scalar response is predicted by means of a functional predictor. Let denote with $\{X_i(t), y_i\}_{i=1,\dots,n}$ the set of observed data, with $y_i \in \mathbb{R}$ being the response, and $X_i(t)$ the functional predictor. A classical formulation in this setting known as functional linear model is given by

$$y_i = \beta_0 + \int X_i(t)\beta(t) + \varepsilon_i \qquad i = 1, \dots, n. \tag{1}$$

Since we are dealing with an infinite-dimensional space of functions, regularization approaches are needed in order to obtain a meaningful and valid estimate of $\beta(t)$. There are two main strategies to address this problem. The first one represents $\beta(t)$ using a p-dimensional basis function, with p carefully chosen in order to effectively apply a regularization, while capturing the relevant features of $\beta(t)$. This lead to an ordinary least square problem, with p playing the role of the smoothing parameter. The second approach involves a penalized least squares estimation procedure to shrink variability in $\beta(t)$. Classical proposals for the penalty take the form of norm of a certain derivative of $\beta(t)$, e.g. $\int \beta(t)^{(d)}(t)^2 dt$, with $d = 2$ being a common choice. The estimation procedure is then recasted as a penalized regression problem, driven by a smoothing parameter λ. Both methodologies have different advantages in different scenarios. Nonetheless, they generates curves that do not adapt in case of different degree of smoothness. Moreover, when for certain regions of the domain $\beta(t) = 0$, the estimates they produce tend to be wiggly and not exactly linear or constant over any region.

Here we propose a flexible approach that is able to overcome these issues, bypassing the basis formulation but still obtaining interpretable estimates. Our proposal leverages on the trend filtering framework developed in Kim et al. [2009]. The idea is to define a proper transformation of the functional predictor $X(t)$ based on the derivative operator, that allows to define an equivalent generalized lasso problem [Tibshirani, 1996]. A similar approach is developed in James et al. [2009]. The proposed method has certain similarity to the basis approach, but removes the constraints on the locations of the basis.

In the next section we outline the methodology, while in Section 3 we show the performances of the proposed methodology via simulations. Section 4 is devoted to a discussion and possible extensions.

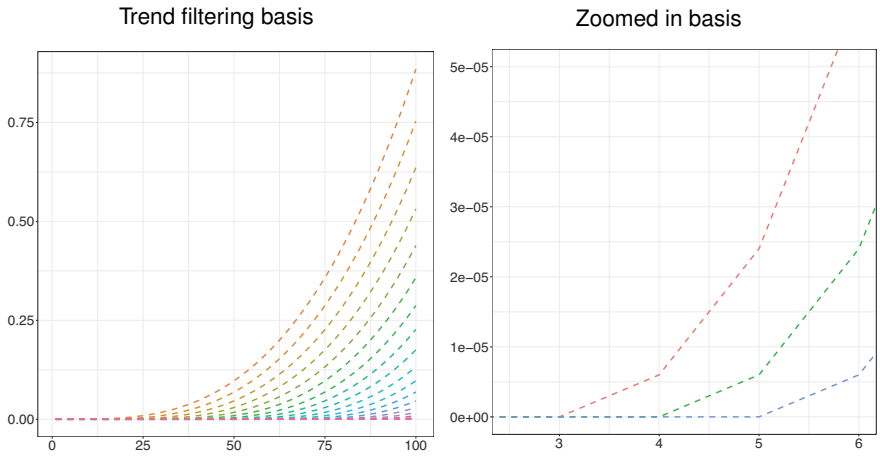


Fig. 1 On the left, the trend filtering basis functions of degree 3. On the right, the same basis are depicted more closely to show their behavior.

2 Methodology

The proposed approach borrows ideas from the classical penalized regression framework but it is rather different in nature. We start by introducing the notation and, afterwards, we discuss the differences and similarities with other methodologies.

Let assume we observe the functional predictors $X_i(t)$ at m distinct input points $(t_1,\dots,t_m)$. We denote by $\mathbf{X}$ the $n \times m$ dimensional matrix of functional predictors, and by $\boldsymbol{y}$ the n dimensional real-valued response vector. Let also $\boldsymbol{\beta} = (\beta(t_1),\dots,\beta(t_m))$ the m-dimensional vector of coefficients at the input points. In order to obtain a valid estimate for the infinite dimensional problem (1), we need to impose some restrictions on the coefficients $\boldsymbol{\beta}$. Here we propose to penalized the ℓ_1 norm of the derivative of $\boldsymbol{\beta}$. This allows to reduce the wiggliness of the estimates while simultaneously shrinking to zero the regions where the functional predictors have no effect on the response.

The resulting penalized least square estimation problem is then defined as

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^m}{\arg\min} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{D}^{(k+1)}\boldsymbol{\beta}||_1 \tag{2}$$

where $\lambda > 0$ plays the role of smoothing parameter and $\boldsymbol{D}^{(i)}$ is the i-th order discrete difference operator.

For general k, the optimization problem in (2) does not have a straightforward solution, due to the complex penalty term. To solve this issue, we propose an equivalent generalized lasso formulation, based on the inversion of the derivative operator. More precisely, let us define $\mathbf{D} \in \mathbb{R}^{m\times m}$ the matrix

$$\mathbf{D} = \begin{bmatrix} \boldsymbol{D}_1^{(0)} \\ \vdots \\ \boldsymbol{D}_1^{(k)} \\ \boldsymbol{D}^{(k+1)} \end{bmatrix},$$

where $\boldsymbol{D}_1^{(i)} \in \mathbb{R}^{1\times m}$ denotes the first row of the i-th discrete difference operator $\boldsymbol{D}^{(i)}$, for $i = 1,\dots,k$ (by convention, $\boldsymbol{D}^0 = \boldsymbol{I}_m$). It is easy to show that the matrix $\mathbf{D}$ is invertible, with inverse $\mathbf{D}^{-1} = \mathbf{M}$ [see Appendix in Tibshirani, 2014]. We then define the vector of transformed coefficients $\boldsymbol{\alpha} = m^k/k! \cdot \mathbf{D}\boldsymbol{\beta}$. By substituting the expression for $\boldsymbol{\alpha}$ in (2), we obtain the equivalent optimization problem

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\arg\min} \left\| \boldsymbol{y} - \frac{k!}{m^k}\mathbf{X}\mathbf{M}\boldsymbol{\alpha} \right\|_2^2 + \lambda \sum_{j=k+2}^{m} |\alpha_j|. \tag{3}$$

Note that (3) takes the form of a lasso problem and can be solved using standard optimization procedures [Tibshirani, 1996].

The proposed formulation (3) has some similarity with locally adaptive regression splines. In particular, the matrix $\mathbf{M}$ takes the form of a basis matrix, where the basis functions are the falling factorial analogues of the truncated power basis in locally adaptive regression splines (see Figure 1). The main difference is that the proposed formulation allows for discountinuities in the derivatives of order $(1,\dots,k-1)$, while adaptive splines maintain continuous derivative. This is a major advantage in our methodology, since it allows to capture different degree of smoothness without introducing oscillatory behaviours in regions where $\beta(t)$ is linear or constant. A similar approach has been proposed in James et al. [2009], where the authors develop a flexible functional regression approach where they combine a classical spline basis formulation with a lasso penalty on the derivative of $\boldsymbol{\beta}$. This differs from our approach, since we do not use spline basis but we model directly the functional predictors.

3 Experimental results

In this Section we present a simulation study to investigate the performances of the proposed procedure. We start from model (1) and we define a function $\beta(t)$ with different degree of smoothness. More precisely, the function is a polynomial of order 3 but is characterized by a flat segment in the central region of the domain (see Figure 2). We set a sample size $n = 500$, with $m = 100$ equispaced input points. We generated 100 samples from model (1) with ε_i i.i.d. Gaussian noise terms with standard deviation 2. We estimated the model via coordinate descent using the `R` package `glmnet` [Friedman et al., 2010] with the order of the penalty set to $k = 0$.

Figure 2 shows the results for two different values of the smoothing parameter. The solid red line represents the true value of $\boldsymbol{\beta}$, while the shaded lines correspond to

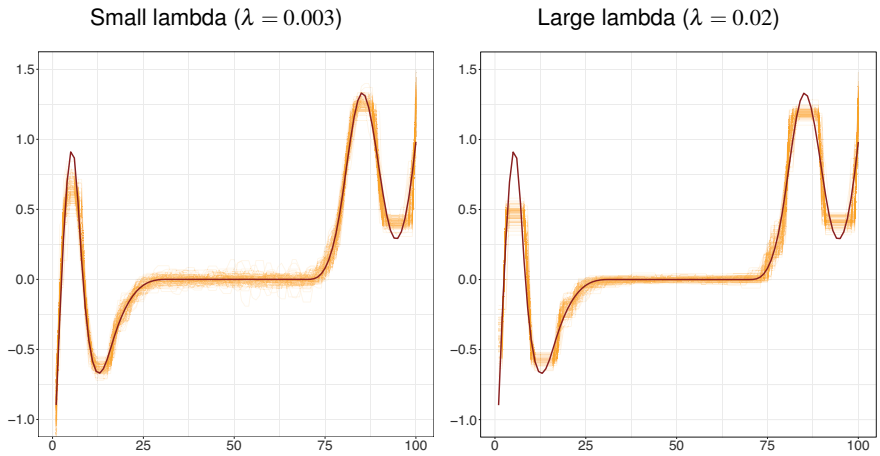


Fig. 2 Estimates for 100 samples from model (1); the solid red line represents the true function, while the background orange lines represent the estimates. Note that for small values of λ, the estimates follow more easily the peaks of the functions, while higher values lead to a more faithful reconstruction of the central regions at the cost of some oversmoothing in the peaks.

the 100 estimates. In both cases, the results look promising as the estimates not only perfectly capture the flat region in the middle of the domain, but also the high curvature peaks. It is worth noting how the choice of λ affects the results. For smaller values, the estimates follow more easily the peaks of the functions, but tend to exhibit more variability in the flat region. On the contrary, higher values of λ lead to a more faithful reconstruction of the central regions at the cost of some oversmoothing in the peaks. This is due to the form of the constraint which, penalizing the first derivative of the estimates, leads to a better estimation of zero-th order functions. We leave the study of higher order constraints to future work.

4 Discussion

We presented a generalization of the trend filtering approach for scalar on function regression. The methodology shows promising results and can be generalized in different ways. First, by imposing appropriate conditions on the functional inputs, we can extend the model to higher order derivatives or even to combinations of multiples derivatives. Given the convex form of the minimization problem, we can also consider alternative optimization procedure to improve the convergence. A possible direction could be given by the ADMM algorithms proposed by Ramdas and Tibshirani [2016]. Future works will also consider the introduction of covariates and the study of inferential procedures such as confidence intervals and hypothesis tests.

References

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL https://www.jstatsoft.org/v33/i01/.

Gareth M James, Jing Wang, and Ji Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108, 2009.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

Aaditya Ramdas and Ryan J. Tibshirani. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

James O. Ramsay and Bernard W. Silverman. *Functional data analysis*. Springer, New York, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Conformal Prediction for Spatio-Functional Regression models

Conformal Prediction per modelli di regressione spazio-funzionale

Andrea Diana, Elvira Romano and Antonio Irpino

Abstract In this work we propose and show a general framework of distribution-free predictive inference for Spatio-Functional Regression models using Conformal Prediction techniques. In particular we focus on Geographically Weighted Regression (GWR) and Heteroskedastic Geographically Weighted Regression (HGWR) with two novel choices of non-conformity measures to compute simultaneous prediction bands.

Abstract In questo lavoro viene proposta e illustrata una generica procedura per fare inferenza non paramentrica su modelli di regressione spazio-funzionale utilizzando tecniche di Conformal Prediction (CP). In particolare l'attenzione viene posta sul modello di regressione geografico pesato (GWR) ed eteroschedastico (HGWR) con due nuove misure di non conformitá per calcolare bande di predizione.

Key words: functional data, spatial dependence, conformal prediction, regression model

1 Introduction

In this work we deal with regression models for spatio-fuctional data [5] and the problem of predictive inference. A crucial challenge for these models, like in classic functional data analysis methods, is quantifying uncertainty in prediction.

In Functional Data Analysis ([7],[9]), methods related to this framework consists in working on parametric bootstrapping techniques ([2], [4]), or in the application of dimensional reduction techniques to manage the naturally infinite dimensional problem ([1], [8]). More recent works are based on a novel approach to forecast by

Andrea Diana
University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: andrea.diana@unicampania.it

Elvira Romano,
University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: elvira.romano@unicampania.it

using Conformal Prediction (CP) ([6]). This last one overcome the previous literature since it is able to output either exact or valid prediction bands under minimal distributional assumptions.
The objective of the present work is to introduce a general framework for distribution-free predictive inference in Geographically Weighted Functional Regression (GWFR) and its heteroskedastic version using conformal inference. A prediction band for the functional response variable is proposed and two new non conformity measures are introduced. Prediction bands with locally varying length are defined, in order to adapt the approach in presence of heteroskedasticity in the data.

2 Spatially dependent Functional Regression models for geostatistical functional data

Geostatistical functional data $(X_{s_1}(t),\ldots,X_{s_i}(t),\ldots,X_{s_n}(t))$ are random functions $X_s(t)$ located in n points $(s_1,\ldots,s_i,\ldots,s_n)$ in $D \subseteq R^d$. Each function is defined on $T=[a,b] \subseteq R$ and is assumed to belong to a Hilbert space with the inner product $\langle X_{s_i}, X_{s_j}\rangle = \int_T X_{s_i}(t)X_{s_j}(t)dt$ [9]. For a fixed site s_i, it is assumed that the observed functions can be expressed according to the model: $X_{s_i}(t) = \mu_{s_i}(t) + \varepsilon_{s_i}(t), \quad i = 1,\ldots,n$ where $\varepsilon_{s_i}(t)$ are zero-mean residuals and $\mu_{s_i}(t)$ is the mean function.
For each $t, t \in T$, the random process is assumed to be second order stationary and isotropic: that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites. It is assumed that the mean function is constant over D and that the semivariogram function $\gamma(h,t) = \gamma_{s_i s_j}(t) = \frac{1}{2}V(X_{s_i}(t) - X_{s_j}(t))$, according to [5], can be expressed by:

$$\gamma(h,t) = \gamma_{s_i s_j}(t) = \frac{1}{2}V(X_{s_i}(t) - X_{s_j}(t)) = \frac{1}{2}E\left[X_{s_i}(t) - X_{s_j}(t)\right]^2. \tag{1}$$

Let's suppose we want to predict a functional response variable $Y_s = \{Y_s(\tau), \tau \in T_1\}$ starting by K functional covariates $\chi_{s_i}(t) = [X_{s_{i,1}}(t),\ldots,X_{s_{i,K}}(t)]^T$. We can consider the geographically weighted regression (GWR) model [11] given by

$$Y_{s_i}(\tau) = \beta_0(\tau, s_j) + \sum_{k=1}^{K}\int_T X_{s_i,k}(t)\beta_k(t,\tau,s_j)dt + \varepsilon_{s_i}(\tau),\ i = 1,\ldots,n, \tag{2}$$

where the function $\beta_0(\tau, s_j)$ is the mean function at location s_j, $\beta_k(t,\tau,s_j)$ is the regression function for the $k-$th covariate at location s_j, and $\varepsilon_{s_i}(\tau)$ is a random error function at point s_i.

In this classical GWR model, the regression coefficients vary geographically and the variance of the error term is assumed fixed. The Heteroskedastic version of this model is obtained by assuming that the variance of the residual model depends on the spatial location. This can be illustrated by observing that local variability of the coefficients in the space may depend on different levels of the spatial variability.

In the HGWR the variance of the model ${\sigma^2}_{s_i}^{GWR}(t)$ is calibrated by replacing $\sigma^2(t)$ with $\sigma^2_{s_i}(t)$. By assuming that ${\sigma^2}_{s_i}(t)$ is a continuous function over the space, it can be estimated by a mean smoother. The final variance $\hat{\sigma}^2_{s_i}(t)$ replaces $\hat{\sigma}^2(t)$ to give

$$ {\sigma^2}_{s_i}^{GWR}(t) = var\{\hat{Y}_{s_i}(t) - Y_{s_i}(t)\} = \hat{\sigma}^2_{s_i}(t)[1 + S_{s_i}(t)] \tag{3} $$

The usual geographical weight at s_i is multiplied by the inverse of $\hat{\sigma}^2_{s_i}(t)$, where this second weighting term corrects for local heteroskedasticity, and resembles the weighted least squares (WLS) in MLR to stabilise a non-constant residual variance. The model has an iterative form, where at each iteration step, a re-weighted H-GWR fit is found.

Statistical inference on these models is strictly related on testing the variability of the coefficients and in constructing confidence interval by using bootstrapping procedures ([11]). In the following we introduce a conformal approach to construct prediction bands such to have finite sample validity, without assumption on the probability models.

3 Conformal Prediction for GWR and HGWR

Consider i.i.d. regression data $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau)) \sim P$, where each $(\chi_{s_i}(t), Y_{s_i}(\tau))$ is a multivariate spatial functional stochastic process or a multivariate functional random field in $\mathscr{L}_2(T)^K \times \mathscr{L}_2(T_1)$, comprised of a response variable $Y_{s_i}(\tau)$ and a k−dimensional vector of features (or predictors, or covariates) $\chi_{s_i}(t) = (X_{s_i,1}(t), \ldots, X_{s_i,K}(t))$. The feature dimension K may be large relative to the sample size n (in an asymptotic model, K is allowed to increase with n). We are interested in predicting a new response $Y_{s_{n+1}}(\tau)$ from a new feature value $\chi_{s_{n+1}}(t)$, with no assumptions on P. Formally, given a nominal miscoverage level $\alpha \in (0,\, 1)$, a prediction band $C \subset \mathscr{L}_2(T)^k \times \mathscr{L}_2(T_1)$ based on $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau))$ is such that

$$ \mathbb{P}(Y_{s_{n+1}}(\tau) \in C(\chi_{s_{n+1}}(t))) \geq 1 - \alpha, \tag{4} $$

where the probability is taken over the s_{n+1} i.i.d. draws $(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau), (\chi_{s_{n+1}}(t), Y_{s_{n+1}}(\tau))) \sim P$, and $C(\chi_s(t)) = \{Y_s(\tau) \in \mathscr{L}_2(T) :\ (\chi_s(t), Y_s(\tau)) \in C\}$ for a point $\chi_s(t) \in \mathscr{L}_2(T)^k$. The defined prediction bands have finite-sample (nonasymptotic) validity, without assumptions on P. These can be obtaining by defining the following algorithm of conformal prediction:

Algorithm of conformal prediction

Input: Data $(\chi_{s_i}(t), Y_{s_i}(\tau))$, $i = 1, \ldots, n$, miscoverage level $\alpha \in (0,1)$, Nonconformity measure $\mathscr{D}$, regression algorithm $\mathscr{A}$, points $\chi_{new}(t) =$

$\{\chi_{s_{n+1}}(t), \chi_{s_{n+2}}(t), \ldots\}$ at which to construct prediction band, and values $Y_{trial}(\tau) = \{Y_{s_{n+2}}(\tau), Y_{s_{n+2}}(\tau), \ldots\}$ to act as trial values
Output: Predictions band, at each element of $\chi_{new}(t)$
for $\chi_s(t) \in \chi_{new}(t)$ **do**
  for $Y_s(\tau) \in Y_{trial}(\tau)$ **do**
    $\hat{Y}_s(\tau) = \mathscr{A}(\{(\chi_{s_1}(t), Y_{s_1}(\tau)), \ldots, (\chi_{s_n}(t), Y_{s_n}(\tau)), (\chi_s(t), Y_s(\tau))\})$
    $R_{Y_s(\tau),s_i} = \mathscr{D}(\hat{Y}_s(\tau), Y_{s_i}(\tau))$, $i = 1, \ldots, n$ and $R_{Ys,s} = \mathscr{D}(\hat{Y}_s(\tau), Y_s(\tau))$
    $\pi(Y_s(\tau)) = \frac{1+\sum_{i=1}^{n} \mathbf{1}_{\{R_{Y_s(\tau),s_i} \leq R_{Y_s(\tau),s}\}}}{n+1}$
  end for
  $C_{conf}(\chi_s(t)) = \{Y_s(\tau) \in Y_{trial}(\tau) : (n+1)\pi(Y_s(\tau)) \leq (1-\alpha)(n+1)\}$
end for
Return $C_{conf}(\chi_s(t)), for\ each\ \chi_s(t) \in \chi_{new}(t)$

The regression algorithm $\mathscr{A}$ in our case is GWR and HGWR for geostatistical functional data. The non-conformity measures are two, an optimally weighted distance for functional data spatially dependent and a distance defined as a covex combination of the functional and the space dimension. $\mathscr{D}$ is an optimally weighted distance for functional data spatially dependent defined as in [10]

$$d_{\omega_s}\left(X_{s_i}(t), X_{s_j}(t)\right) = \sqrt{\int_T \omega_s(t)(X_{s_i}(t) - X_{s_j}(t))^2 dt} \tag{5}$$

where the weight ω_s satisfies $\omega_s \geq 0$ and $\int \omega_s dt = 1$. Using distance in 5, we consider weight functions including both the spatial and functional component. It is a generalisation of [3] to the spatial functional framework for two different spatial domains: the georeferenced and the directed network.
The spatio-functional smooth function is obtained by the following minimisation problem:

$$\omega_s(t) = argmin_{||\omega_s||=1} \frac{\sum_{1 \leq i < j \leq n} V(||\theta_{i,j})||^2_{\omega_s})}{\sum_{1 \leq i < j \leq n} [E(||\theta_{i,j})||^2_{\omega_s})]^2}; \tag{6}$$

with $\theta_{i,j}(t) = a_{i,j}X_i(t) - a_{j,i}X_j(t)$, where $a_{i,j}$ and $a_{j,i}$ are obtained starting from the structure of the spatial domain of interest. The coefficient $a_{j,i}$ is the element reflecting the spatial dependence among functional data and changes according to the spatial grid on which the functional data are observed. In our case, we choose a weight function depending on the spatial variability expressed by a trace-variogram function. Formally we define: $a_{i,j} = a_{j,i} = \hat{\gamma}(h_{i,j})$ where $\hat{\gamma}(h)$ is the estimated trace-variogramm. In alternative, $\mathscr{D}$ is a convex combination of classical functional distance and spatial distance.
An application on environmental studies shows main characteristics and advantages of both the proposed approaches.

References

1. Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.M.: A prediction interval for a function-valued forecast model: Application to load forecasting. Int. J. Forecast. **32**, 939–947, (2016)
2. Cao, G., Yang, L., Todem, D.: Simultaneous Inference For The Mean Function Based on Dense Functional Data. J. Nonparametr. Stat. **24**, 359–377 (2012)
3. Chen H., Reiss, P.T., Tarpey, T.:Optimally Weighted L2 Distance for Functional Data. Biometrics **70(3)**, 516–525, (2014)
4. Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statist. Sinica **21**, (2011)
5. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric **21**, 224–239, (2010)
6. Diquigiovanni, J., Fontana, M., Vantini, S.: The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. arXiv:2102.06746 (2021)
7. Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Verlag (2006)
8. Hyndman, R.J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Statist. Data Anal. **51**, 4942–4956, (2007)
9. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)
10. Romano, E., Diana, A., Miller, C., ODonnell, R.: Optimally weighted L2 distances for spatially dependent functional data. Spatial Statistics **39**, (2020)
11. Yamanishi, Y. and Tanaka, Y.: Geographically weighted functional multiple regression analysis: A numerical investigation. Journal of Japanese Society of Computational Statistics **15**, 307–317, (2003).

Tourism and sport studies

Assessing satisfaction of tourists visiting Italian museums: evidence from the eWOM

Valutare la soddisfazione dei turisti che visitano i musei italiani: evidenze dall'eWOM

Daria Mendola and Valentina Oddo*

Abstract Museum visitors' satisfaction is a priority asset for those museums aiming at being competitive in the cultural tourism sector. In this study, evaluations by visitor-tourists, left in their reviews on museums on the TripAdvisor platform, are analysed. Through web scraping, we collected comments for twelve Italian museums left during the year 2019. The content analysis focussed on components of the tourists' satisfaction among Italian and foreign museum visitors, with emotional response, management, and exhibitions emerging as key concepts in museum's evaluation. Noticeably, some elements emerged as divisive: i.e., they appeared both among less and most satisfied tourist-visitors.

Abstract *La soddisfazione dei visitatori dei musei è un asset prioritario per quei musei che vogliano avere una posizione dominante nel turismo culturale. In questo studio vengono analizzate le valutazioni del visitatore-turista rilasciate come recensioni sulla piattaforma TripAdvisor. Attraverso il web-scraping, abbiamo raccolto i commenti su dodici musei italiani per l'anno 2019. La content analysis ha consentito l'esplorazione delle componenti di soddisfazione di visitatori italiani e stranieri. Le componenti emozionale, di management e le mostre hanno evidenziato un ruolo chiave. Sorprendentemente, alcuni elementi sono apparsi divisivi in quanto compaiono tra i principali motivi sia di soddisfazione che di insoddisfazione.*

Key words: Satisfaction, eWOM, Web-Scraping, Content Analysis, TripAdvisor.

[1] Daria Mendola, Department of Psychology, Educational Science and Human Movement (SPPEFF) University of Palermo, Viale delle scienze, ed. 15, 90128, Palermo, Italy; email: daria.mendola@unipa.it

Valentina Oddo, Department of Economics, University of Messina, Piazza Pugliatti, 98122 Messina, Italy; email: voddo@unime.it, valelisaoddo@gmail.com * corresponding author

1 Introduction

Cultural tourism is one of the main drivers of the Italian tourism flows and plays a crucial role in smoothing tourism seasonality [3]. Moreover, as from Ministry of Culture and Heritage data [9], richness of Italian artistic and cultural heritage is mainly concentrated in museums.

Museum managers have often to deal with the scarceness of funds and need to implement funding initiatives to meet conservation and maintenance costs. A recent trend is to base those strategies on visitors and their needs [4].

Satisfaction of visitors develops loyalty, increases museum reputation (e.g., through word-of-mouth among visitors), and is strictly linked to prospective revenues and funding opportunities [11]. Therefore, monitoring visitors' satisfaction provides museums with tools to plan competitive strategies, and to increase museums' attractiveness. Hence, the satisfaction of visitors is crucial to museums' survival over time.

Social media have dramatically changed the way the reputation of museums build itself. Visitors are currently able to exchange information, experiences, and emotions and communicate their satisfaction in a very easy and globally widespread way. At the same time, or consequently, social media have renewed the museum's approach to communication and marketing, besides revitalizing the learning, entertainment, and fundraising activities [1, 8].

In the last decade, user-generated content (UGC) has been gaining reliability for consumers, thus becoming an important input in their decision-making process, especially for the museum sector [8]. By producing UGC, e.g., via the so-called electronic word-of-mouth (eWOM), non-professional reviewers can influence stakeholders' opinions and share their views on products, services, and other commercial aspects. Museums' performances are continually affected by (and have real effects on) positive and negative eWOM, with a complex interplay still under-investigated. However, although studies using UGC for assessing tourism satisfaction are recently increasing, a relatively small number deal with museum visitors' satisfaction [1,11]. The relevance of opportunities raised by UGC and eWOM and the scarceness of studies on museums visitors' satisfaction through user-generated content were among the main motivations to undertake the present study.

This paper focuses on twelve Italian museums and their tourists-visitors' satisfaction, relying on comments left on the TripAdvisor platform during 2019 (the last year before the start of the pandemic crisis of the Covid-19 virus).

2 Data and Sample Selection

TripAdvisor (TA) is the largest travel site on the web which collects more than 860 million reviews. Through web-scraping, which is the process of extracting data from a website [5], it is possible to download visitors' reviews. Reviews are made by text

describing the visit and the museum facilities, refer to a specific year (the one of the visit), and are associated to an overall rating of the experience (on a 5-point scale).

We evaluated TA as a proper source for the aim of this study: it allowed us to study the attributes associated with the museum visitors' satisfaction by harvesting the information directly from the visitors.

In this study, we selected twelve museums[1] from six well-known Italian cultural tourism destinations, spread over the country: Turin, Venice, Florence, Rome, Naples and Palermo. For each city, museums were chosen from the top of TripAdvisor's ranking. The availability of an adequate number of tourists' comments for the year 2019 was also considered as a selection criterion.

Our target population is made up of comments left on the TA website. We restricted our analysis to those comments left by tourists who visited the twelve Italian museums in 2019. Given that the motivation of the trip/visit is unknown, we assumed as tourists those living outside the city where the museum is. The frame list obtained is made of 9,212 reviews. The frame population was stratified, first by the museum and then by the comment language. Particularly, two lists have been identified for each museum: one for reviews written in the Italian language and one for those in the English language. Hence, a probability sample of reviews (comments) for the twelve museums was drawn (n= 1,200).

3 Methods

The research aims were addressed via "content analysis" [6] and "word frequency analysis" [6, 10]. The former proved to be useful to detect areas of satisfaction and dissatisfaction with museums, gathering comments into nodes. The latter defines the main attributes that generate satisfaction or dissatisfaction by counting how many times a word is present in comments previously arranged by the overall museum experience rating. Hence, through the information collected, we assessed which concepts/words of the UGC are the most nominee in the museum evaluation.

In performing content analysis, we defined six *nodes*, that are logical containers where data is collected and coded following the specific core concept based on the existing literature [2]. Based on the existing literature [4,7], the following nodes have been defined:

1) Permanent and temporary exhibition: this category includes impressions related to exhibitions;

2) Museum management: this category highlights the visitor evaluation of museum management (ticket price, queue, marketing actions, etc.);

3) Staff: this category includes visitors' impressions of the staff competence (e.g., the front desk staff's ability to provide personalized information or the guides' competency in interpreting artworks);

[1] The selected museums are: *Egyptian Museum, Cinema Museum, Doge's Palace, Uffizi, Accademia Gallery, Vatican Museum, Castel S. Angelo, Borghese Gallery, Archaeological Museum, San Severo Chapel, Bourbon Tunnel, Norman Palace*.

4) No-core services: this category collects the impressions given to side-services (shops, cafes, restaurants, etc.);

5) Motivation: this category collects the reasons why the tourist-reviewer visited that museum (to spend time with family, to see a specific exhibition, etc.);

6) Emotional response: this category describes the visitors' feelings, their visit experience, and the assessments of the overall museum product.

The coding procedure (through the NVivo 12 software) worked on the assignment of parts of the comment's text to the specific node [2].

In addition, the word frequency analysis was useful to assess which attributes were more recurrent by the level of the overall rating that tourists gave to museums on the TripAdvisor platform.

Finally, coding of the 1,200 comments and their division according to the museum's final rating between the Italian and English samples gave us information on the opinion expressed by visitors on each specific node.

4 Results and Discussion

For the sake of brevity, this section reports content analysis results only on a selection of six major museums spread over all Italy. One hundred reviews (50 in Italian and 50 in the English language) were randomly drawn for each of the following museums: Turin Egyptian Museum and Venice Doge's Palace (Northern Italy), Florence Uffizi and Vatican Museum (Central Italy), Naples Archaeological Museum and Palermo Norman Palace (Southern Italy). For this comparison, the number of "references" (i.e., how many times an aspect has been mentioned in the overall comments) has been used. A chi-squared test for the homogeneity between the two linguistic groups of comments with respect to the node distribution was done on the pooled sample of the six museums. It assessed that there is no statistically significant difference between the two groups at 1% (χ^2=13.03, p-value=0.02). Table 1 shows the percentage distribution of references by node and museum, pooling together the two languages.

Table 1: *Percentage distribution of the number of references by node (column percentage by museum)*

Node	***Egyptian M.***	***Doge's Palace***	***Uffizi***	***Vatican M.***	***Archaeo-logical M.***	***Nor-man Palace***
Emotional response	30.65	32.14	30.85	29.29	30.6	38.81
Management	29.65	31.12	22.34	33.33	23.28	21.39
Exhibition	19.10	25.51	23.40	18.69	29.74	31.84
Staff	9.55	7.14	9.57	12.12	5.60	3.48
Motivation	7.54	1.02	8.51	3.54	8.62	3.48
No-core services	3.52	3.06	5.32	3.03	2.16	1.00
Total	100.00	100.00	100.00	100.00	100.00	100.00

Hence, the content analysis showed that, regardless of the museum, aspects most frequently mentioned in tourists' comments are those related to emotional response, management, and exhibitions.

When it comes to the word frequency analysis, Tables 2 and 3 show the most common (key)words (respectively in the Italian and the English language comments) mentioned in comments divided by rating.

Table 2: *Most common concepts found in Italian comments by rating (expressed in row percentage) assigned to the museum visit*

	Rating					
Keyword	*	**	***	****	*****	***n***
Visit	13.10	21.82	31.34	29.20	33.79	233
Guide	7.14	14.55	11.94	10.95	16.78	111
Tour	4.76	10.91	14.93	10.22	17.01	109
Exhibition	3.57	10.91	5.97	5.84	7.71	55
Queue (short)	1.19	0.00	2.99	7.30	8.16	49
Ticket (booked)	9.52	12.73	0.00	13.14	3.63	49
Queue (long)	19.05	9.09	11.94	4.38	1.13	40
Staff (bad service)	19.05	3.64	4.48	2.92	0.91	29
Staff (good service)	3.57	0.00	2.99	2.92	4.31	28
Ticket (expensive)	11.90	7.27	2.99	2.92	1.81	28
Audio guide	5.95	1.82	5.97	4.38	2.04	25
Crowded	0.00	7.27	1.49	1.46	1.59	14
Cafeteria	0.00	0.00	1.49	4.38	0.91	11
Shop	1.19	0.00	1.49	0.00	0.23	3
Total	100	100	100	100	100	784

Among Italian comments, the three more frequently reported sources of maximum dissatisfaction (one star) are, in decreasing order: the long queue, bad service from the staff and visits. Noticeably, "visit" is also the most frequently reported word in comments of most satisfied tourists (five stars), immediately followed by tour and guide. A similar ranking was found in English comments of most satisfied tourists: tour, visit, guide. While for the worst rating (one star), English comments focus on bad service from the staff, long queues, and tour.

In the end, the contribution of this paper was to assess the museum tourist-visitors' satisfaction via comments released on TripAdvisor. Although selection bias issues for web-based data are widely acknowledged in literature [1, 10, 11], among their advantages, is the possibility for managers to quickly and easily monitor the museum's performance very frequently and at a relatively low cost.

Further development of this study will regard the assessment of the post-pandemic period to investigate the evolution of visitors' experience.

Table 3: *Most common concepts found in English comments by rating (expressed in row percentage) assigned to the museum visit*

	Rating					
Keyword	*	**	***	****	*****	***n***
Tour	14.71	17.33	18.00	20.67	27.17	221
Visit	11.76	13.33	17.00	20.67	22.08	187
Guide	8.82	6.67	15.00	10.00	16.04	129
Ticket (booked)	6.86	2.67	0.00	6.00	8.11	61
Audio guide	4.90	5.33	7.00	7.33	6.04	59
Ticket (expensive)	8.82	17.33	3.00	10.67	2.08	52
Queue (long)	14.71	8.00	12.00	2.67	1.89	47
Staff (good service)	1.96	8.00	9.00	6.00	3.96	47
Staff (bad service)	17.65	8.00	12.00	4.67	0.00	43
Exhibition	0.98	2.67	3.00	3.33	3.96	32
Queue (short)	0.00	0.00	1.00	3.33	4.34	29
Cafeteria	0.00	0.00	3.00	4.00	2.83	24
Crowded	7.84	10.67	0.00	0.00	0.38	18
Shop	0.98	0.00	0.00	0.67	1.13	8
Total	100	100	100	100	100	957

References

1. Alexander, V., Blank, G., Hale, S.A.: TripAdvisor Reviews of London Museums: A New Approach to Understanding Visitors. Museum Int., 70, 154 – 165 (2018)
2. Bazeley, P., Jackson, K.: Qualitative data analysis with NVIVO. SAGE Publications Ltd, London (2013)
3. Di Lascio, F., M., L., Giannerini, S., Scorcu, A., E., Candela, G.: Cultural tourism and temporary art exhibitions in Italy: a panel data analysis. Stat. Methods. Appl., 20, 519–542 (2011)
4. Gil, S. M., Ritchie, J. R. B.: Understanding the museum image formation process—A comparison between residents and tourists. J. Travel Res., 47(4), 480–493 (2009)
5. Hanretty, C.: Scraping the web for arts and humanities. University of East Anglia, Norwich (2013)
6. Hsieh, H. F., Shannon, S. E.: Three approaches to qualitative content analysis. Qual Health Res, 15(9), 1277-1288 (2005)
7. Kotler, N., Kotler, P.: Can Museums Be All Things to All People? Missions, Goals, and Marketing Role. Mus. Manag. Curatorship., 18(3), 271-287 (2001)
8. Kydros, D., Vrana, V.: A Twitter network analysis of European museums. Mus. Manag. Curatorship., 36(6), 569-589 (2021)
9. MiBACT - Ministero dei beni e delle attività culturali e del turismo: Piano strategico di Sviluppo del Turismo (PST) 2017-2022 -Italia Paese per Viaggiatori (2017). Retrieved at: https://www.ministeroturismo.gov.it/wp-content/uploads/2021/11/Piano-Strategico-del-Turismo-2017-2022.pdf
10. O'Connor, P.: Managing a Hotel's Image on TripAdvisor. J. Hosp. Tour. Manag., 19(7), 754-772 (2010)
11. Su, Y., Teng, W.: Contemplating museums' service failure: Extracting the service quality dimensions of museums from negative on-line reviews. Tour. Manag, 69, 214-222 (2018)

COVID-19 pandemic and tourism demand: a comparison between Spain and Italy

La pandemia da COVID-19 ela domanda turistica: un confronto tra Spagna e Italia

Caterina Sciortino, Ludovica Venturella, Stefano De Cantis

Abstract Studies relating to the COVID-19 pandemic and its effects on tourism demand have been widely described in the literature. The main objective of this work is to quantify the drop in tourism demand, the drastic reduction in travel and changes in travel behavior. In order to undertake this task, a comparison between two important European countries is made: Spain and Italy. The impact of the pandemic on the tourism demand of residents was analyzed by using the two main surveys linked to tourism demand (*Viaggi e Vacanze in Italia e all'Estero* and *Encuesta de turismo de residentes*). Departing from an analysis of micro-data, an overview of the data obtained is also included in the work. The results demonstrate a significant reduction in demand with a consequent change in travel behavior.

Abstract *Gli studi relativi alla pandemia da COVID-19 e i suoi effetti sulla domanda turistica sono stati ampiamente descritti in letteratura. L'obiettivo principale di questo lavoro è quello di quantificare il calo della domanda turistica, la drastica riduzione dei viaggi e i cambiamenti nel comportamento di viaggio. Per fare questo, viene fatto un confronto tra due importanti paesi europei: Spagna e Italia. L'impatto della pandemia sulla domanda turistica dei residenti è stato analizzato utilizzando le due principali indagini legate alla domanda turistica (Viaggi e vacanze in Italia e all'estero e Encuesta de turismo de residentes). Partendo da un'analisi dei microdati, nel lavoro è inclusa anche una panoramica dei dati ottenuti. I risultati dimostrano una significativa riduzione della domanda con un conseguente cambiamento nel comportamento di viaggio*

Key words: Tourism demand; pandemic; microdata; Italy; Spain

Caterina Sciortino
Department of Economics, Statistics and Business. University of Palermo;
e-mail: caterina.sciortino@unipa.it

Ludovica Venturella
Department of Economics, Statistics and Business.University of Palermo;
e-mail: ludovica.venturella@unipa.it

Stefano De Cantis
Department of Economics, Statistics and Business. University of Palermo;
e-mail: stefano.decantis@unipa.it

1 Introduction

The outbreak of the COVID-19 pandemic occurred towards the end of 2019, convulsing the whole world and shifting social, political, and economic equilibria. Since the outbreak of the pandemic, many sectors in the economy began to collapse and, of these, tourism was one of the most affected.. Indeed, it can be said that the tourist sector has suffered a significant downturn in economic activity. According to the UNWTO (2021), international arrivals fell by 72% in the first ten months of 2020, compared with 2019. This drop determined a loss of US$ 935 billion in export revenues (UNWTO, 2021). The Covid-19 pandemic must be considered as a crisis of unprecedented proportions. Indeed, it is characterized by various features and implications unlike other economic or natural disasters (for example, SARS in 2002 or H1N1 in 2009, Skare et al, 2021,). As stated by ISTAT (2020), the cessation in tourist flows was one of the immediate effects of the pandemic crisis. According to DPCM (Prime Ministerial Decree) No. 19 of 25 March 2020, non-hotel accommodation was considered a dispensable part of the tourist economy while hotels could formally continue to operate (ISTAT, 2020).
Since the 1980s, there has been increasing growth in studies investigating the determinants of tourism demand and its changes (Jong, 2020). The academic community is also interested to understand the various factors determining the outbreak of the COVID-19 pandemic on tourism demand (Nguyen et al., 2020; Skare et al., 2021). For instance, Skare et al. (2021) have recognized the following as the most distinctive factors in this regard: geographic distance from ground zero (infection epicenter), the transmission rate and finally, media attention and the associated hysteria which can be considered more pertinent today. Furthermore, the literature regarding COVID-19 and tourism demand can be considered relatively heterogeneous (Schmude et al., 2021). There are several important issues related to this topic: air passenger demand (Gallego and Font, 2020), travel behavior when the pandemic can be said to have 'abated' (Ivanova et al., 2020) and the potential effects on lifestyle and travel in general (Wen et al., 2021).
Furthermore, as the pandemic has affected trends in international tourism demand, it is also important to study the effects of COVID-19 pandemic on national tourism and to understand how travel behaviour has changed. Indeed, according to Zhang et al. (2020) COVID-19 has modified travel behavior by reducing the willingness to travel, thereby increasing the desire to go from point A to point B by private means of transport and to nearer regions. In order to understand the effects of the COVID-19 pandemic on tourism demand, the aim of this paper is to investigate the domestic tourism demand in Italy and Spain between 2019 and 2020. Microdata from the *Trips and Holidays* survey relating to Italy and the *Encuesta de turismo de residentes* survey regarding Spain have been analysed. Various research questions were addressed in achieving this aim: quantifying the drop in the demand for tourism services in Italy and Spain (in terms of trips and excursions); and
understanding how travel behavior has changed in the 2019-2020 period in terms of domestic and foreign destinations, motivations, expenditure, accommodation choice etc.

2 Data processing and description

The analysis was conducted by considering the surveys relating to tourism demand, which had been conducted by ISTAT (in Italy) and INE (in Spain). A focus of the survey relating to the expenditure in Italian households is *Trips and Holidays*. The rationale for this focus is to collate information regarding the tourist movements of Italian residents. The survey is a quarterly, and the object of the research is the number and the features of trips made for holiday or work in Italy and abroad. The survey also includes information relating to the expenditure incurred. The sample comprises household members, which are selected either from a *municipal registry office-LAC* or via the *Anagrafe Nazionale della Popolazione Residente (National Registry Office of the Resident Population or ANPR)* of approximately 32,000 households in 540 Italian municipalities. The data are collected by means of a questionnaire, using the CAPI (Computer-Assisted Personal Interviews) methodology. The *Encuesta de Turismo de Residentes* survey provides monthly, quarterly and annual estimates of trips made by the resident population. Thereafter, the main features of the survey are collated, including expenses, socio-demographic characteristics, etc. The collection method used in this case is by means of telephone calls. The sample method makes reference to stratification criteria in terms of census sections. Microdata for 2019 and 2020 were downloaded from the ISTAT (2022) and INE (2022) websites; the data are included in Tables 1-5 below, considered in million.[1]

Table 1 Trips and Overnight stays in Italy and Spain (2019-2020)

(Millions, unless otherwise specified)

		2020			2019			percentage variation		
		Trips	Overnight	Average stay	Trips	Overnight	Average stay	Trips	Overnight	Average stay
ITA	Hotel	34.130	202.673	5.9	54.253	278.558	5.1	-37.1%	-27.2%	15.7%
	Non-hotel	3.398	28.524	8.4	17.001	130.715	7.7	-80.0%	-78.2%	9.2%
	TOT.	37.527	231.197	6.2	71.254	409.273	5.7	-47.3%	-43.5%	7.3%
SPA	Hotel	96.449	450.689	4.7	173.755	649.516	3.7	-44.5%	-30.6%	25.0%
	Non-hotel	5.074	42.720	8.4	20.120	144.392	7.2	-74.8%	-70.4%	17.3%
	TOT.	101.524	493.409	4.9	193.875	793.908	4.1	-47.6%	-37.9%	18.7%

Table 2 Trips and Overnight stays in Italy and Spain (only Domestic)

(Millions, unless otherwise specified)

		2020			2019			percentage variation		
		Trips	Overnight	Average stay	Trips	Overnight	Average stay	Trips	Overnight	Average stay
ITA	Hotel	14.665	68.100	4.6	24.870	105.577	4.2	-41.0%	-35.5%	9.4%
	Non-hotel	19.464	134.573	6.9	29.384	172.981	5.9	-33.8%	-22.2%	17.4%
	TOT.	34.130	202.673	5.9	54.253	278.558	5.1	-37.1%	-27.2%	15.7%
SPA	Hotel	14.988	48.055	3.2	36.268	115.023	3.2	-58.7%	-58.2%	1.1%
	Non-hotel	81.461	402.634	4.9	137.487	534.494	3.9	-40.8%	-24.7%	27.1%
	TOT.	96.449	450.689	4.7	173.755	649.516	3.7	-44.5%	-30.6%	25.0%

[1] According to EU Regulation (692/2011) all EU countries have to collect information about domestic tourism demand. However, only few countries provide the corresponding microdata quickly available and sufficiently updated.

Table 3 Expenditure per trip, per night and total

(Millions in €, unless otherwise specified)

		2020			2019			Percentage variation		
		Expenditure	Average expenditure		Expenditure	Average expenditure		Expenditure	Average expenditure	
		Total	Per trip	Per night	Total	Per trip	Per night	Total	Per trip	Per night
ITA	Dom	10,722.786 €	314.18 €	52.91 €	16,513.562 €	304.38 €	59.28 €	-35.1%	3.2%	-10.8%
	Out	2,052.475 €	604.05 €	71.96 €	11,435.537 €	672.63 €	87.48 €	-82.1%	-10.2%	-17.7%
	TOT.	12,775.261 €	340.42 €	55.26 €	27,949.100 €	392.24 €	68.29 €	-54.3%	-13.2%	-19.1%
SPA	Dom	17,558.445 €	182.05 €	38.96 €	32,014.006 €	184.25 €	49.29 €	-45.2%	-1.2%	-21.0%
	Out	3,576.611 €	704.82 €	83.72 €	16,051.838 €	797.82 €	111.17 €	-77.7%	-11.7%	-24.7%
	TOT.	21,135.057 €	208.18 €	42.83 €	48,065.844 €	247.92 €	60.54 €	-56.0%	-16.0%	-29.2%

Table 4 Trips and Overnight stays classified by primary motivations

(Millions, unless otherwise specified)

		2020			2019			percentage variation		
	Motivation	Trips	Overnights	Average stay	Trips	Overnights	Average stay	Trips	Overnights	Average stay
ITA	Personal	35.024	417.208	6.4	63.467	382.004	6.0	-44.8%	-41.8%	5.5%
	Pleasure	25.352	158.324	6.2	44.075	253.185	5.7	-42.5%	-37.5%	8.7%
	VFR	8.903	59.888	6.7	17.224	113.708	6.6	-48.3%	-47.3%	1.9%
	Religion	82.539	195.077	2.4	855.572	6.248	7.3	-90.4%	-96.9%	-67.6%
	Health	687.314	3.919	5.7	1.312	8.863	6.8	-47.6%	-55.8%	-15.6%
	Professional	2.503	8.871	3.5	7.788	27.269	3.5	-67.9%	-67.5%	1.2%
	TOT.	37.527	9.288	6.2	71.254	409.273	5.7	-47.3%	-43.5%	7.3%
SPA	Personal	346.701	1.343	4.9	177.748	736.775	4.1	-48.2%	-38.2%	19.3%
	Pleasure	51.250	275.556	5.4	106.631	463.500	4.3	-51.9%	-40.5%	23.7%
	VFR	38.636	168.742	4.4	66.891	256.466	3.8	-42.2%	-34.2%	13.9%
	Religion	254.820	888.442	3.5	796.774	2.778	3.5	-68.0%	-68.0%	0.0%
	Health	1.995	10.370	5.2	3.429	14.030	4.1	-41.8%	-26.1%	27.0%
	Professional	9.388	37.852	4.0	16.127	57.133	3.5	-41.8%	-33.7%	13.8%
	TOT.	356.089	1.381	4.9	193.875	793.908	4.1	-47.6%	-37.9%	18.7%

Table 5 Excursions for destination and for main reasons

(Millions, unless otherwise specified)

Excursions					Excursions			
		2020	2019	Percentage Variation	Motivation	2020	2019	Percentage Variation
ITA	Dom.	40.524	78.604	-48.4%	Personal reasons	40.147	78.191	-48.7%
	Out.	670	1.416	-52.7%	Professional reasons	1.047	1.829	-42.8%
	Tot.	41.194	80.021	-48.5%	Total	41.194	80.021	-48.5%
SPA	Dom.	160.340	257.619	-37.8%	Personal reasons	156.928	250.805	-37.4%
	Out.	1.296	2.971	-56.4%	Professional reasons	4.707	9.785	-51.9%
	Tot.	161.636	260.591	-38.0%	Total	161.636	260.591	-38.0%

3 Results

The total number of trips and overnight stays for domestic and outgoing trips in Italy and Spain are reported in Table 1. The total number of trips in the 2019-2020 period in Italy decreased by 47% (from 71.254 to 37.527 billion); the number of overnight stays also decreased (-43%). A similar result was observed for Spain: -47% regarding trips, -37% for overnight stays. The ratio between domestic and outgoing trips in Italy was 3:1 in 2019 and 10:1 in 2020. A consistency variation in outgoing trips, which was equal to -80%, was also observed. Moreover, 76% of the total recorded trips in Italy in 2019 were domestic. This percentage increased to 91% in 2020.
The ratio between domestic and outgoing trips in Spain was 8:1 in 2019 and 20:1 in 2020. The percentage variation in outgoing trips in the 2019-2020 period was -75%.

Furthermore, the percentage of domestic on total trips in Spain was approximately 90% for 2019 and 95% for 2020. Looking only at the domestic trips (Table 2), relating to Italy, the ratio between nights in hotel and non-hotel establishments is equal to 2:3 in 2019 and 1:2 in 2020.
Moreover, a diminution in nights spent in hotels of -36% was also observed. The percentage loss for Spain was approximately -58% with a ratio between nights spent in hotel and non-hotel establishments of 1:4 in 2019 and 1:8 in 2020. Table 3 regards expenditure that is, total expenditure, expenditure per trip, and expenditure per night. The authors of this paper also proposed a comparison between Italy and Spain for the 2019-2020 period regarding domestic and outgoing trips. And, considering total expenditure, it was observed that there was a similar decrease in the expenditure between Italy (-54%) and Spain (-56%). The highest decreases relating to the expenditure for trips abroad was -82% for Italy and -78% for Spain.
Table 4 details the motivations for trips and overnight stays. Personal motivations are classified as: *Pleasure* (Pleasure, leisure and holiday); *VFR* (Visiting friends and relatives); *Religious* (pilgrimage); and *Health* (treatment, including spas). Table 4 highlights a marked decrease in trips in Italy for personal and professional motivation (-45% and -68% respectively). And an observation of the four aforementioned categories reveals that the highest decrease was recorded in trips for religious reasons and pilgrimages with a decrease of 90%. It would appear that trips in Spain for personal motivation were most affected by the COVID-19 pandemic than trips for professional motivation (-48% and -42% respectively). Finally,Table 5 reports the number of excursions decreased considerably from 2020 to 2019 both in Italy (-48.5% from 80,021 million in 2019 to 41,194 million in 2020) and in Spain (-38.0% from 260.591 million in 2019 to 161.636 million in 2020).

4 Discussion and Conclusion

The COVID-19 pandemic has caused dramatic changes in societies worldwide, also impacting on the tourism industry, in turn inducing a substantial reduction in tourism demand. Utiilizing microdata [2] obtained from surveys relating to tourism demand in Spain and Italy, the aim of this paper has been to quantify the decrease in total trips but also to assess this decrease by considering different aggregates: travel behavior, excursion activity and, particularly, the decrease in terms of expenditure and the effects on tourism markets of these reductions (Plzáková and Smeral, 2021). This decrease concerns all tourism components: incoming and outgoing, hotel and non-hotel establishments, and varying motivations. However, a greater decrease was recorded in the outgoing component relating to hotels and in specific segments (religious trips). These changes in tourism trends during the COVID-19 pandemic derive from different reasons. Indeed, there are several determinants affecting travel decisions and, of these, it is possible to identify various travel constraints, such as: an intrapersonal constraint (a risk associated with 'health'), an interpersonal constraint

[2] However, the nature of this work is mainly descriptive and, it represents a preliminary analysis that will be overcomed through a methodological approach in which the change in tourism demand caused by COVID-19 will be analyzed.

(the absence of friends or relatives as travelling companions), and a structural constraint (social distancing) (Shin et al., 2022).

Only a partial answer to the original research question (to quantify the drop in tourism demand, the drastic reduction in travel and changes in travel behavior post-pandemic) is currently available with the preliminary 2021 data. On the one hand, the main issues of the pandemic concern its long-term effects and, specifically, a diminished degree of willingness to travel to long-haul destinations. Indeed, regional or national destinations (proximity tourism) are currently preferred. On the other hand, of note is the matter of a lasting change in tourism behavior: closer attention to non-hotel establishment solutions (typically smaller and more intimate), aspects relating to the health, and customer service. Changes in travel behavior have been highlighted by Jeon and Yang (2021), according to whom the COVID-19 pandemic has transformed and restricted tourist activities. In the post-pandemic, tourists may well prefer destinations offering safe spaces from the COVID-19 virus. Furthermore, many tourists will wish to avoid human contact with other tourists, investing their time in preventative measures. Changes in travel behavior post-pandemic are currently being extensively studied but it will only be possible to evaluate their impact and effects on the tourism market in the years to come.

References

1. Gallego, I., Font, X. (2021). Changes in air passenger demand as a result of the COVID-19 crisis: Using Big Data to inform tourism policy. Journal of Sustainable Tourism, 29(9), 1470-1489.
2. Jeon, C. Y., Yang, H. W. (2021). The structural changes of a local tourism network: Comparison of before and after COVID-19. Current Issues in Tourism, 24(23), 3324-3338.
3. INE (2022). Residents Travel Survey Methodology https://www.ine.es/dyngs/INEbase/en
4. ISTAT (2022). TRIPS AND HOLIDAYS: MICRODATA FOR RESEARCH PURPOSES https://www.istat.it/it/archivio/178674
5. Ivanova, M., Ivanov, I. K., Ivanov, S. (2021). Travel behaviour after the pandemic: the case of Bulgaria. Anatolia, 32(1), 1-11.
6. Jong, M. C. (2020). Empirical review on tourism demand and COVID-19.
7. Nguyen CP, Schinckus C and Su TD (2020) Economic policy uncertainty and demand for international tourism: An empirical study. Tourism Economics. Epub ahead of print 22 January 2020. DOI: 10.1177/ 1354816619900584.
8. Plzáková, L., Smeral, E. (2021). Impact of the COVID-19 crisis on European tourism. Tourism Economics, 13548166211031113.
9. Škare, M., Soriano, D. R., Porada-Rochoń, M. (2021). Impact of COVID-19 on the travel and tourism industry. Technological Forecasting and Social Change, 163, 120469.
10. Schmude, J., Filimon, S., Namberger, P., Lindner, E., Nam, J. E., Metzinger, P. (2021). COVID-19 and the Pandemic's Spatio-Temporal Impact on Tourism Demand in Bavaria (Germany). Tourism: An International Interdisciplinary Journal, 69(2), 246-261.
11. Shin, H., Nicolau, J. L., Kang, J., Sharma, A., Lee, H. (2022). Travel decision determinants during and after COVID-19: The role of tourist trust, travel constraints, and attitudinal factors. Tourism Management, 88, 104428.
12. UNWTO (2021) Vaccines and reopen borders driving tourism's recovery https://www.unwto.org/taxonomy/term/347)
13. Wen, J., Kozak, M., Yang, S., Liu, F. (2020). COVID-19: potential effects on Chinese citizens' lifestyle and travel. Tourism Review.
14. Zhang, K., Hou, Y., Li, G. (2020). Threat of infectious disease during an outbreak: Influence on tourists' emotional responses to disadvantaged price inequality. Annals of tourism research, 84, 102993.

A compositional analysis of tourism in Europe
Un'analisi composizionale del turismo in Europa

Francesco Porro

Abstract Tourism plays an important role in the economy of many European countries. For this reason many analyses have been performed to improve the understanding of such a relevant sector. In the literature many innovative statistical methodologies have been applied to tourism. In this paper the Decompositional Data (CoDa) approach is used for analyzing the touristic presence in Europe over the years 2016-2021. The considered time range also includes the last two years (2020 and 2021) affected by the COVID-19 pandemic.
Abstract *Il turismo gioca un ruolo importante nell'economia di molti paesi europei. Per questa ragione molte analisi sono state effettuate per migliorare la conoscenza di questo rilevante settore. In letteratura, molte innovative metodologie statistiche sono state applicate al turismo. In questo lavoro l'approccio basato su Dati Compositionali (CoDa) è utilizzato per analizzare la presenza turistica in Europa negli anni 2016-2021. Tale periodo comprende anche gli ultimi due anni (2020 e 2021) in cui l'emergenza sanitaria ha avuto un impatto molto rilevante.*

Key words: Aitchison geometry, Compositional Data, Hospitality, Logratio coordinates

1 Introduction

Since for many European countries, tourism is a very important economic sector, in the literature there are many statistical analyses on it. The (CoDa) methodology is a set of quite recent statistical techniques, that is getting more and more attention in the literature (cfr. [1] and [2]), and it begins to be successfully applied in many fields. An interesting review of many CoDa applications in tourism and hospitality

Francesco Porro
Dipartimento di Matematica
Università degli Studi di Genova, e-mail: francesco.porro@unige.it

can be found in [3]. In the following, after a brief recall of the main definitions, an application about the touristic presence in Europe over the years 2016-2021 based on the CoDa methodology is provided.

2 The Compositional Data (CoDa) methodology

The Compositional Data are multivariate observations where relative rather than absolute information is relevant. For this reason they represent a quantitative description of the parts of some whole. The basic definition in the CoDa methodology refers to the compositions.

Definition 1. A composition vector is a real-valued vector with all (strictly) positive components. A D-part composition is a class of equivalence which contains all the *compositionally equivalent* vectors in $\mathbb{R}^D$. Two compositions $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_D)$ are compositionally equivalent whenever there is a positive constant $c \in \mathbb{R}^+$ such that $\mathbf{x} = c \cdot \mathbf{y}$.

For dealing with equivalence classes, a representative of each class has to be selected. For sake of simplicity, usually in Compositional Data analysis, the vectors of proportions are chosen as representative elements, since their sum is equal to 1. The sample space for the equivalence classes is the D-part simplex $\mathbb{S}^D$, defined by:

$$\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D \ : \ x_i > 0 \ \forall i; \ \sum_{i=1}^{D} x_i = \kappa\}, \tag{1}$$

In formula (1), κ is an arbitrary constant. For further details on the D-part simplex, see [7], [5], [4] and the references therein.

Another important definition is about the *closure* of a composition.

Definition 2. Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ be a the D-part composition. The closure (to κ) of $\mathbf{x}$ is a composition defined as:

$$\mathscr{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right)$$

In the literature the so-called Aitchison geometry on the simplex has been introduced for analyzing compositional datasets in a suitable framework. More details can be found in [7] and in the reference therein.

In CoDa analysis, a dataset $\mathbf{X}$ is a sample consisting of n observations of D-part compositions $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$, with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, $i = 1, 2, \ldots, n$. As described in [7], the standard statistical descriptive measures, based on the real Euclidean structure, should be used with attention in such a dataset, since they can lead to erroneous conclusions. To overcome this issue, the compositional approach suggestes an alternative set of descriptive measures based on the Aitchison geometry. In the following, just the most common two are presented.

Definition 3. For the compositional dataset $\mathbf{X}$, an indicator of central tendency is the closed geometric mean. This vector is called *center*, and it is defined by:

$$\text{cen}(\mathbf{X}) = \mathscr{C}(g_1, g_2, \ldots, g_D),$$

with g_j is the geometric mean of the n observations related to the j-th component of the vectors in $\mathbf{X}$:

$$g_j = \sqrt[n]{\prod_{i=1}^{n} x_{ij}},\ j = 1, 2, \ldots, D.$$

The dispersion level of a compositional dataset $\mathbf{X}$, is evaluated by the normalized variation matrix, defined by:

$$T^* = \begin{pmatrix} 0 & t^*_{12} & \cdots & t^*_{1D} \\ t^*_{21} & 0 & \cdots & t^*_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D1} & t^*_{D2} & \cdots & 0 \end{pmatrix}, \quad \text{where } t^*_{ij} = var\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right).$$

By definition, the normalized variation matrix T^* is symmetric and its diagonal contains only zeros. A synthesis of the variability of a dataset by a single value, is the *Total Variance*, defined as follows.

Definition 4. The Total Variance of the compositional dataset $\mathbf{X}$ is a measure of its global dispersion. It is defined as

$$\text{TotVar}(\mathbf{X}) = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} var\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right) = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} t^*_{ij},$$

where the values t^*_{ij} are the entries of the normalized variation matrix T^*.

As the information convoyed by a composition is relative, an usual practice is to apply transformations, mapping the compositions into real vectors for exploiting the usual Euclidean structure. In the literature there are several transformations based on the logratios: the additive logratio (*alr*), the centered logratio (*clr*), and the isometric logratio (*ilr*). Unfortunately, an overview of these transformations is out of the scope of this paper: the interested reader can see [7] and [5], among others, for a detailed discussion. For the following analysis in this paper, it can just be reported that the *clr*-transformation is basically characterized by two important features: the first one is that it does not modify the number of parts, since a D composition is mapped into a vector in $\mathbb{R}^D$. The second one is that it preserves both the distances and the angles: for this reason, the Aitchison distance in the simplex of two compositions is equal to the distance of the corresponding transformed vectors in $\mathbb{R}^D$ (see [7] for details). This characteristic is very important in exploratory analyses based on metrics, like *clr*-biplots and ternary diagrams. The following one is the definition of the centered logratio transformation.

Definition 5. The centered logratio transformation (clr) of a D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ is given by

$$clr(\mathbf{x}) = \ln\left(\frac{x_1}{g_m(\mathbf{x})}, \frac{x_2}{g_m(\mathbf{x})}, \ldots, \frac{x_D}{g_m(\mathbf{x})}\right),$$

where $g_m(\mathbf{x})$ denotes the geometric mean of the D parts: $g_m(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$.

Graphics are an important instrument for the visualization and the interpretation of data. The most common in Compositional Data analysis are the PCA biplots. The biplots are a representation of the data in dimension 2: they are based on the Single Value Decomposition (SVD) of the centered (or standardized) data matrix. In CoDa analysis, there are two basic kinds of biplots: the *form biplot*, which favours the display of the units, and the *covariance biplot*, which favours the display of the variables. More details on them can be found in [7], [6], and in [5].

3 Application

The data used for the application are provided by Eurostat, and they are available in the section *Industry, Trade and Services*[1]. The considered variable is "*Nights spent at tourist accommodation establishments*", and it is evaluated for all European countries in the time range between Jan-2016 and Nov-2021. There are no data for December 2021, because at the moment, the detection of such month is not yet available. In order to make the analysis more understandable, the original set of countries has been reduced, restricting the observed population to 12 countries: Finland (FIN), Germany (D), Greece (GR), Iceland (IS), Italy (I), Malta (MT), Netherlands (NL), Norway (NOR), Portugal (PT), Slovenia (SLO), Spain (E), and Sweden (SWE). The choice of the countries is motivated by the data availability, the geographic location and the population size. France and United Kingdom are not included because of the large amount of missing data, especially in the last two years (2020 and 2021). For the presence of missing data, also 8 compositions have been removed from the dataset. Since the considered variable is monthly detected, after the described data cleaning, the dataset can be seen as a compositional sample of 63 different 12-part compositions. In order to have a measure of central tendency, the Table 1 shows the sample center: there is a predominance of Germany (D), with the greatest value of the geometric mean equal to 0.267 over the sample period. The Table 2 reports the variation matrix of the dataset. The values in the last column (*Clr variances*) indicate that the maximum variability is associated with Greece (GR), and the Total Variance is equal to 1.426. In order to identify patterns in the data,

Table 1 The center of the 12-part compositions

	D	GR	E	I	MT	NL	PT	SLO	FIN	SWE	IS	NOR
Center	0.267	0.046	0.247	0.230	0.005	0.076	0.042	0.009	0.014	0.038	0.005	0.022

[1] Eurostat website: https://ec.europa.eu/eurostat/databrowser

Table 2 The Logratio variances of the 12-part composition, and the Total Variance.

	D	GR	E	I	MT	NL	PT	SLO	FIN	SWE	IS	NOR	Clr variances
D	0	0.786	0.248	0.142	0.298	0.052	0.091	0.136	0.078	0.080	0.153	0.079	0.089
GR	0.786	0	0.552	0.401	0.762	0.729	0.549	0.540	1.058	0.872	0.745	0.828	0.326
E	0.248	0.552	0	0.139	0.046	0.302	0.060	0.144	0.262	0.337	0.153	0.325	0.107
I	0.142	0.401	0.139	0	0.282	0.133	0.083	0.051	0.218	0.159	0.120	0.137	0.078
MT	0.298	0.762	0.046	0.282	0	0.408	0.106	0.263	0.295	0.436	0.253	0.438	0.149
NL	0.052	0.729	0.302	0.133	0.408	0	0.122	0.168	0.121	0.062	0.203	0.064	0.099
PT	0.091	0.549	0.060	0.083	0.106	0.122	0	0.105	0.147	0.178	0.128	0.174	0.073
SLO	0.136	0.540	0.144	0.051	0.263	0.168	0.105	0	0.192	0.202	0.143	0.178	0.088
FIN	0.078	1.058	0.262	0.218	0.295	0.121	0.147	0.192	0	0.073	0.158	0.077	0.112
SWE	0.080	0.872	0.337	0.159	0.436	0.062	0.178	0.202	0.073	0	0.144	0.009	0.106
IS	0.153	0.745	0.153	0.120	0.253	0.203	0.128	0.143	0.158	0.144	0	0.128	0.097
NOR	0.079	0.88	0.325	0.137	0.438	0.064	0.174	0.178	0.077	0.009	0.128	0	0.102
												Total Variance	1.426

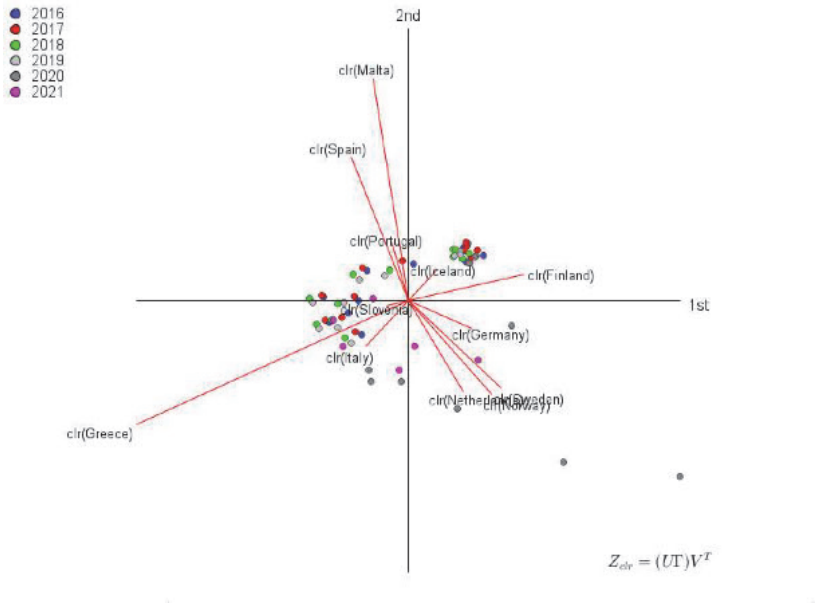


Fig. 1 The Covariance Biplot for the dataset of the 12 countries.

in Figure 1 has drawn the covariance biplot. The first two Principal Components can be considered quite representative of the dataset variability, since the first one explains the 53.33% of the total variance, and the second one bears the proportion of the cumulative explained variance to 80.84%. Following the interpretation rules of the compositional biplots, some remarks can be made from Figure 1.

- The first one regarding the interpretation of the Principal Components. The first Component divides the countries into two subsets: Malta, Spain, Portugal, Slovenia, Greece and Italy (on the left side) and the other countries, Iceland, Finland, Germany, Sweden, Norway and Netherlands (on the right side). This component seems thus related to a ”geographic” factor: on the left there are countries of South Europe, while on the right there are contries located in the North.
- The second Principal Component acts differently: it discriminates between Malta, Spain (at the top) and Greece, Netherlands, Norway and Sweden (at the bottom).
- The length of the rays (the segments joining each vertex to the center) shows that the smallest one is related to Slovenia, suggesting that the logratios of such country provide the smallest contribution to the total variability.

An analysis of the links (segments joining two vertices) is usually explanatory, since orthogonal links suggest a check of the corresponding logratios for zero correlation.

- The link Greece-Finland is (quite) orthogonal to any link with an extreme in the set {Malta, Spain, Portugal} and the other one in {Germany, Netherland, Sweden, Norway}.
- The sames happens for the links Italy-Finland, and Slovenia-Finland.
- The vertices of Sweden and Norway are very close: this means that the variances of the corresponding logratios are small, thus the ratios of their values are (quite) constant. For this reason, these two countries have a common behaviour, and eventually one can be discarded in further analyses.
- Since the projections of many points of the years 2020 and 2021 on each link are far from the projection of the center on the same link, it can be stated that the observations in these two years have a value of the logratio (corresponding to the link) very different from their average on the whole dataset. A reasonable motivation for this is the COVID-19 pandemic occurred exactly in those years.

Further findings can be obtained by the analysis of other plots (as ternary or De Finetti diagrams) and by using the Sequential Binary Partitions (SBP) with their related CoDa-dendrograms.

References

1. Aitchison, J., The statistical analysis of compositional data. J R Stat Soc Series B Stat Methodol 44(2), 139-160 (1982)
2. Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, London, UK (1986)
3. Coenders, G. and Ferre-Rosell, B., Compositional data analysis in tourism: review and future directions. Tourism Analysis 25, 153–168 (2020)
4. Egozcue, J., Pawlowsky-Glahn, V., Compositional data: the sample space and its structure. Test 28, 599-638 (2019)
5. Filzmoser, P., Hron, K., Templ, M. Applied compositional data analysis. Springer Nature (2018)
6. Greenacre, M., Compositional data analysis in practice. CRC Press (2018)
7. Pawlowsky-Glahn, V., Egozcue, J., Tolosana-Delgado, R. Modeling and analysis of compositional data. John Wiley & Sons (2015)

Improving administrative data quality on tourism using Big Data.

Migliorare la qualità dei dati amministrativi sul turismo attraverso l'uso dei Big Data.

Antonella Bianchino, Armando d'Aniello, Daniela Fusco.

Abstract National Statistic Institutes (NSIs) have used the Administrative Data for official statistics production for many years. However, they are characterized by typical errors of the source (e.g., coverage, delay in data registration) that could be solved by integrating other sources. The study shows how the use of Big Data can partially solve some of the errors. Particularly, the information about touristic accommodation obtained through web scraping, has been integrated with regional Administrative Data to improve the final quality assessment.

Abstract *L'uso dei dati amministrativi caratterizza da tempo le statistiche ufficiali degli Istituti di statistica. Tuttavia, essi si caratterizzano per tipologie di errore proprie della fonte (es. copertura, ritardi di registrazione) che possono essere risolti con l'ausilio di altre fonti. Lo studio dimostra come l'uso dei Big Data può intervenire nella risoluzione parziale di tali errori. In particolare, allo scopo di incrementare la qualità dei dati regionali sul turismo, questi sono stati integrati con le informazioni delle strutture ricettive presenti sul web.*

Key words: Administrative data, Big Data, Multisource, Tourism, Data Quality

[1] Antonella Bianchino, Istat; email: bianchin@istat.it.

Armando d'Aniello, Università degli studi di Napoli Parthenope; email: armando.daniello@studenti.uniparthenope.it

Daniela Fusco, Istat; email: dafusco@istat.it.

Introduction

The NSIs ordinarily use the Administrative Data (AD) sources for statistical purposes. By AD we mean the information produced by the Public Administration for administrative purposes which, processed by research institutes, become usable for statistical purposes. The rapid Information Technology (IT) development of Public Administrations (PA), which provides back structured and easily usable information on companies, institutions and individuals, made this integration process possible. However, non-sampling errors characterize the AD. For example, the time lag between the date of presentation of the deed and its reference period: the loss of information due to delay in updating the source and the presence of units cataloged with obsolete classifications. These aspects can reduce the informative power of administrative archives.

This study demonstrates how Big Data can improve the coverage of AD. Particularly we analyzed the Campania Region touristic accommodation affected by over coverage error. The activity status of the structures was evaluated using combining data: two AD sources (Rilevatore Turistico Regionale and Turismo Web), the Big Data source TripAdvisor and the Italian Business Register (Asia). The study shows the effect of the introduction of the Big Data source in the quality assessment of the integrated dataset.

The quality framework

1.1 Quality issues referred to different sources

Using a transformation function, the administrative objects are derived in Statistical units [10]; in this way the AD are aligned with statistical data firstly at the metadata level (through comparison and connection of definitions) and then at the data level. A mismatch between the statistical population of interest and the population identified by the administrative source may depend on several factors:

a) Definitional difference between the target population and the population covered by the administrative source.
b) Incorrect identification of statistical units.
c) Delays and / or missing administrative registrations.

The three factors cause coverage errors (over or under).
Where is possible, more administrative sources should be used. This would allow to reduce both the under-coverage error (by integrating sources that cover different portions of the population) and the over-coverage error (having the possibility of more information to establish which units correctly belong to the desired field of observation) [1].

For the quality assessment, Accuracy has a specific importance: for a long time, it was the main defining characteristic of the quality of a data product [3]. It represents the degree of closeness of computations or estimates to the exact or true values what the statistics were intended to measure. The bias and the variance of the estimates influence the accuracy of the final estimates. Theoretically, in multisource statistics, the use of additional sources should have a positive effect on the accuracy of the output, particularly on the coverage.

In this direction, Big Data play an important role. Big Data sources can generally be described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making" [8]. In addition to generating new commercial opportunities in the private sector, Big Data are potentially a very interesting data source for official statistics, either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers [4].

Among the differences between Big Data and traditional sources is the degree of control that NSIs have on the data acquisition and recording. In traditional surveys, NSIs can plan, design and carry out the acquisition procedures; for the use of administrative data sources, they may have agreements with the data providers [9]. Indeed, in a survey life cycle, coverage errors attain to the dimension of the representation line, i.e. to the target population or the set of units to be studied ("who") [6]. Instead, Big Data could not be representative of the whole population but just a fraction of it, which will probably have specific characteristics that differ from the broader population. In other words, the problem is similar to the one faced by studies dealing with non-probabilistic samples.

1.2 The sources quality evaluation

The administrative data quality assessment was done using the method proposed by Statistics Netherlands [5]. The AD sources used in the study had been evaluated using specific qualitative and quantitative indicators calculated on the three hyper-dimensions Source, Data and Metadata (table 1).

When Big Data sources are used, evaluation plays an important role [2]. Most of the specificities of Big Data are related both to its quick pace of change and in terms of the population covered. To evaluate the population coverage of touristic accommodation hubs, they are compared with Italian statistical institute (Istat) official statistics (table 2), used as benchmark. How table two shows, TripAdvisor has unit numbers similar to Istat data.

Table 1: Quality assessment of Administrative Data, year 2019.

INDICATORS	***SOURCE***	
	Turismo web	***Rilevatore T. Regionale***
Contact	+	+
Purpose	+/o	+
Respons Burden	+	+
Feedback	+	+
Security	+	+
INDICATORS	***METADATA***	
	Turismo web	***Rilevatore T. Regionale***
Clarity	+	+
Comparability	+/o	+
Unique key	+	+
Data treatment	-	-
INDICATORS	***DATA***	
	Turismo web	***Rilevatore T. Regionale***
Technical check	+	+
Over coverage	+	+
Under coverage	?	?
Data treatment	-	-
Not responding units	+	+
Missing partial response	+/o	o/-
Measurment	+/o	o/-
Sensibility	+	o

Legend: + Good, o Reasonable, - Insufficient, ? Not Clear, / Intermediate between values indicated

Table 2: Number of touristic accommodation for source and Province, year 2019.

Province	***Source***					
	Istat (2018)	***TripAvisor***	***Google***	***Booking***	***Kayak***	***Hotels.com***
Caserta	439	353	671	160	5,871	2,515
Benevento	653	244	415	63	160	343
Napoli	3,453	3,714	7,408	4,148	7,103	4,805
Avellino	403	231	399	406	108	1,660
Salerno	2,255	2,517	4,613	680	1,592	2,873
Total	**7,185**	**7,059**	**13,506**	**5,457**	**14,815**	**12,196**

Methodology and results

The AD sources (Turismo Web and Rilevatore Turistico Regionale), the Big Data source (TripAdvisor) and the statistical source (Asia) are integrated applying the Fellegi-Sunter probabilistic model [7] combined with the deterministic model with

rules. Using the integrated dataset containing a new variable valorized by the presence of each unit in each source called SOURCE, a model finalized to establish the activity status of unit was built. The sources have been classified considering the reliability: the Business Register is the primary source for the quality building process; between the two AD sources, Turismo Web resulted the most reliable, according with the results show in paragraph 2.2; lastly, TripAdvisor is the least reliable concerning Big Data's typical errors. It was then attributed a score on each unit considering their presence in each source (variable SOURCE): Asia 0.45, Turismo Web 0.25, Rilevatore Turistico Regionale 0.20, TripAdvisor 0.10.

Summarizing the scores on each unit, the thresholds for the activity status of units have been defined as follows:

- More than 0.56: Active;
- Between 0.45 and 0.55: Eligible;
- Less than 0.44: Not eligible.

The results are shown in table 3. Considering Istat official data as a benchmark, the total difference with the Eligible units into the integrate Dataset is of 36%. In particular, the difference has been determined by the quantity of the structures in the province of Salerno, which, although decreased in number compared to the initial situation, is always twice the value found in the survey on the capacity of the accommodation facilities. This could depend on the presence of non-entrepreneurial managed structures that are not part of the Istat survey field of observation, but could be present in the integrated dataset with score equal to 0.45.

Table 3: Integrated Database touristic accommodation for typology.

Province	***DB units***	***Not Eligible***	***Eligible***	***Actives***	***Eligible+ actives***	***Istat***
Caserta	566	88	199	279	478	439
Benevento	764	151	374	239	613	635
Napoli	5,111	1,898	1,104	2,109	3,213	3,453
Avellino	438	10	198	230	428	403
Salerno	6,395	1,289	2,952	2,154	5,106	2,255
Total	**13,274**	**3,436**	**4,827**	**5,011**	**9,838**	**7,185**

Conclusion

The results show how Big Data can be introduced in statistical productive processes as a partial solution to some administrative data limits. Particularly, they can contribute to reduce the AD time lag error and help the identification of closed accommodation. A step forward could be using Big Data as the main source, supplementing the administrative source. This would allow to identify the structures not yet registered due to delays in the registration of administrative acts by the PA.

The table 4 shows some quality indicators calculated on the integrated dataset. The coverage rate is calculated as the following ratio: (Not eligible units+(1-alpha) Units

solved)/(Units solved+Not solved units+alpha Not solved units). Units solved are the number of units which was possible determine the eligibility; Not solved units are the unknowing eligibility units; alpha is the ratio of not solved units that may be eligible (Eurostat suggests to set alpha=1).

The coverage rate for Campania region is +25.9% (over coverage), respect to +75.2% of over coverage obtained used only the AD sources. If the integrated register is used as a frame, it could be accepted. It's important to underline that probably the over coverage depends on different referring data of the sources. This limit could be overcome by running the model in an ordinary process.

Table 4: Integrated dataset quality indicators by province.

Province	***Coverage rate***	***Eligibility rate***	***Activity rate***
Caserta	15.5%	35.2%	49.3%
Benevento	19.8%	49.0%	31.3%
Napoli	37.1%	21.6%	41.3%
Avellino	2.3%	45.2%	52.5%
Salerno	20.2%	46.2%	33.7%
Total	**25.9%**	**36.4%**	**37.8%**

References

1. Ascari, G., Alexander, P. J., Brancato, G., Burg, T., Zhang, L. C., Waldner, C. : Quality Guidelines for Multisource Statistics (QGMSS). Version 0.8.1. ESSnet Quality of multisource statistics – Komuso, Eurostat (2018)
2. Bogdanovits, F., Degorre, A., Gallois, F., Fischer, B., Georgiev, K., Paulussen, R., Quaresma, S. Scannapieco, M., Summa, D., Stoltze, P. BREAL: Big Data Reference Architecture and Layers. Business Layer. ESSnet Big Data II. Work Package F. Process and architecture. Eurostat. Bruxelles (2019).
3. Brancato G., Boggia A., Barbalace F., Cerroni F., Cozzi S., Di Bella G., Di Zio M., Filipponi D., Luzi O.,Righi P., Scanu M.: Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi. Istat (2016)
4. Cheung, P. Big Data, Official Statistics and Social Science Research: Emerging Data Challenges. Presentation at the December 19th World Bank meeting, Washington (2012)
5. Daas P. On the quality of registers. Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses. The Hague, The Netherlands, 10-11 May 2010 (2016).
6. Groves R. M., Fowler F.J.Jr, Couper M, Lepkowsky J.M, Singer E., Tourangeau R.,: Survey Methodology. Wiley, New York. (2004).
7. Fellegi, I. P., Sunter, A. B. A theory of record linkage, Journal of the American Statistical Association: Vol. 64, No. 328, pp. 1183-1210 (1969).
8. Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. Report of the McKinsey Global Institute, McKinsey & Company (2011).
9. Quaresma, S., Maślankowski, J., Salgado, D., Ascari, G., Brancato, G., Di Consiglio, L., Righi, P., Tuoto, T., Daas, P., Six, M., Kowarik, A.: Revised Version of The Quality Guidelines for Acquisition and Usage of Big Data. ESSnet Big Data II. Eurostat (2020)
10. Wallgren A. and Wallgren B. Register-based Statistics: Administrative Data for Statistical Purposes. Second Edition. John Wiley & Sons, Chichester, UK. (2014)

Geographical variations of socio-demographic issues

Elderly HCE and health care need: comparing spatially unexplained levels

Spesa e bisogno assistenziale delle popolazioni anziane: un confronto spaziale dei livelli non predetti

Irene Torrini, Laura Rizzi and Luca Grassetti

Abstract The increasing trend of HCE calls for a deeper evaluation of geographical variations of elderly expenditures for health services compared with their health care burden. Using 2019 data at the municipality level on the elderly population (65+) in Friuli-Venezia Giulia, we estimate Spatial Error Models for the total per-capita health expenditures and the percentage of individuals with high health care burden (classified into Resource Utilization Bands). The determinants of HCE and health care burden are investigated and the models' residuals are examined to derive health-economic scenarios of great relevance for policy-makers.

Abstract *La continua crescita delle spese assistenziali rende necessaria un'analisi geografica approfondita della spesa (quale proxy della domanda) e del bisogno assistenziale della popolazione anziana. Utilizzando i dati 2019 della popolazione anziana a livello comunale in Friuli-Venezia Giulia sono stati stimati Spatial Error Models per la media della spesa sanitaria pro-capite e per la percentuale di anziani con maggiore severità clinica (misurata attraverso la scala dei Resource Utilization Bands). I modelli hanno identificato quali sono i fattori predittivi delle variabili di studio, mentre i residui sono stati esaminati per trarre differenti scenari economico-sanitari di grande rilevanza per le scelte di politica economica e sanitaria.*

Key words: Elderly HCE, ACG-RUB, Spatial Error Models, Residuals' Spatial Comparison

---

Irene Torrini
Dipartimento di Scienze Economiche e Statistiche, Università di Udine, via Tomadini 30/A, e-mail: irene.torrini@uniud.it

Laura Rizzi
Dipartimento di Scienze Economiche e Statistiche, Università di Udine, via Tomadini 30/A, e-mail: laura.rizzi@uniud.it

Luca Grassetti
Dipartimento di Scienze Economiche e Statistiche, Università di Udine, via Tomadini 30/A, e-mail: luca.grassetti@uniud.it

1 Aim and Background

The rise in health care expenditures (HCE) observed over the past 50 years is among the major concerns for the stability of public finances. This trend calls for continuous evaluation of health system efficiency, with unwarranted geographic variations in HCE widely assessed to identify under-use or over-use of services [15]. Shares of heterogeneity that are not explained by differences in the demographic structure of geographical units may, in fact, reflect inefficiencies. According to the existing literature, such unexplained variation is accounted for by differences in supply of services [6, 7, 4], socio-economic characteristics [5], technology [8, 11, 14], and policy makers' behaviour, political ideologies and contextual factors [13, 4].

In this paper, we investigate the heterogeneity at the municipality-level in health care burden and expenditures by comparing spatial patterns of unexplained shares of HCE and care need. The aim is the identification of the geographical units showing abnormal levels of HCE with respect to health care need or vice-versa.
We focus on the context of the Italian Region Friuli-Venezia Giulia (FVG) and the over-65 population. This subpopulation is characterised by severe health conditions, high care needs and, consequently, high health care expenditures. Several studies highlight the age-dependent nature of individual health care costs [12, 9]. They exhibit a J-shaped curve that increases slowly through adulthood and then exponentially after age 50 [1], with a large share of HCE taking place near the time of death [10, 16]. With a population average age of 47.31 years, a share of individuals over-65 equal to 29.08%[1] and a population that ranks as the second oldest in Italy, FVG represents the ideal setting for analyzing the elderly cohort.
By using administrative data at the municipality level, we analyse the per-capita total health care expenditures (total HCE hereafter) and the percentage of over-65 individuals with high health care needs accounting for their main determinants and the spatial correlation among disturbances. The sign and magnitude of models' residuals are then exploited to discuss the different health-economic scenarios.

2 Data Description

For our analysis, we use two datasets recording information on the over-65 population observed in the year 2019. The first one is drawn from the Health Care Information System (SISSR) of the Region Friuli-Venezia Giulia and provides information on age, gender, survival status, municipality of residence and individual health care expenditures covered by the public health care system for different types of services (hospital, outpatient, pharmaceutical and home care), aggregated into total health care expenditures (total HCE). The second dataset provides information on resource use by assigning each individual to an Adjusted Clinical Group (Johns Hopkins ACG System [2]). ACGs are mutually exclusive health status cate-

[1] Statistics drawn from the Health Care Information System of the Region Friuli-Venezia Giulia.

gories defined by morbidity, duration, severity, diagnostic certainty, aetiology, age, and gender [3]. Specifically, within each morbidity pattern, ACGs define clinical groups of individuals expected to require similar levels of health care resources. ACGs are combined into six Resources Utilization Bands (RUBs), where individuals with higher bands are expected to present higher clinical severity and consequently higher care needs, being more likely multi-morbid. We aggregate individual data at the municipal level and add to our dataset information on municipalities' socio-economic, health care supply and geographical characteristics, which may contribute to the territorial variation in expenditures and health care burden. They are obtained from, respectively, records of the National Institute of Statistics (ISTAT), the Italian Internal Revenue Service (*Agenzia delle Entrate*), and the Italian Ministry of Health.

3 Empirical Strategy and Results

The main aim of this analysis is to describe the spatial heterogeneity of total HCE, expressed in logarithm, and percentage of RUB equal to four or five (RUB_{45} hereafter) in the elderly population at the municipality level. Our approach is based on two stages. In the first step, we describe the behaviour of the two study variables through Spatial Error Models (SEMs). In the second step, we study the geographical patterns of the unexplained shares of elderly health care expenditure and burden across regional municipalities. The spatial models are estimated separately for the total HCE and the percentage of older adults with RUB_{45}. While spillover effects do not appear significant in preliminary analyses, the presence of unexplained similarities across elderly residing in neighbouring municipalities supports the introduction of spatial autocorrelation in the models' error component. Table 1 shows the results of statistically significant parameters[2]. One of the most interesting finding concerns the spatial autocorrelation in the error component, which is significant for HCEs but not for RUB_{45}. The different spatial pattern of unexplained shares of expenditures and health burden deserves further analysis, described in the next section.

To analyse the unexplained part of the geographical variation in our interest variables, we plot in Figure 1 the spatial distribution of unpredicted total HCE (left panel) and RUB_{45} (right panel). Green areas indicate municipalities where observed values are higher than the predicted (positive residuals - under-estimation), while red areas indicate the opposite scenario (negative residuals - over-estimation). In both cases, increasing colour tones correspond to increasing residuals' absolute values, while residuals near to zero are reported in white. This plot gives the first insight into the geographical pattern of unexplained levels of HCE and health care burden. For example, the mountainous area in the north-eastern part of the region presents relevant positive residuals for the percentages of RUB_{45} and negligible residuals for total HCE, pointing out the excess in the observed needs with respect to the pre-

[2] The coefficients of the disease dummies are omitted for space reasons.

Table 1 RUB_{45} and Total HCE estimation.

	RUB_{45}			**Total HCE**		
	Demo.	Diseases	Income	Demo.	Diseases	Income
Average age	0.0172***	0.0072**	0.0063*	0.0677***	0.0490**	0.0597***
	(.0032)	(.0025)	(.0026)	(.017)	(.016)	(.016)
Female	-0.1817*	-0.0797	-0.0588	-0.8307	-0.7430	-0.9844*
	(.081)	(.062)	(.064)	(.44)	(.4)	(.41)
Chronic diseases	No	Yes	Yes	No	Yes	Yes
Low income[a]			0.0493			-0.4864*
			(.034)			(.21)
Constant	-1.0259***	-0.5074**	-0.4628**	3.0526*	3.6342***	3.0855**
	(.23)	(.17)	(.17)	(1.2)	(1.1)	(1.1)
Spatial correlation	0.5425***	0.3596**	0.3598**	-0.0322	0.0003	-0.0316
	(.083)	(.13)	(.13)	(.13)	(.15)	(.14)
Var(ε_*)	0.0007***	0.0003***	0.0003***	0.0204***	0.0144***	0.0141***
	(.0001)	(< .0001)	(< .0001)	(.0020)	(.0014)	(.0014)
AIC	-927	-1052	-1055	-215	-262	-265
BIC	-1055	-991	-988	-199	-198	-198

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
[a] Percentage of individuals with annual income below 10000€.

dicted ones. This result does not correspond to a coherent excess in the observed expenditures. For the systematic comparison among all municipalities, we classify our geographical units according to the simultaneous concordance or discordance in the signs of the residuals of total HCE and RUB_{45}. We obtain four scenarios: municipalities with concordant over-estimated values, municipalities with concor-

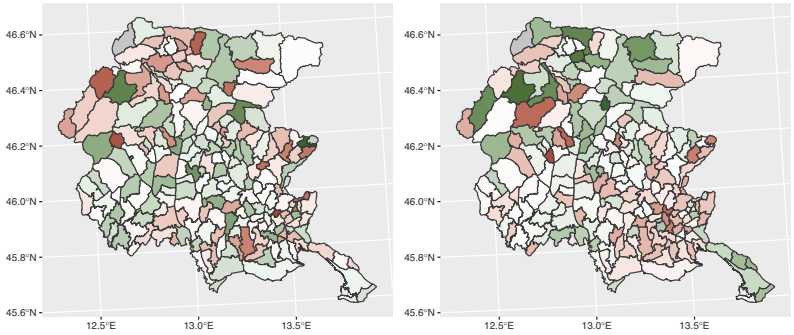


Fig. 1 Distribution of residuals at municipality level of the SEM model for total HCE (left panel) and RUB_{45} (right panel). The residuals shade from red (negative values) to green (positive values) identify over and under-estimated outcomes, respectively.

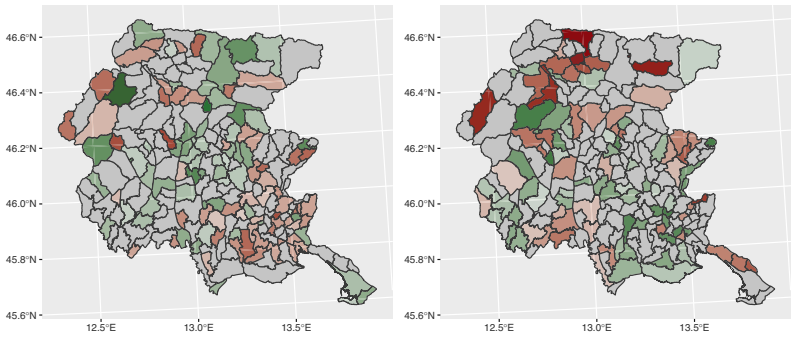


Fig. 2 Municipalities presenting concordant (left panel) and discordant (right panel) residuals. Red and green areas correspond to negative and positive total HCE residuals, respectively.

dant under-estimated values, municipalities with under-estimated HCE and over-estimated burden, and municipalities with over-estimated HCE and under-estimated burden. Figure 2 shows the spatial distribution of the scenarios with concordant (left panel) and discordant residuals (right panel). In both panels, red tones correspond to over-estimated HCE (hence, over-estimated RUB_{45} in the left panel and under-estimated RUB_{45} in the right panel), while green tones to under-estimated HCE (hence, under-estimated RUB_{45} in the left panel and over-estimated RUB_{45} in the right panel). Increasing colour tones correspond to an increasing joint magnitude of standardised residuals, measured through the Manhattan distance of the residuals coordinates from zero. The left panel indicates coherent patterns between expenditures and care needs. It points out several areas where the excess in observed HCE with respect to the estimated values is justified by a similar relationship between observed and estimated health care burden. A few cases present the opposite behaviour. The right panel is instead more relevant from a policy perspective, as it shows scenarios in which an excess in observed HCE is accompanied by under-estimated RUB_{45} or vice-versa. Such critical patterns require further investigations in each specific area in which they occur. In particular, the mountainous area in the northern part of the region is characterised by observed HCE lower than expected and an excess in the observed health burden, indicating partial coverage of the need for medical care. Similar situation is faced in the municipalities near the urban area of Trieste. In this city, instead, a concordance in the excess of observed HCE and burden is revealed. Finally, there are no critical scenarios in the other coastal municipalities and in plain or hilly areas.

4 Conclusions

In this work, we analyse the geographic variation in total HCE and health care burden at the municipality level by estimating two separate SEMs to account for spatial heterogeneity among unobserved factors. The geographical distribution of the residuals is then investigated to identify areas with inconsistent relationship between care needs and expenses, with particular remark for municipalities denoting partial coverage of the elderly need for medical care. Further developments concern the separated analysis of different kinds of health services (SUR spatial models) and time patterns (Panel spatial models).

References

1. Alemayehu, B, Warner, K. E.: The lifetime distribution of health care costs. Health services research, 39(3):627–642 (2004)
2. The Johns Hopkins ACG Case-Mix System Technical Reference Guide. Version 11.0. (2014). https://www.hopkinsacg.org/document/acg-system-version-11-technical-reference-guide/. Accessed 9 Feb 2022
3. Baltimore, M.D.: Johns Hopkins Bloomberg School of Public Health, The Johns Hopkins ACG, version 12.1., (2020)
4. Costa-Font, J., Pons-Novell, J.: Public health expenditure and spatial interactions in a decentralized national health system. Health economics, 16(3):291–306 (2007)
5. Costa-Font, J., Rico, A.: Devolution and the interregional inequalities in health and healthcare in spain. Regional studies, 40(8):875–887 (2006)
6. Eibich, P., Ziebarth, N.R.: Analyzing regional variation in health care utilization using (rich) household microdata. Health Policy, 114(1):41–53 (2014)
7. Filippini, M., Masiero, G., Moschetti, K.: Socioeconomic determinants of regional differences in outpatient antibiotic consumption: evidence from switzerland. Health policy, 78(1):77–92 (2006)
8. Finkelstein, A., Gentzkow, M., Williams, H.: Sources of geographic variation in health care: Evidence from patient migration. The quarterly journal of economics, 131(4):1681–1726 (2016)
9. Gabriele, S., Cislaghi, C., Costantini, F., Innocenti, F., Lepore, V., Tediosi, F., Valerio, M., Zocchetti, C.: Demographic factors and health expenditure profiles by age: the case of Italy. CEPS (2006).
10. Geue, C., Lorgelly, P., Lewsey, J., Hart, C., Briggs, A.: Hospital expenditure at the end-of-life: what are the impacts of health status and health risks? Plos one, 10(3):e0119035 (2015)
11. Göpffarth, D., Kopetsch, T., Schmitz, H.: Determinants of regional variation in health expenditures in germany. Health economics, 25(7):801–815 (2016)
12. Hogan, S., Lise, J.: Life expectancy, health expectancy, and the life cycle. Horizons, 6(2):17–20 (2003)
13. Moscone, F., Knapp, M., Tosetti, E.: Mental health expenditure in england: a spatial panel approach. Journal of Health Economics, 26(4):842–864 (2007)
14. Moscone, F., Skinner, J., Tosetti, E., Yasaitis, L.: The association between medical care utilization and health outcomes: A spatial analysis. Regional Science and Urban Economics, 77:306–314 (2019)
15. Srivastava, D.: Geographic variations in health care: what do we know and what can be done to improve health system performance? OECD Publishing (2014)
16. Zweifel, P., Felder, S., Meiers, M.: Ageing of population and health care expenditure: a red herring? Health economics, 8(6):485–496 (1999)

Measuring sustainable development at the regional level. The case of Italy

Misurare lo sviluppo sostenibile a livello regionale. Il caso Italia

Marianna Bartiromo, Enrico Ivaldi[1]

Abstract The first objective of this work is to provide an innovative tool for sustainable development through the construction of a composite index. The second objective is to quantify the level of sustainable development achieved by Italian regions. Moreover, once the composite index has been constructed, the relationship between regional sustainable development and gross domestic product and the Gini index will be tested. The index of sustainable development will be constructed through a series of indicators made freely available by ISTAT through the method of Pena's Distance (DP2). The results of this study confirm the strong territorial gap present within the Italian Peninsula.

Abstract *Gli obiettivi di questo lavoro sono quello di fornire una chiave di lettura innovativa sullo sviluppo sostenibile attraverso la costruzione di un indice composito. Il secondo obiettivo è quello di quantificare il livello di sviluppo sostenibile raggiunto dalle Regioni italiane. Inoltre, una volta costruito l'indice composito si andrà a testare il rapporto tra lo sviluppo sostenibile regionale e il prodotto interno lordo e l'indice di Gini. L'indice di sviluppo sostenibile verrà costruito attraverso una serie di indicatori resi disponibili gratuitamente dall'ISTAT attraverso il metodo di Pena's Distance (DP2). I risultati di questo studio confermano il forte divario territoriale presente all'interno della Penisola italiana.*

Key words: Sustainable development, DP2, index

[1] Channel & Retail Lab, SDA Bocconi School of Management, Milan, Italy

e-mail: marianna.bartiromo@sdabocconi.it

University of Genova, Department of Political Science:, enrico.ivaldi@unige.it

1 Introduction

Sustainable development represents one of the greatest challenges posed by globalization. Sustainable development consists in reconciling economic and social development with environmental protection, social equity and the rights of future generations. This orientation, expressed in the Bruntland Report (WCED, 1987), has the merit of having clarified the aspects on which global and national policies must focus. In particular, it highlighted the close relationship between economic, social and environmental development, proposing a holistic approach in which green sustainability is given central importance. In addition, the Report introduces the long-term perspective, the criterion of equity and inter- and intra-generational justice and, finally, the principle of efficiency in the use of natural resources. These principles have been translated into the rule of the balance of the three Es: Economic, Equity and Environment. Economic E is the ability to increase the gross domestic product per capita and improve competitiveness without affecting income and employment. Equity E, on the other hand, is the ability to increase human well-being in terms of security, nutrition, education and respect for fundamental individual freedoms. Finally, Environment E is the ability to provide resources, assimilate negative externalities and provide utility.
The objectives of this work can be outlined as follows:

1. to provide a new key to understanding sustainable development;
2. to quantify the level of sustainable development achieved by the various Italian regions;
3. to study the relationship between sustainable development, per capita GDP and the Gini index.

2 Material and Methods

Sustainable development, therefore, is a complex and multidimensional phenomenon. For this reason it has been necessary to use indicators capable of reading this complexity. To construct an aggregate composite index, data from the National Institute of Statistics (ISTAT) are considered, taking regional SDG values as a spatial reference.
The construction of the sustainable development index can be schematized as follows:

1. Preliminary survey of data availability and elimination of indicators that were clearly incomplete or manifestly unreliable. thanks to this phase, 98 indicators were selected that should ensure sufficient completeness of information (OECD, 2008);
2. a principal component analysis was then conducted using Varimax rotation (Kaiser,1958; Stevens, 1986). An explained variance level of 70% is considered significant (Pituch and Stevens, 2016).

3. Once the indicators had been chosen through Principal Component Analysis, it was possible to proceed to the elaboration of the index and then pursue the phases of normalization and weighting, and aggregation (OECD, 2008). The elaboration of the index was carried out using a non-compensatory index - the DP2 Distance - which is a parametric index since its formulation is based on the application of the linear regression model (Pena, 1977; Somarriba and Pena, 2009; Penco et al, 2020).
   The DP2 index provides the distance of each region from a reference base, which corresponds to the theoretical area that reaches the lowest value of the studied indicators, and it is defined for area j as follows:
   $DP2_j = \sum_{i=1}^{n} \left\{ \left(\frac{d_{ij}}{\sigma_i} \right) \left(1 - R^2_{i,i-1,i-2,\dots,1} \right) \right\}$ with i=1,…, n (area) e j=1,2,….,m (indicators)
   $d_{ij} = |x_{ij} - x^*_{ij}|$ = the difference between the value taken by the i-th indicator in area j and the minimum of the indicator in the least desirable theoretical scenario, i.e. the reference value of the X matrix.
   σ_I is the standard deviation of indicator I;
   $R^2_{i,i-1,i-2,\dots,1}$ is the squared multiple linear correlation coefficient in the linear regression of X_i on X_{i-1}, X_{i-2}, …, X_1 and indicates the part of the variance of X_i explained linearly by the X_{i-1}, X_{i-2}, …, X_1 indicators. This coefficient is an abstract number and is independent of the units of measurement of the various indicators.

The composite index obtained by applying the DP2 Index represents sustainable development at the Italian regional level. Higher scores represent situations of good "sustainable development" index compared to lower scores.

3 Results

The results obtained from the application of the method previously illustrated are as follows.
Overall, Emilia-Romagna (9.90) is the Region which has achieved the highest level of sustainable development in Italy. This is followed by Tuscany (9.56), Trentino-Alto Adige (9.14), Piedmont (9.13) and Umbria (8.50). Overall, the regions lagging furthest behind are geographically located in the South and they are: Basilicata (4.81), Campania (4.64), Puglia (4.06), Calabria (3.29) and Sicily (1.91). The results, therefore, converge towards a clear prevalence of central and northern Italy over the South in terms of achieving the best levels of sustainable development (Figure 1).

Figure 1: Results map for Sustainable development index

Emilia-Romagna	9.9
Tuscany	9.56
Trentino-Alto Adige	9.14
Piedmont	9.13
Umbria	8.5
Lombardy	8.48
Liguria	8.43
Marche	8.33
Friuli-Venezia Giulia	7.83
Abruzzo	7.76
Veneto	7.53
Lazio	7.43
Aosta Valley	7.41
Molise	6.78
Sardinia	5.34
Basilicata	4.81
Campania	4.64
Apulia	4.06
Calabria	3.29
Sicily	1.91

Once the sustainable development index was obtained and the results commented on, it was decided to test the link between the composite index, the regional Gross Domestic Product per capita (GDP per capita) and the Gini index to analyse the relationship between these variables.

In this work it was decided to use Kendall's tau-b coefficient, (Kendall, 1962). To better analyze the link between sustainable development, GDP per capita and the Gini index, it was decided to calculate the delta of Somers to understand the causal relationship between the variables. In fact, Somers' delta is a non-parametric index which allows to evaluate the asymmetric association between two ordinal variables (Somers, 1962). Between the independent variable of sustainable development and the dependent variable of GDP there is a positive concordance (0.547). This means that GDP positively influences levels of sustainable development. Regarding the relationship between the sustainable development index and GDP per capita ($\tau b = 0.550$, $p = 0.001$), it is observed that the levels of GDP per capita influence the levels of sustainable development. This means that regions that possess higher levels of GDP per capita also have high levels of sustainable development. Therefore, it can be understood how higher levels of wellbeing correspond to greater attention to sustainable development issues. On the other hand, as far as the Gini index is concerned, there is a negative concordance (-0.342), which means that the levels of sustainable development influence the levels of this indicator. ($\tau b = -0.349$, $p = 0.034$). This result shows how sustainable development is a solution in the fight against inequality: high levels of sustainable development, in fact, correspond to lower levels of inequality.

4 Discussion and Conclusion

To bridge the gap between North and South, the Ministry for the South and Territorial Cohesion adopted the so-called "Plan for the South 2030" at the beginning of 2020. This plan, in addition to reducing the gaps in the Italian peninsula between citizens and territories, wants to relaunch Italy in Europe, as the continuous and constant disinvestment in the South has done nothing but push the country back in the European Union rankings. The South 2030 Plan is structured around five missions: a) a South for young people; b) a connected and inclusive South; c) a South for the ecological turnaround; d) a South at the frontier of innovation; e) a South open to the world in the Mediterranean. As can be seen, these objectives are perfectly in line with both the goals of the 2030 Agenda and those of the European Green Deal. The year after the Plan for the South 2030 entered into force, the Italian government extended a new plan: the National Recovery and Resilience Plan (PNRR)(2021). The PNRR, as reported on the website of the Ministry of Economy and Finance "is part of the Next Generation EU (NGEU) program, the 750-billion-euro package, about half of which consists of grants, agreed by the European Union in response to the pandemic crisis. The total funds provided for the NGEU are around 248 billion euros, of which 26 billion are earmarked for the implementation of specific works and for the replenishment of resources from the Development and Cohesion Fund".

This measure aims to repair the economic and, above all, social damage caused by the pandemic crisis by helping to resolve Italy's structural weaknesses - including the strong territorial gap between North and South - and accompanying the country towards an ecological and environmental transition.

This study provides some relevant findings that can be summarized in the following propositions: first, a deep divide persists between the North-Central and the Southern regions of Italy. Southern regions tend to cluster systematically at the bottom of the rankings, unlike those in the North; second, regions where levels of sustainable development are higher show higher levels of GDP per capita and lower inequality; finally, lower levels of GDP per capita and higher levels of inequality may be obstacles to achieving higher levels of sustainable development.

The strength of this work lies in having created a composite index of sustainable development useful for measuring the levels of this phenomenon at the regional level. In fact, the complexity and multidimensionality of this phenomenon makes it very difficult and almost impossible to measure it (Gidding et al, 2002) and for this reason there are still few studies focused on the analysis of sustainable development at the regional level (Niemeijer, 2002; Salvati and Carlucci, 2014). In addition, the chosen method - the DP2 Index - allows the comparison of data over time. Therefore, it will be very interesting to repeat this study once the effects of the Plan for the South 2030 and the PNRR are visible so as to analyze regional differences as a result of the application of these policies. In terms of the critical aspects of this work, a major limitation is that of subjectivity in the choice of dimensions analyzed. However, sustainable development indices "remain invaluable in terms of their

ability to simplify complex measurement constructs, focus attention, and capture attention" (Booysen, 2002).
Finally, the results of this study only confirm the complexity of the sustainable development phenomenon and underscore the need for intervention especially in all those contexts with low levels of well-being and high levels of inequality.

References

1. Booysen, F. (2002). An Overview and Evaluation of Composite Indices of Development. Social Indicators Research 59, 115–151. https://doi.org/10.1023/A:1016275505152
2. Giddings, B., Hopwood, B., & O'Brien, G. (2002). Environment, economy and society: fitting them together into sustainable development. Sustainable Development, 10(4), 187–196. https://doi.org/10.1002/sd.199
3. Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, Vol. 23, pp. 187–200.
4. Kendall, M.G. 1962. Rank Correlation Methods, New York: Hafner
5. Niemeijer, D. (2002). Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. Environmental Science & Policy, 5(2), 91–103. https://doi.org/10.1016/s1462-9011(02)00026-6
6. Organisation for Economic Co-Operation and Develop (OECD). 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide (Science et Technologies de L'information). Pap/Ado ed. Paris: OECD Publishing.
7. Pena, J. (1977). Problemas de la mediciòn del bienestar y conceptos afines (Una aplicacion al caso espanol), INE, Madrid.
8. Penco, L., Ivaldi, E., Bruzzi, C., Musso, E. (2020). Knowledge-based urban environments and entrepreneurship: Inside EU cities. Cities, Vol. 96, 102443, doi:10.1016/j.cities.2019.102443.
9. Pituch, K.A., and Stevens, J.P. (2016). Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS, Sixth Edition. New York, Routledge.
10. Salvati, L., and Carlucci, M. (2014). A composite index of sustainable development at the local scale: Italy as a case study. Ecological Indicators, Vol. 43, pp. 162-171.
11. Somarriba, N., and Pena, B. (2009). Synthetic Indicators of Quality of Life in Europe. *Social Indicators Research*, Vol. 94, pp. 115-133.
12. Somarriba, N., Zarzosa, P., Pena B. (2015). The economic crisis and its effects on the quality of life in the European Union. *Social Indicators Research*, Vol. 120, pp. 323-343.
13. Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, Vol. 27, pp. 799-811*
14. Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale.
15. WCED (1987). Our Common Future. London: Oxford University Press.

Socio-economic deprivation and COVID-19 infection: a Bayesian spatial modelling approach

Deprivazione socio-economica e infezione da COVID-19: un approccio spaziale Bayesiano

Antonino Abbruzzo, Andrea Mattaliano, Alessandro Arrigo, Salvatore Scondotto, Mauro Ferrante

Abstract This paper aims to analyse the effect of socio-economic deprivation on COVID-19 incidence at the sub-urban level. Given the availability of information on monthly incidence rates from COVID-19 at census tract level for the two municipalities of Palermo and Catania (Italy), a Bayesian spatial model with zero-inflated binomial distribution is proposed. Results show an association between deprivation levels and incidence from COVID-19 in the two municipalities, also by controlling for the spatial structure of the territorial units. In the light of the results, health policy actions are needed, focusing interventions on the most deprived population groups.

Abstract *Il presente articolo ha l'obiettivo di analizzare l'effetto della deprivazione socio-economica sull'incidenza da COVID-19 a livello sub-comunale. Grazie alla disponibilità di informazioni sui tassi di incidenza mensili da COVID-19 a livello di sezione di censimento per i due comuni di Palermo e Catania (Italia), viene proposto l'utilizzo di un modello spaziale Bayesiano con distribuzione binomiale zero-inflated. I risultati mostrano un'associazione tra livelli di deprivazione e incidenza da COVID-19 nei due comuni, controllando per la struttura spaziale delle unità areali considerate. Alla luce dei risultati, si rendono necessarie azioni di politica sanitaria focalizzando gli interventi su segmenti di popolazione maggiormente a rischio.*

Key-words: COVID-19, Socio-economic inequalities, Bayesian Inference, Laplace approximation, Spatial-temporal models

---

A. Abbruzzo (*corresponding author*)
Department of Economics, Business and Statistics, University of Palermo e-mail: antonino.abbruzzo@unipa.it

A. Mattaliano
Department of Engeneering, University of Palermo e-mail: andrea.mattaliano@unipa.it

A. Arrigo; S. Scondotto
Dipartimento Attività Sanitarie e Osservatorio Epidemiologico Regione Siciliana e-mail: alessandro.arrigo.stat@gmail.com; salvatore.scondotto@regione.sicilia.it

M. Ferrante
Department of Culture and Society, University of Palermo e-mail: mauro.ferrante@unipa.it

1 Introduction

The spread of SARS-CoV-2 has strongly impacted urban centres characterised by high population density. Consequently, pandemic containment policies have focused on the closure of some economic activities described by a higher risk of contagion and on interventions aimed at limiting mobility, by taking into account the spread of the infection at the regional and the sub-regional level. Indeed, recent studies conducted in the USA [1] and South America [2] have shown an association between the risk of COVID-19 infection and socio-economic conditions.

This work aims to analyse the relationship between socio-economic deprivation and incidence of COVID-19 cases in the two Sicilian municipalities of Palermo and Catania (Italy), at the intra-municipality level, by considering the spatial structure of the territorial units under analysis. To this aim, a hierarchical Bayesian model with zero-inflated binomial distribution for the number of cases in each area is proposed by using the integrated nested Laplace approximation (INLA).

2 Material and Methods

Aggregate monthly data on incident cases at the census tract level for the two major Sicilian municipalities of Palermo and Catania, from October 2020 to April 2021, has been made available by the Regional Epidemiological Observatory. The data include a variety of demographic, comorbidity-related, test-related and disease-specific variables that track the case's progression until its recovery or death. In this paper, we focus on the effect of socio-economic deprivation on COVID-19 incidence as measured by the deprivation index. This index is a composite indicator that considers five dimensions of deprivation, namely: the share of the population with at least primary level education, the proportion of unemployed people, the share of families with one parent and dependent offspring living together, the proportion of rented households, and the number of people by 100 m^2 at the census tract territorial level [3]. A categorisation of the deprivation index based on quantiles of the regional population has been used in this study. The collected data can be described as areal spatial data, where Y_i is a random aggregate value over an areal unit i with well-defined boundaries in D_i, defined as a countable collection of n-dimensional spatial units. Therefore, the data have a specific spatial structure that needs to be considered in the inferential process. Generally, spatial models are computationally demanding, given the complexity of the parameter space. The Integrated Nested Laplace Approximation (INLA) proposed by [7] is a computationally efficient algorithm designed for latent Gaussian models, which incorporates a vast and flexible class of models ranging from generalized linear mixed to spatial and spatio-temporal models. Example of application on disease mapping includes [4, 5, 6].

Let's consider a region of interest divided into n non-overlapping areas. Let $Y_i, i = 1, \ldots, n$, denote the number of cases in region i and assume that its distribution belongs to the exponential family, in which the linear predictor η_i can include terms

of covariates in an additive way,

$$\eta_i = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(w_{ji}), \tag{1}$$

where β_0 is the overall intercept, $\mathbf{z} = (z_1, \ldots, z_{n\beta})$ are fixed covariates with linear effects $\{\beta_k\}$. The terms $\{f^{(j)}\}$ are used to represent specific Gaussian process. These latter terms can be accounted for by several statistical approaches, such as random (*iid*) effects, spatially or temporally correlated effects, smoothing splines, by varying the functional form of the $f(\cdot)$s. We consider the Besag-York-Mollie model to keep into account the spatial dependence.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{1+n_\beta+n_f}) = (\boldsymbol{\beta}, \mathbf{f})$ denote the vector of the model parameters, called latent field, and assume that the multivariate distribution belongs to a Gaussian Markov Random Field (GMRF), with zero mean and with a sparse precision matrix $Q(\boldsymbol{\psi})$, where $\boldsymbol{\psi}$ denotes a vector of hyperparameters, which are not necessarily Gaussian. A GMRF is simply a Gaussian with additional, conditional independence properties, meaning that x_i and x_j are conditionally independent given the remaining elements $x_{\{-ij\}}$ for quite a few $\{i, j\}$. In the Gaussian case, a practical consequence of conditional independence is that this results in zeros for pairs of conditional independence in the precision matrix. The main goal of the INLA-methodology is to obtain the marginal distributions for the elements of the latent field $p(\boldsymbol{\theta}|\mathbf{y})$ and the hyperparameters $p(\boldsymbol{\psi}|\mathbf{y})$, and use these to compute posterior summary statistics. This goal is achieved by exploiting the computational properties of the GMRF and the Laplace approximation for multidimensional integration, assuming the conditional independence of the observed variables $\mathbf{y}$ given the latent field $\boldsymbol{\theta}$ and the hyperparameters $\boldsymbol{\psi}$. For further details see [7, 8].

Besag–York–Mollié (BYM). The BYM is a discrete domain model used for describing count data per spatial unit [9]. The main idea is to borrow strength over neighbouring regions to get more reliable region-specific estimates. Due to inherent sampling variability, it is not recommended to inspect crude rates directly. We formulate a zero-inflated binomial model for the random variable Y_i, representing the number of cases in the region $D_i, i = 1, \ldots, n$. Moreover, the spatial component is assumed to follow a BYM distribution such that

$$Y_i \sim ZIBinom(\pi_i, N_i), \quad \text{logit}(\pi_i) = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + u_i, \quad u_i \sim BYM(\sigma_v^2, \sigma_\delta^2),$$

where N_i is the corresponding area population for the i-th spatial unit, z_{ki} are region level covariates (e.g. deprivation score, month). The random spatial process u_i is the sum of two independent Gaussian processes δ_i with variance σ_δ^2 and v_i with variance σ_v^2. In particular, for each region, the value of the v_i component depends on the average from the neighbouring regions

$$v_i | v_{\{-i\}}, \sigma_v^2 \sim N\Big(\sum_{j \in d_i} v_j / d_i, \sigma_v^2 / d_i \Big),$$

where d_i is the number of areas which share boundaries with the i-th one. Referring to the notation in Equation (1), $v_i = f^{(1)}(i)$ and $\delta_i = f^{(2)}(i)$ are two area-specific effects. The joint distribution for $\mathbf{v} = (v_1, \ldots, v_n)$ is a GMRF. Furthermore, the parameter δ_i represents the unstructured residual, modelled as

$$\delta_i | \sigma_v^2 \sim N(0, \sigma_\delta^2).$$

Inference for the latent field represented by $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{u}]^T$ and the hyperparameters $\boldsymbol{\psi} = [\sigma_v^2, \sigma_v^2]^T$ can be carried out with INLA (see [8, Chapter 6] for further details).

3 Results

During the considered period, a total of 29492 cases in 1952 census tracts for the city of Palermo (total population 733780) and 11163 cases in 1484 census tracts in Catania (total population 329746) were observed. Due to lack of space, in Figures, 1 and 2 only maps for the deprivation index, and the percentage of cases in Palermo municipality are reported, respectively. Deprivation areas appear in peripheral areas, and a similar spatial distribution for the percentage of COVID-19 cases appears. After model estimation, performed by using the R-INLA package [7], in Table 1 the mean, standard deviation and quantiles (0.025,0.975) for the parameters β_j related to the linear predictor $\eta_i = \text{logit}(\pi_i)$, for the two municipalities of Palermo and Catania are reported, where π_i represents the probability of being a case in the area i. For both municipalities under analysis, a relationship between deprivation and risk of COVID-19 infection appears, with higher values of risk for areas with higher deprivation levels compared to less deprived areas. However, its effect seems to be much stronger in the case of Palermo municipality compared to Catania. By looking at the temporal effect, the risk of infection is, as expected, consistently higher in all the months, compared to the reference period (October 2020), with peak values observed in January and April 2021 for Palermo municipality and November 2020 and January 2021 in the case of Catania.

4 Conclusion

It is acknowledged that socioeconomic conditions strongly affects various health outcomes. This paper proposes an analysis of its effect on the risk of contagion by COVID-19 at the sub-urban level in two Sicilian municipalities. From a methodological perspective, the proposed approach allows for considering the spatial structure of the territorial units under analysis by using a computationally efficient algorithm. Similar to previous studies [1, 2], the results showed a positive association between socioeconomic deprivation, as measured by the deprivation index, and risk of infection by COVID-19. Nonetheless, the dynamic of such an association is

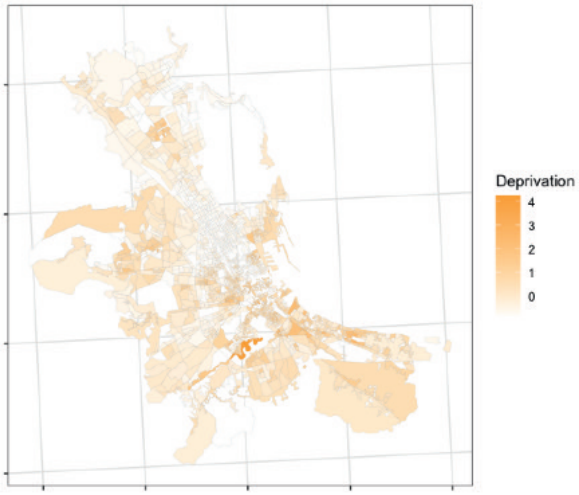


Fig. 1 Spatial distribution of deprivation index at the census tract level in Palermo municipality

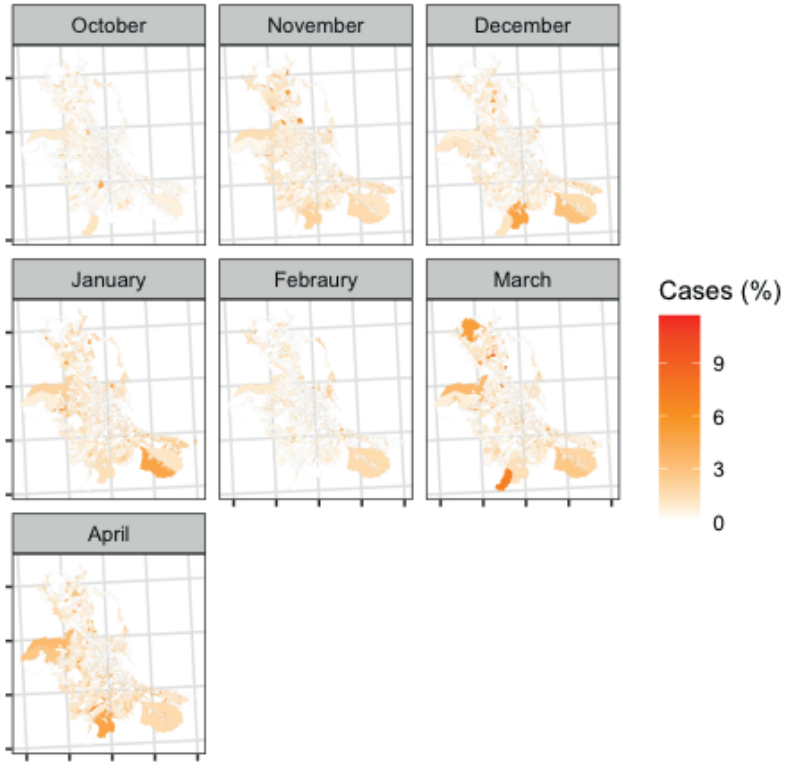


Fig. 2 Monthly Cases of COVID-19 as % of population by census tract of residence in Palermo municipality, October 2020 – April 2021.

Table 1 Posterior summary statistics of hierarchical Bayesian model for the city of Palermo and Catania

		Palermo				Catania			
		Mean	sd	Quantiles		Mean	sd	Quantiles	
				0.025	0.975			0.025	0.975
	Intercept	-5.90	0.04	-5.98	-5.82	-5.51	0.05	-5.60	-5.41
Deprivation	Middle	0.18	0.04	0.09	0.26	0.18	0.05	0.09	0.28
	High	0.36	0.05	0.26	0.45	0.13	0.06	0.01	0.25
Months	November	0.82	0.03	0.76	0.89	0.70	0.05	0.61	0.79
	December	0.67	0.03	0.60	0.73	0.48	0.05	0.38	0.58
	January	1.04	0.03	0.98	1.11	0.66	0.05	0.56	0.75
	February	0.48	0.04	0.41	0.55	-0.04	0.06	-0.16	0.08
	March	0.94	0.03	0.88	1.01	0.24	0.06	0.13	0.35
	April	1.11	0.03	1.05	1.17	0.34	0.05	0.24	0.44
Zero-probability		0.44				0.57			
Precision (iid component)		3.423				8.14			
Precision (spatial component)		27.977				18.30			

less clear. Lower cultural contexts may influence self-protection behaviour. Worst housing and working conditions may lead to overcrowded households and working environments where transmission is more likely, and suggested self-isolation procedures are challenging to follow. Consequently, health policy should consider the socioeconomic characteristics of the resident population and the area of residence in the development of infection control measures.

References

1. Hawkins, R. B., Charles, E. J., Mehaffey, J. H.: Socio-economic status and COVID-19–related cases and fatalities. Public health **189**, 129-134 (2020).
2. Mena, G. E., Martinez, P. P., Mahmud, A. S., Marquet, P. A., Buckee, C. O., Santillana, M.:Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. Science, **372**(6545), eabg5298 (2021).
3. Caranci, N., Biggeri, A., Grisotto, L., Pacelli, B., Spadea, T., Costa, G.: The Italian deprivation index at census block level: definition, description and association with general mortality. Epidemiologia e prevenzione, **34**(4), 167-176 (2010).
4. D'Angelo, N., Abbruzzo, A., Adelfio, G.: Spatio-Temporal Spread Pattern of COVID-19 in Italy. Mathematics **9**(19), 2454 (2021).
5. Goicoa, T., Ugarte, M.D., Etxeberria, J., Militino, A.F.: Age–space–time CAR models in Bayesian disease mapping. Statistics in medicine **35**(14), 2391–2405 (2016).
6. Riebler, A., Sørbye, S.H., Simpson, D., Rue, H.: An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Statistical methods in medical research **25**(4), 1145–1165 (2016).
7. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations: A national population-based study. Journal of the royal statistical society: Series b (statistical methodology) **71**(2), 319–392 (2009).
8. Blangiardo, M., Cameletti, M., Baio, G., Rue, H.: Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology **4**, 33–49 (2013).
9. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics **43**, 1–20 (1991).

Applications in Economics

The measurement of Economic Security through relative indicators

La misura della Sicurezza Economica attraverso indicatori relativi

Alessandro Gallo, Silvia Pacei and Maria Rosaria Ferrante

Abstract Economic Security is a topic which focused the attention of many researchers in the past years. With the COVID-19 pandemic, the interest has been grown drastically because economic security influences the lives of many individuals and consequently the government's economic and political choices. In this framework the individual insecurity indicator proposed by Bossert and D'Ambrosio [1] it's a landmark, because of its useful analytical properties. We embrace their notion of economic insecurity and we propose relative security indicators starting by the absolute one they suggested. We compare the indicators in terms of analytical properties and in terms of applicability to European countries using the EU-SILC (2019) data-set.

Abstract *La Sicurezza Economica è un fenomeno che ha attirato l'attenzione di molti ricercatori negli ultimi anni. L'interesse per questo tema è cresciuto molto con l'arrivo della pandemia del COVID-19, poiché la sicurezza economica ha un impatto sulla vita degli individui e di conseguenza sulle scelte economiche e politiche di governi. In questo ambito, l'indicatore assoluto d'insicurezza economica suggerito da Bossert e D'Ambrosio [1] è considerato un punto di riferimento, grazie alle sue numerose proprietà analitiche. Supportando la loro idea di base per misurare l'insicurezza, in questo lavoro proponiamo nuovi indicatori relativi. Gli indicatori sono poi confrontati in termini di proprietà analitiche soddisfatte e in termini di applicabilità ai dati EU-SILC (2019) di alcuni paesi europei.*

Key words: Economic Insecurity, Well-Being Indicators, Index numbers...

1 Introduction

In the existing literature, there are few examples of Economic Insecurity Indicators, such as Hacker et al. [3], Osberg and Sharpe [4], Bossert and D'Ambrosio [1] and

Alessandro Gallo, Silvia Pacei, Maria Rosaria Ferrante
University of Bologna, Statistical Sciences Departement, via Belle Arti,
e-mail: alessandro.gallo20@unibo.it, silvia.pacei@unibo.it, maria.ferrante@unibo.it

Bossert et al. [2]. While few measures have been proposed there is some debate on which index is more adequate to measure a multi-faced issue as economic insecurity. Economic insecurity may be seen as the perception of insecurity experienced by the individuals, measured through the confidence with which individuals face future potential economic changes. Bossert and D'Ambrosio [1] gave a fundamental contribution to the literature with their insecurity indicator, continuing with the collaboration of Clark and Lepinteur in [2]. Their indicator's latest version is based on a weighted summation of absolute changes between the individual resource levels observed in consecutive years [2].
Considering $\mathbb{R}^{(T)}$ as the $(T+1)$-dimensional Euclidean space with component labeled $(-T,...,0)$, where 0 is the current period and $-T$ is the furthest period in the past and acknowledging that $x=(x_{-T},...,x_0)\in\mathbb{R}^{(T)}$ is an individual income stream and that the economic insecurity is evaluated only considering x, the indicator proposed by Bossert et al. [2] is given by:

$$AI^T(x)=l_0\sum_{\substack{t\in\{1,...,T\}\\ x_{-t}>x_{-(t-1)}}}\delta^{t-1}(x_{-t}-x_{-(t-1)})+g_0\sum_{\substack{t\in\{1,...,T\}\\ x_{-t}<x_{-(t-1)}}}\delta^{t-1}(x_{-t}-x_{-(t-1)}) \quad (1)$$

where δ is the inter-temporal weight, l_0 and g_0 are weights assigned to losses and gains and $AI^{(T)}$ (Absolute Indicator) is a function $AI^{(T)}:\mathbb{R}^{(T)}\to\mathbb{R}$ for each $t\in\mathbb{N}$, and $AI^T(x)=\langle AI^{(T)}\rangle_{T\in\mathbb{N}}$ is a measure of individual economic insecurity.
With this work, we give our contribution by proposing relative versions of the individual economic insecurity indicator in (1). We rather suggest considering relative change between resources stock, instead of the absolute variation, as we think that individuals are more likely to judge changes in their economic status in relative terms, being also used to relative salary increases. Therefore we think relative changes may allow to capture the real perceived individual security better. Moreover, indices based on relative changes have the advantages of being independent of the order of magnitude of the phenomenon considered and of the units of measurement, thus favoring the comparison between individuals and countries adopting different currencies. Section 2 shows relative economic security indicators and the respective analytical properties, in Section 3 an empirical application to EU-SILC(2019) is presented and in Section 4 we carry out conclusions and suggest further applications.

2 Relative indicators

In this section, two different indicators are proposed and compared with the AI. The first one considers the relative pair differences with respect to the further period. Considering the same environment as before, but acknowledging that $x=(x_{-T},...,x_0)\in\mathbb{R}^{(T)}-\{0\}$, the first proposal is formulated as follows:

$$RI^T(x) = g_0 \sum_{\substack{t \in \{1,...,T\} \\ x_{-(t-1)} > x_{-t}}} \delta^{t-1} \frac{x_{-(t-1)} - x_{-t}}{|x_{-t}|} + l_0 \sum_{\substack{t \in \{1,...,T\} \\ x_{-(t-1)} < x_{-t}}} \delta^{t-1} \frac{x_{-(t-1)} - x_{-t}}{|x_{-t}|} \quad (2)$$

where δ is the inter-temporal weighting parameter, l_0 and g_0 are weights assigned respectively to losses and gains, $RI^{(T)}$ is a function $RI^{(T)} : \mathbb{R}^{(T)} - \{0\} \to \mathbb{R}$ for each $t \in \mathbb{N}$, therefore $RI^T(x) = \langle RI^{(T)} \rangle_{T \in \mathbb{N}}$ is a measure of individual economic security, from now RI (Relative Indicator).
The second indicator we propose is a Logarithmic Relative indicator (LRI), where the relative change is obtained by applying the logarithm, as suggested in Törnqvist et al. (1985) [5], to the ratio between resources levels in two subsequent points in time, taking advantage of the useful properties of the logarithm.
Considering the same setup as before, with $x = (x_{-T}, ..., x_0) \in \mathbb{R}^{(T)}_{++}$, the $LRI^{(T)}$is a function $LRI^{(T)} : \mathbb{R}^{(T)}_{++} \to \mathbb{R}$ for each $t \in \mathbb{N}$. The security indicator $LRI^T(x) = \langle LRI^{(T)} \rangle_{T \in \mathbb{N}}$ is formulated as follows:

$$LRI^T(x) = g_0 \sum_{\substack{t \in \{1,...,T\} \\ x_{-(t-1)} > x_{-t}}} \delta^{t-1} \ln\left(\frac{x_{-(t-1)}}{x_{-t}}\right) + l_0 \sum_{\substack{t \in \{1,...,T\} \\ x_{-(t-1)} < x_{-t}}} \delta^{t-1} \ln\left(\frac{x_{-(t-1)}}{x_{-t}}\right) \quad (3)$$

where δ is the inter-temporal weighting parameter, l_0 and g_0 are weights assigned respectively to losses and gain.
In this indicator the relative pair differences are expressed as logarithm of the ratio and the relative difference is considered with respect to the logarithmic mean $L(x_{-t}, x_{-(t-1)})$ [5]:

$$\ln\left(\frac{x_{-(t-1)}}{x_{-t}}\right) = \frac{x_{-(t-1)} - x_{-t}}{L(x_{-t}, x_{-(t-1)})} \quad (4)$$

where $L(x_{-t}, x_{-(t-1)})$ is defined as [5]:

$$L(x_{-t}, x_{-(t-1)}) = \begin{cases} \frac{x_{-(t-1)} - x_{-t}}{\ln(x_{-(t-1)}/x_{-t})} & \text{for } x_{-(t-1)} \neq x_{-t} \\ x_{-t} & \text{for } x_{-(t-1)} = x_{-t}, \end{cases} \quad (5)$$

2.1 Properties

It has been proved that the two relative indicators satisfy some important properties. They both fulfill some properties considered also in [2], in particular *Gain-loss monotonicity*, which ensures that a gain in the further period is related to a higher level of security compared to a loss that occurred in the same period; *Homogeneity*, which establishes that if each value of an individual stream is multiplied by the

same constant, the security score remains unchanged; *Quasilinearity*, assures that it is possible to establish a link between streams of different lengths, and *Stationarity*, an important property when dealing with indicators adopting the geometric discounting. Moreover, LRI satisfied further useful properties: *Proximity monotonicity*, important to order the different individual streams and *Simmetry* [5], according to which inverting the ratio is equivalent to change the sign of the indicator, and necessary to prove that the LRI satisfies the *Proximity monotonicity* (while RI does not).

3 Application to Eu-Silc Data

To highlight the advantages offered by the indicators proposed, we estimate security indicators based on information produced by the EU-SILC (2019) survey, a widely used database to study poverty, inequality and social exclusion in EU member countries. We consider four different panels: from 2012 to 2016, from 2013 to 2017, from 2014 to 2018 and from 2015 to 2019. The variable used in the analysis is the equivalent disposable income at an individual level. Income values have been discounted to remove inflation. Streams of four years with three delays are considered and economic security scores are computed for all three indicators in all panels. As regards

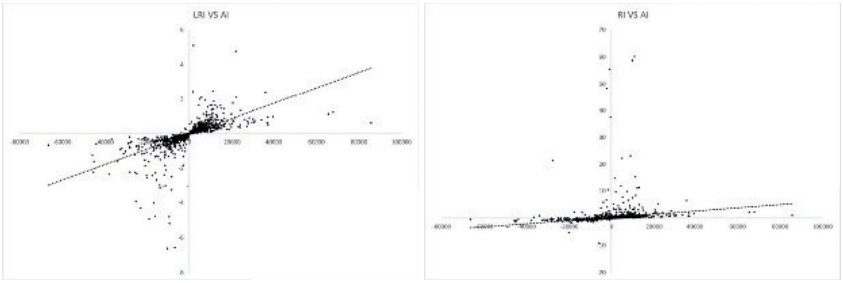


Fig. 1 Scatter-plots comparing relative indicators and the absolute one in 2019. On the left side the comparison among LRI and AI while on the right side among RI and AI. The X axis contains values from the AI and, the Y axis from the relative ones. In both graphs is shown the dotted interpolation line

the values used for the parameters included in the equations, we use the same values proposed in [2], therefore $\delta = 0.9, l_0 = 1$ and $g_0 = 15/16$. The AI is transformed into a security indicator by multiplying the scores times (-1), to carry out the comparison.

The correlation between LRI and AI appears high in the four periods considered (as an example the correlation coefficient is equal to 0.60 in 2019), while the correlation between RI and AI is much lower (0.04 in 2019). Scores obtained for the three indicators are compared in Figure 1. In the X-axis we put the AI scale, while in the

Y-axis the relative indicators ones. On the left figure, we see how dots allocate just on the first and the third quadrants, therefore individuals seen as insecure for the AI are classified in the same way from the LRI. On the other hand, dots look quite spread in the quadrants. This means that there are many cases where high values of security assigned by the AI coincide with much lower values of security according to the LRI and vice-versa. In the right figure, we see how the security allocation of the AI and the RI looks quite different. Dots are spread all over the X-axis in all four quadrants, confirming the lower correlation detected before.
To highlight the trend shown by the three indicators at an aggregate level, we calculate the weighted median of the individual scores. Figure 2 shows that results obtained from LRI are more coherent, at least on the median, with those obtained from AI. Hence LRI appears preferable to the RI as it satisfies more properties and is more coherent with AI, which has been already largely considered and applied in the literature.

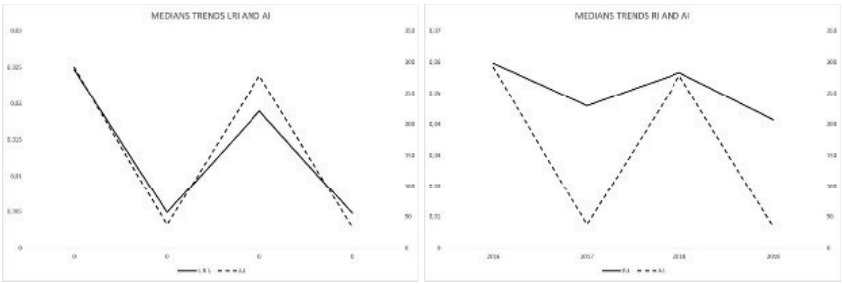


Fig. 2 Comparison between relative indicators weighted medians trends and the absolute indicator one.

Analyzing more in deep two indicators, LRI and AI, we notice that they lead to some differences in the "scores" calculated for the individuals in the sample. Taking individual scores from the year 2016 as an example, if we divide them into four quartile classes, we notice that 15% of the observations belong to different quartiles according to the two indicators. A sensible part of the observations belonging to different classes (45%) exchange from the third to the fourth class: differences regard above all high positive scores, that for the absolute indicator are close to median while for the relative indicator are close to the maximum and vice-versa. Finally, we use both LRI and AI to compare economic security in some European countries: Spain, Italy, Portugal, Greece, France, and the Netherlands. Figure 3 shows, as an example, results obtained from one synthetic indicator, the weighted median. The trends look comparable but, it's possible to notice how Spain and Portugal's trends in the AI diverge while in the LRI converge from 2016 to 2017. It is also notable how the countries' trends change rank order in two out of four years considered (2018 and 2019).

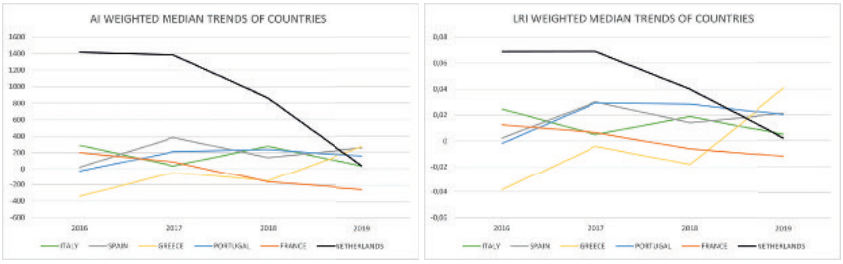


Fig. 3 Comparison between security indicators trends among different European countries.

4 Conclusions and further applications

In this work, two relative indicators of economic security are proposed and compared with the absolute one suggested by Bossert et al. [2]. Both are based on the use of relative change: in the RI it's applied by dividing the intertemporal difference of resource levels by the further period, while in the LRI is implemented by applying the logarithm to the ratio between resources levels in two subsequent points in time. Among these two relative indicators, the logarithmic version appears preferable: in terms of properties satisfied and in terms of consistency with the absolute indicators. Nevertheless, the logarithmic relative indicator rightly provides different results from the absolute indicator in some important cases, where considering a relative change amplifies or reduces the importance of the variation compared to the absolute change. The indicator suggested may be used to study the relationship between economic insecurity and other phenomena such as mental health, criminality and fertility.

References

1. Bossert, W. and D'Ambrosio, C. :Economic insecurity and variations in resources. International Economic Review. **54**, 1017–1030 (2013)
2. Bossert, W. and Clark, A.E. and d'Ambrosio, C. and Lepinteur, A.C. :Economic Insecurity and the Rise of the Right. In: CEP Discussion Papers, LSE (2019). http://eprints.lse.ac.uk/103442/1/dp1659.pdf
3. Hacker, J. S. and Huber, G. A. and Rehm, P. and Schlesinger, M. and Valletta, R.: The Economic Security Index: A New Measure for Research and Policy Analysis. Review of Income and Wealth. **60**, S5-S32 (2014)
4. Osberg, I. and Sharpe, A.:New estimates of the index of economic well-being for selected OECD countries, 1980-2007. Centre for the Study of Living Standards (CSLS) Research Report 2009-11 (2009)
5. Törnqvist, L. and Vartia, P. and Vartia, Y.O.: How should relative changes be measured? The American Statistician. **39**, 43–46 (1985)

A regional analysis of the efficiency by energy's producers in Italy

Un'analisi regionale dell'efficienza dei produttori di energia in Italia

Gianna Greca, Giuseppe Cinquegrana and Giovanni Fosco

Abstract The energy market in Italy is characterized by a competitive framework where regulatory Authority checks the conditions in which operate enterprises that produce, store, distribute and trade energy. The presence of government subsidies towards the energy's production Italian enterprises from two decades up now days in different ways (*Energy Account*, *Green Certificates*, *GRIN*) points out the issue if the efficiency of this industry was affected by these incentives. In this study we focus on the efficiency of the energy's production SMEs taking into account the territorial gap among the Italian Regions. An analysis on the impact of the subsidies to the energy's producers has been conducted regarding the efficiency of the industry on the basis of data collected (Source ISTAT, GSE).

Abstract Il mercato dell'energia in Italia è caratterizzato da un quadro competitivo in cui l'Autorità di regolamentazione verifica le condizioni in cui operano le imprese che producono, immagazzinano, trasmettono distribuiscono e commercializzano energia. La presenza di contributi pubblici alle imprese italiane di produzione di energia da due decenni ad oggi, con modalità diverse (Conto Energia, Certificati Verdi, GRIN), pone la questione se l'efficienza di questo settore sia stata influenzata da tali incentivi. In questo studio ci concentriamo sull'efficienza delle PMI produttrici di energia elettrica tenendo conto del divario territoriale tra le Regioni italiane. Sulla base dei dati raccolti (Fonte ISTAT, GSE) è stata condotta un'analisi sull'impatto degli incentivi ai produttori di energia sull'efficienza del settore.

Key words: energy, producer, market, subsidies.

[1] Gianna Greca, Istat; greca@istat.it:
Giuseppe Cinquegrana, Istat; gicinque@istat.it:
Giovanni Fosco, Università degli Studi di Salerno; gfosco@unisa.it

1 Introduction

The framework of the Italian energy market is based on a competitive dimension regulated by law and monitored by the Italian Regulatory Authority for Energy, Networks and Environment (ARERA). The regulation involves the production, storage, distribution and trade of energy and regards enterprises different by size and industry activity. A relevant characteristic of the energy's production market is due to the fact that in Italy from two decades up to now days a frequent flow of government subsidies has been provided to the energy's producers.

With the exception of a few relevant multinational companies, the vast majority of enterprises that produce electric energy are SMEs and in this work we focus on them, in particular we conduct an efficiency analysis of these companies. In paragraph 1.1 we report a picture of the government subsides regarding *Energy Account* in the last two decades. In paragraph 2. a detailed description of data available for the analysis of the energy's producers (SMEs) and a formulation of the model used for the estimates have been presented. Government subsidies to energy producers

1.1 Government subsidies to energy producers

For plants that generate electricity through the conversion of solar energy, a specific incentive system called *Energy Account*, EC, has been provided for a period of twenty years that rewards the energy produced by photovoltaic systems with incentive rates. This mechanism, already provided for by Legislative Decree 387/2003, became operational following the entry into force of the interministerial decrees of July 28, 2005 and February 6, 2006 (I EA). The activity carried out by the GSE (*Gestore dei Servizi Elettrici*) consisted in the management and examination of the documentation sent by the responsible parties, in the monitoring of deadlines related to the obligations required by law and in the administrative management of the incentives related to the energy produced by the plants. The incentive was ruled by other decrees and in the following table 1 we report the amounts disbursed on *Energy Account* by decree and by year starting from 2007 up to 2020.

Table 1: Amounts disbursed on *Energy Account*, EA, by decree from 2007 up to 2019-2020, mil. Euro.

Decree	2007	2019	2020
I EA (issued in 2003-2005-2006)	17	71	80
II EA (issued in 2007)	2	2.903	3.005
III EA (issued in 2010)	…	581	599
IV EA (issued in 2011)	…	2.181	2.275
V EA (issued in 2012)	…	208	227
Total	**19**	**5.945**	**6.197**

Source: GSE

An other relevant mechanism of energy incentives is GRIN (Incentive Recognition Management) that allows you to take advantage of the incentives provided for by the D.M. 06/07/2012 for all the plants powered by renewable sources qualified to be subsided by Green Certificates up to 2015. Since 2016, the green certificate mechanism has been replaced by this new form of incentive. In this work we focus on GRIN mechanism in order to evaluate the impact of these subsides on the energy sector.

2 Data, Methods and Results

In order to estimate the efficiency of the energy's producers we have built a balanced panel of balance sheet data of energy producers, in particular SME's, from 2014 up to 2020 (source ISTAT). We report in the table 2.1 the number of units per size according to the employees in the activity 35.11 (NACE REV 2).

Table 2.1 : Number of units of the energy's producers by size included in the panel 2014-2020.

Size by employee	Number of units
Up to 9	3.819
10-49	85
50-249	6
Total	**3.910**

A territorial detail is available in the data base and in table 2.2 we report the distribution of energy's producers among the Italian macro territorial divisions.

Table 2.2: Number of units of the energy's producers by size included in the panel 2014-2020.

Size by employee	Number of units	Share % on Italy
North-Est	1.522	38.93
North-West	1.144	29.26
Center	611	15.63
South	470	12.02
Inlands	163	4.17
Total	**3.910**	**100**

The opportunity to use the territorial FRAME SBS (Source Istat), an integrated system of administrative and statistical data by local unit of each enterprise, and ENERDATA database (Source Istat) on economic-legal units that receive subsidies and/or buy and sell allowances[1], provide the tools to territorially explore the performance of energy producers.

[1] The "ENERDATA" database was built as part of the Istat thematic research Project "I meccanismi di incentivazione energetico-ambientale (Energy-environmental incentive mechanisms)". The database, integrating Istat and public data, collects information on the economic-legal units that, in the role of producers and/or users of energy, receive subsidies and/or buy and sell allowances (CO_2 allowances, Energy Efficiency Certificates, etc.) (Cfr. Greca G., Istat, Progetto di ricerca tematica "I meccanismi di incentivazione energetico-ambientale" and Greca G. (2021)).

In the model used to estimate the efficiency of energy producers, we consider a single output (Q) and two inputs: *labor* (L) and *capital* (K). The *total cost* depends on two prices: the price of the *labor* (p_w) and the price of the *capital* (p_r, *user cost of capital*). In order to estimate the cost function Cobb-Douglas, CB, we set a linear homogeneity condition, so the total cost and the price of capital are normalized with the price of labor.
In the following equation, the model considered is reported:

$$\ln\left(\frac{TC}{p_e}\right)_{i,t} = \alpha_i + \gamma_t + \delta_c + \beta_Q \ln Q_{i,t} + \beta_K \ln\left(\frac{p_r}{p_e}\right)_{i,t} + \beta_L \ln\left(\frac{p_w}{p_e}\right)_{i,t} + \beta_1 t + \beta_2 t^2 + U_{i,t} + V_{i,t}$$

$$U_{i,t} = \varphi_1 + \varphi_1 t + \varphi_2 t^2 + Z_{i,t}$$

$$U_{i,t} = \varphi_1 + \varphi_1 post_{2016} + \varphi_2 treated_i + \underbrace{\varphi_3 treated_i \cdot post_{2016}}_{\text{Treatment Effect}} + Z_{i,t}$$

TC= Total costs
p_e= energy price (based on the deflator of value added in energy activity)
p_r= capital price (depreciation plus financial charges / capital
p_w= average wage per employee (wages/ employees)
Q= value added at chained value (2015)
$\alpha_i,\ \gamma_t,\ \delta_c$ fixed effects: firm, year and municipality
t= trend
U= inefficiency term

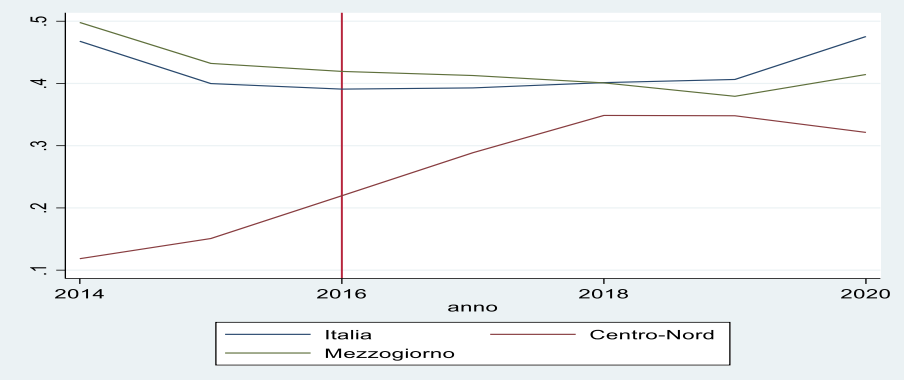
Table 2.3: Frontier Cost (panel data, years 2014-2020)

	(1) Italy	(2) Center-North	(3) South and Islands
Frontier			
y	0.96*** [0.00]	0.91*** [0.00]	0.95*** [0.01]
Pr	0.04*** [0.01]	0.04* [0.02]	0.10*** [0.03]
Pw	0.03*** [0.00]	0.04*** [0.00]	0.02*** [0.00]
t	0.13*** [0.02]	-0.05** [0.01]	0.09*** [0.03]
t^2	-0.02*** [0.00]	0.00 [0.00]	-0.01** [0.00]
Mu			
t	-2434.72** [799.23]	3.88e+08*** [0.00]	-1825.89 [4199.07]
t^2	310.72** [96.45]	-3.52e+07*** [0.00]	165.98 [380.08]
Lambda	243.70	19718.58	238.59

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: our elaborations
All the variables considered have been deflated on regional level by the deflator of value added at chained value (2015) of the energy activity (NACE REV2). The analysis has been conducted on Small enterprises with only 6 units have more than 50 employees on the total of 3910 firms considered. The results of the estimation of the economic efficiency of energy's producers for a balanced panel from 2014 to 2020 have been reported in the table 2.3, where the control variable is the municipality. The northern firms that operate in the energy sector are less inefficient of the ones that are located in the South of Italy; in particular the regional origin affect the efficiency of the capital: the coefficient β_K is significant both Italy and Center-North and South (and Islands) but for the southern firms signals more inefficiency where it is equal to 0.10 as regards to 0.04 of the northern ones. On the side of output the coefficient β_Q is 0.95 in South and Island while it is 0.91 for the northern units.
In the Figure 2.1 we report the cost efficiency by Italian macro regions from 2014 to 2020, where it's pointed out that the energy's producers of southern regions are largely less efficient than the northern ones, but the gap has started to reduce from 2016.

Figure 2.1 Cost efficiency by Italian macro regions (Years 2014-2020).



Source: our elaborations
In a second step we identify the firms that receive GRIN subsidies in order to evaluate the impact of these incentives on the efficiency of the energy's producers by region and consider these units subjected to treatment. In particular these treated enterprises are 347 in 2016, 373 in 2017, 358 in 2018, 334 in 2019 and 294 in 2020. In the table 2.4 we report for a balanced panel from 2014 to 2020 the estimate of frontier cost of energy's producers considering a treatment for the firms that receive GRIN subsidies using DIFinDIF method to estimated the GRIN's policy effect.
In the table above, we implemented the same frontier cost but a different inefficiency cost model, which use the diff-in-diff method to estimate the GRIN subsidies' effect.

$$U_{i,t} = \varphi_1 + \varphi_1 post_{2016} + \varphi_2 trattati_i + \varphi_3 trattati_i \cdot post_{2016} + Z_{i,t}$$

In this different formulation of cost inefficiency, $post_{2016}$ is an indicator that defines years after the 2016. Whereas, *trattati* is an indicator that defines who takes GRIN subsidies. Thus, the interaction between these two variable indicator gives us the policy effect estimate φ_3.
The analysis with treated firms that receive GRIN subsidies shows that the incentives received do not improve the efficiency of energy's producers, rather the inefficiency increases both in terms of coefficient β_Q and of the one β_K and the southern firms is always worse than the northern ones considering these parameters. The regional gap is also clear looking at the figure

2.2 where it is reported cost efficiency by Italian macro regions with treatment for firms with GRIN incentives (Years 2014-2020). The firms in South and Islands decrease efficiency from 2015 to 2019 while the northern ones improve in that period. In 2020 all the areas show a large rebound.

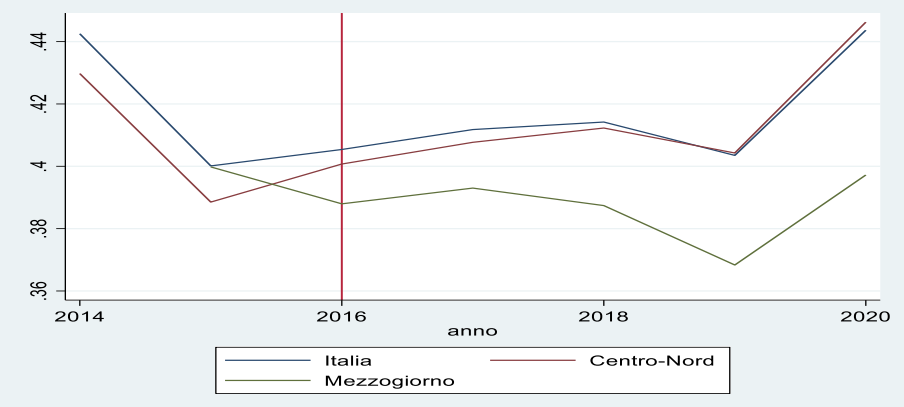
Table 2.4: Frontier Cost with treatment for firms with GRIN incentives (panel data, years 2014-2020)

	(1) Italy	(2) Center-North	(3) South and Islands
Frontier			
y	0.96***	0.96***	0.99***
	[0.00]	[0.00]	[0.01]
P_r	0.04***	0.03**	0.11***
	[0.01]	[0.01]	[0.03]
P_w	0.03***	0.04***	0.02***
	[0.00]	[0.00]	[0.00]
t	0.09***	0.09***	0.08
	[0.01]	[0.01]	[0.04]
t^2	-0.01***	-0.01***	-0.01*
	[0.00]	[0.00]	[0.00]
Mu			
post	-410.81	-148.02	-813.69
	[279.52]	[462.83]	[479.60]
treatment	-4799.30***	-9073.60***	-2451.59
	[1091.90]	[2108.50]	[2133.82]
Post*treatment	4318.72***	8355.83***	512.69
	[1134.26]	[2168.50]	[2189.47]
lambda	175.29	212.89	164.81

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: our elaborations

Figure 2.2 Cost efficiency by Italian macro regions with treatment for firms with GRIN incentives (Years 2014-2020).



Source: our elaborations

The results presented give some implications for the policy makers that decide to subside the energy's producers in order to check if this kind of incentives provide positive impact on the efficiency of the energy production sector. From Figure 2.1 we can see a convergence of the efficiency path between the Center-North and South (and Islands), but considering an analysis DIFinDIF with treatment of the firms that receive GRIN subsidies, we can conclude that this kind of incentives do not improve the efficiency of the energy's producers, anyway there is a differential impact on the Italian macro regions. In general the northern firms that operate in the energy sector are more efficient than the ones in the South and Islands as showed in Table 2.3 and in Figure 2.1, but the impact of GRIN subsidies show in Figure 2.2 a marked deterioration in the southern firms.

3 Citations and References

ANGRIST, J. D., and J-S. PISCHKE. 2009. Mostly Harmless Econometrics: An Empiricist Companion. Princeton: *Princeton University*

BATTESE G., COELLI T. (1992). Frontier Production Functions, Technical Efficiency and Panel Data with Applications to Paddy Farmers in India. *Journal of Productivity Analysis*, 3, 153-169

BATTESE G., COELLI T. (1995). A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data. *Empirical Economics*, 20, 325-332

BELOTTI F., DAIDONE S., ILARDI G., ATELLA V. (2012). Stochastic Frontier Analysis using STATA. *CEIS Tor Vergata, Research Paper Series*, 10 (12), n.251, Roma

BRIPI F., CARMIGNANI A., GIORDANO R. (2011). La qualità dei servizi pubblici in Italia. *Banca d'Italia Questioni di Economia e Finanza (Occasional papers) n. 22*

GRECA G. (2019). La filiera dell'energia elettrica: il valore aggiunto e le sue principali caratteristiche strutturali, "Relazione sulla situazione energetica nazionale del 2018", Ministero dello sviluppo economico, giugno 2019

Greca G. (2021). #Incentiviallerinnovabili #EfficienzaenergeticaeCO2 #Mercatodell'energiaecompetitività #ENERDATA", "I meccanismi di incentivazione energetico-ambientale", Poster scientifico, 14ª Conferenza nazionale di Statistica, 30 novembre-1 dicembre 2021

On Investigating Social and Financial Aspects of Cardano

Analisi degli Aspetti Sociali e Finanziari di Cardano

Stefano Vacca, Marco Ortu, Gianpaolo Zammarchi and Giuseppe Destefanis

Abstract The goal of this paper is to investigate the reasons and speculations surrounding the Cardano cryptocurrency (ADA) from two points of view: (I) opinion mining, analyzing the comments of expert users on the r/cardano subreddit; (II) financial, through the analysis and correlation of the price trend with the metrics extracted from Reddit. The main purpose of this study is to measure any cause-effect relationships between the social variables and the financial variables, in order to demonstrate whether it is possible to predict the price trend of the Cardano starting from the specialized discussions of the redditors, or if it is possible to predict the user behavior depending on the price trend in different periods.
L'obiettivo di questo lavoro è lo studio delle cause delle speculazioni che riguardano la criptovaluta Cardano (ADA), sotto due punti di vista: (I) l'opinion mining, analizzando i commenti di utenti esperti sul subreddit r/cardano; (II) finanziaria, attraverso l'analisi e la correlazione dell'andamento dei prezzi con le metriche estratte da Reddit; Lo scopo principale è quello di misurare eventuali relazioni di causa-effetto tra le variabili sociali e le variabili finanziarie, al fine di dimostrare se è possibile prevedere l'andamento dei prezzi del Cardano partendo dalle discussioni specialistiche dei redditors, oppure se è possibile prevedere il comportamento dell'utente in base all'andamento del prezzo in periodi diversi.

Key words: Blockchain, Cryptocurrencies, Hawkes Model, Causal Inference

---

Stefano Vacca
Eustema SpA, IT, e-mail: s.vacca@eustema.it

Marco Ortu
University Of Cagliari, IT, Dept. of Economics and Business Sciences e-mail: marco.ortu@unica.it

Gianpaolo Zammarchi
University Of Cagliari, IT, Dept. of Economics and Business Sciences e-mail: gianpaolo.zammarchi@unica.it

Giuseppe Destefanis
Brunel University, UK, Dept. of Computer Science e-mail: giuseppe.destefanis@brunel.ac.uk

1 Introduction And Methodology

The analysis of cryptocurrencies price behaviours is one of the most discussed topics of the last few years on social media environments. In this study we focused our attention on Cardano (ADA), which is a proof-of-stake blockchain platform that, since January 2021, has seen a price increase from 15 euro cents on December 31st 2020 to the current 1.93 euro on September 23rd 2021 (+ 1271.55%). However, after the absolute peak of 2.42 euros reached on September 2nd, ADA suffered a price collapse until this analysis. The purpose of this work is to ascertain the existence of possible relationships between cryptocurrency market prices and social media [6] discussions in order to understand what topics have the potential to predict price movements. The first step was to retrieve the discussions comments from the social media platform *Reddit*, which has been shown to be one of the most valuable sources of information relating to cryptocurrency markets. Secondly, the occurrence of particular topics from social media content were evaluated, using dynamic topic modelling, that is an extension of Latent Dirichlet Allocation (LDA). Finally, a Hawkes model was implemented to identify hidden interactions between topics and cryptocurrency market prices. This family of statistical models is often used in economics analysis [2]. *Reddit* is an American social news aggregation, web content rating, and discussion website that reaches about 8 billion page views per month. It is built over multiples subreddits, where each subreddit is dedicated to the discussion of a particular subject and there are specific subreddits related to major cryptocurrency projects. We used the Pushift API[1] to extract comments from posts written on *r/cardano*, a subreddit with over 619,000 subscribers. The period considered runs from 1st July to 23rd September 2021, for a total of 72,090 comments extracted. Thanks to Natural Language Processing, Sentiment Analysis and Topic Modeling techniques, we extracted the following metrics: i) 10 topics using Latent Dirichlet Allocation; ii) 6 up and down VAD metrics, of which: 2 Valence metrics, 2 Arousal metrics and 2 Dominance metrics; iii) 6 price metrics that count the number of low, high and stable events derived by the common financial metrics, closing price minus opening price in the minute, and high price minus low price in the minute. Valence, Arousal and Dominance (VAD) represent conceptualized affective dimensions that respectively describe the interest, alertness and control a subject feels in response to a certain stimulus. Warriner et al. [7] has created a reference lexicon containing 14,000 English words with VAD scores for Valence, Arousal, and Dominance.

Topic Modeling. A topic model is a specific statistical model used to identify the abstract topics within a collection of documents. Topic modelling is a frequently used text-mining tool for discovery themes occurring within a corpus automatically, finding a distribution of words in each topic and the distribution of topics in each document. In this work, we used the *Latent Dirichlet Allocation* (LDA) [4], which is a popular unsupervised learning technique for topic modelling, introduced in 2003. This type of topic model assumes that each document contains multiple topics to different extents. It has been proven that the standard LDA model can not understand both the ordering of words within a document and the ordering of documents within

[1] https://github.com/pushshift/api

a corpus. For this reason, in 2006, an extension of this model was developed [3]. This LDA model extension is known as *dynamic topic model* and even if it still has no understanding of the order of words in a document, at least the order of documents in the corpus is accounted for.

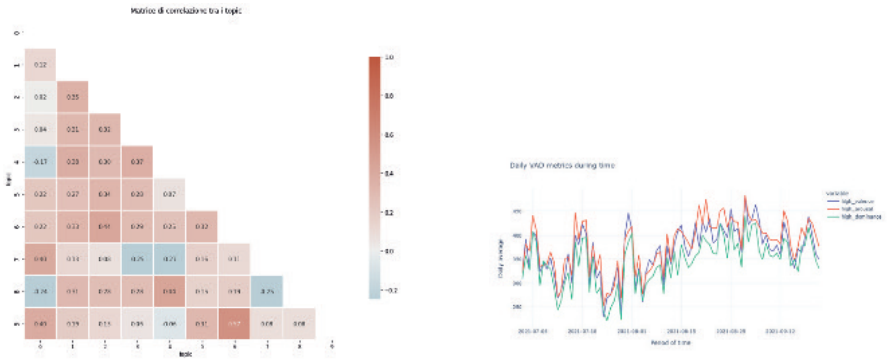
Topic	Keywords	Label	# comments
0	stake, pool, reward, ada, epoch	Staking & rewards	9932
1	question, answer, word, ask, dip	Uncertainty	4189
2	video, comment, link, watch, youtube	General Information	5030
3	fee, people, transaction, nice, invest	Trading and Smart Contracts	4404
4	project, cardano, people, think, know	Cardano Community Information	9036
5	ada, token, cardano, binance, nft	NFT & Investments	4017
6	buy, crypto, good, coin, ada	Pure Trading	8130
7	wallet, thank, yoroi, ledger, use	Stacking Wallets	9008
8	cardano, contract, smart, eth, ethereum	Smart Contracts & Crypto Market	12233
9	ada, year, pay, tax, hold	Crypto, Taxes & Law	6111

Table 1: Reddit Topics.

Table 1 shows the 10 topic extracted with the dynamic LDA along with the five top words, label and number of comments of each topic. Figure 1a shows the linear correlation between topics (that primarily indicates positive linear correlations, except for 6 out of 45 combinations). The highest correlation is between topics 6 (Pure Trading) and 9 (Crypto, taxes & law), with a positive correlation of 0.57. The highest negative correlation is instead between topic 4 (Cardano Community information) and topic 7 (Stacking wallets) with -0.27.
Figure 1b shows the daily trend of VAD metrics. The growing trend of all three metrics starting from August follows the increase in the ADA price. Financial data was extracted from Binance[2]. Each line of the dataset refers to the price trend with a frequency of 1 minute. We finally created six final metrics used in our model, they count the events of *low*, *high* and *stable* of price and volatility changes (delta of high minus low values).

Hawkes Model. The Hawkes process is a point process class [5], also known as a self-exciting counting process, in which the impulse response function explicitly depends on past events. In this type of process, the observation of an event causes the increase of the process impulse function. From a mathematical point of view, a point process is a Hawkes process if the impulse function $\lambda\left(t|H_t\right)$ of the process takes the form of Equation 1.

[2] https://www.binance.com/it/markets



(a) Correlation among topics

(b) VAD daily trend.

Fig. 1: Topics correlation and VAD metric trends.

$$\lambda\left(t|H_t\right) = \lambda_0(t) + \sum_{i:t_i<t} \phi\left(t - t_i\right) \tag{1}$$

In Equation 1, H_t represents the history of given past events, $\lambda_0(t)$ is a positive function that determines the basic intensity of the process and ϕ is another positive function known as *memory kernel*, since it depends on past events occurred before time t. Hawkes models can be used to identify dynamic interactions among a group of K processes. Once the Hawkes model is fitted, it will contain some weights representing the directional strength of any interaction between processes, interpreted as the expected number of events on a specific process resulting from an event on another process.

2 Main Findings

The goal of the analysis was to investigate and identify any cause-and-effect relationships between the social variables (the online world) and the financial variables of the cryptocurrency market (the financial world). Hawkes models made it possible to relate these two (apparently) unrelated worlds. Here the cause-and-effect relationships are intended as *predictive causality*, a particular case of Humean causality such as Granger causality [1], which could indeed be reconstructed through Hawkes models [2, 8]. In particular, we found: *i)* The relationships between the variables are self-excitement both before and after the event in which the price reaches its maximum (2nd September), with a loss of effect when the max_lag exceeds 24; *ii)* In the period leading up to the maximum point, the topic "trading and smart contracts" (Topic 3) causes increases in the delta_high_volatility metric where, on the contrary, the topic linked to uncertainty (Topic 1) generates a reduction; *iii)* From 3rd to 8th September the substantial price reduction (delta_volatility_low) generated a reduction in high volatility events, which in turn records a negative self-excitement relationship (-1.53). Figure 2 shows that there are self-excitement movements, with

slight mutual-excitement movements between the low delta price events towards other price metrics and topics.

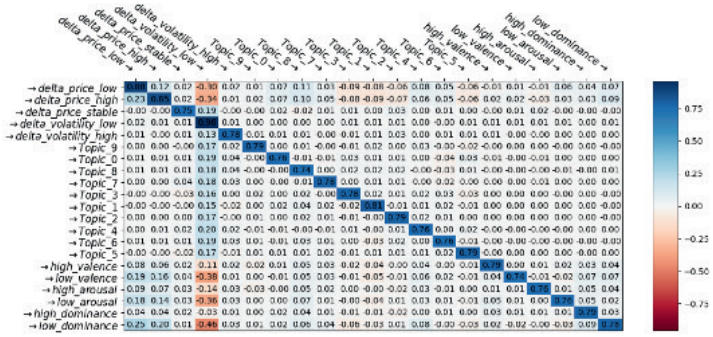
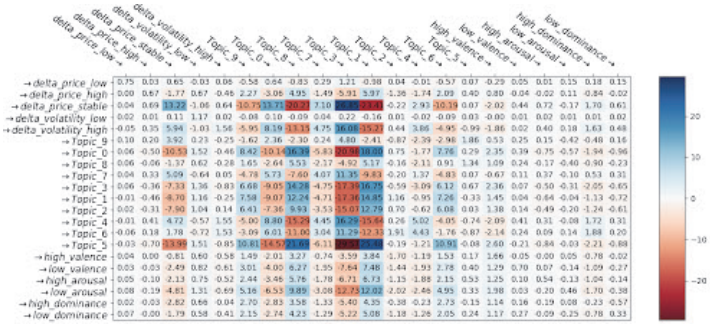


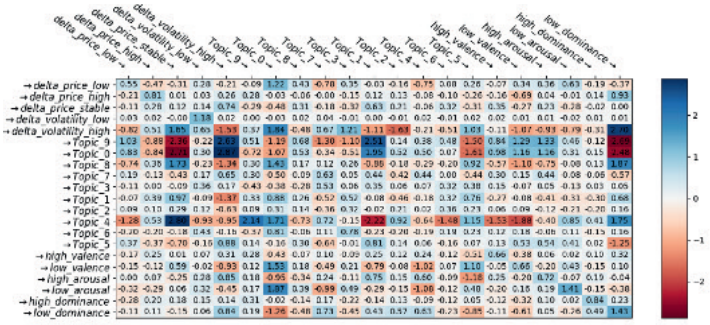
Fig. 2: Max_lag 24 for the entire considered period

The low volatility metric (delta_volatility_low) generates slight increases in all 10 topics, while a reduction in VAD metrics, but above all a reduction in financial metrics linked to high and low price reduction events. By reducing the number of days in the analysis, it is possible to note an important change in the behaviour of the matrix (Figure 3a). In fact, the diagonal of self-excitement disappears, in favour of important mutual-excitement recordings. The most important are those that link topic 3 (which deal with the themes of "trading and smart contracts") with the increase in price volatility (delta_volatility_high), while the topic linked to uncertainty (Topic 1) causes a reduction.
The coefficient matrix lights up differently (Figure 3b) considering the period 3-8 September 2021 (on 7th September the price settles down after the collapse occurred). In this case, events of high volatility suffer important effects, in particular due to the price reduction (delta_price_low) which generates the reduction together with itself (self-excitement effect with coefficient at -1.53). The low value of the comments also generates highly variable events, as well as the topic on NFTs and Trading (Topic 3 and Topic 5).

The purpose of this work was to identify the relationships between topic discussion occurrences on social media and Cardano (ADA) market price changes. To achieve this goal, dynamic topic modelling was first applied to social media content, and then a Hawkes model was used to decipher relationships between topics and cryptocurrency price movements. The results highlighted an interesting cause-effect (in the Granger causality sense) relationships between the social media metrics extracted from comments of domain expert redditors and the trend of ADA in two moments of price rise and fall. Also, the results show how social media plays an important role in these decentralized systems.



(a) Max_lag 7 from 27th of August till 2nd of September



(b) Max_lag 7 from 3rd of September till 8th of September

Fig. 3: Hawkes coefficients matrix of sub-periods.

References

1. Achab, M., Bacry, E., Gaıffas, S., Mastromatteo, I., Muzy, J.F.: Uncovering causality from multivariate hawkes integrated cumulants. In: International Conference on Machine Learning, pp. 1–10. PMLR (2017)
2. Bacry, E., Mastromatteo, I., Muzy, J.F.: Hawkes processes in finance. Market Microstructure and Liquidity **1**(01), 1550,005 (2015)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning, pp. 113–120 (2006)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
6. Phillips, R.C., Gorse, D.: Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In: 2017 IEEE symposium series on computational intelligence (SSCI), pp. 1–7. IEEE (2017)
7. Warriner, A.B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior research methods **45**(4), 1191–1207 (2013)
8. Zhang, W., Panum, T., Jha, S., Chalasani, P., Page, D.: Cause: Learning granger causality from event sequences using attribution methods. In: International Conference on Machine Learning, pp. 11,235–11,245. PMLR (2020)

Combined permutation test on the effect of age of micro enterprises on the propensity to Circular Economy

Test di permutazione combinato sull'effetto dell'età di micro imprese sulla propensione all'Economia Circolare

Stefano Bonnini[1] and Michela Borghesi[2]

Abstract Circular Economy (CE) is emerging as a possible strategy that companies might adopt to engage with environmental sustainability challenges. An interesting hypothesis concerns the effect of companies' age on the propensity of micro enterprises to undertake CE activities. Among the main difficulties in testing this hypothesis, are the multidimensional nature of the concept of CE, the small number of observations and the unbalanced sample sizes, typical situations of the empirical studies on this topic. We propose a solution based on a combined permutation test. Through a Monte Carlo simulation study, we investigate the power of this test. Finally, we consider a case study concerning Italian companies in the metal sector.

Abstract *L'Economia Circolare (CE) sta emergendo come una possibile strategia che le aziende di tutte le dimensioni potrebbero adottare per consentire loro di affrontare le sfide della sostenibilità ambientale. Un'ipotesi interessante riguarda l'effetto dell'età delle aziende sulla propensione delle micro imprese a intraprendere attività di CE. Tra le principali difficoltà nel verificare questa ipotesi, vi sono la natura multidimensionale del concetto di CE, le piccole dimensioni campionarie e lo sbilanciamento delle stesse, tipiche situazioni degli studi empirici sull'argomento. Proponiamo una soluzione basata su un test di permutazione combinato. Attraverso uno studio di simulazione Monte Carlo, indaghiamo sulla potenza di questo test. Infine consideriamo un caso studio riguardante aziende italiane nel settore della produzione e lavorazione dei metalli.*

Key words: permutation test, multivariate test, Monte Carlo simulation, Circular Economy, SMEs

[1] Stefano Bonnini, University of Ferrara; email: bnnsfn@unife.it

[2] Michela Borghesi, University of Ferrara; email: michela.borghesi@unife.it

1 Introduction

This paper deals with a multivariate two-sample test on proportions, based on the comparison of two multivariate Bernoulli distributions. For instance, it may be of interest to compare the propensity to Circular Economy (CE) of young and old companies. "Propensity to CE" is a multidimensional concept that concerns a set of attitudes and behaviors on the part of enterprises. The response represents presence or absence of such attitudes or behaviors, hence its components are dichotomous variables and the underlying probability distribution is a multivariate Bernoulli.

The aim of CE is to reduce the resources escaping from the circle so that the system functions in an optimal way. One goal of CE is to keep the added value in products for as long as possible and eliminate waste. When a product has reached the end of its life, resources are kept within the economy so that they can be productively be used again to create further value [7].

According to the existing literature, small-and-medium-sized enterprises are lagging behind on this topic. Many authors indicate that older firms are better to be involved in CE activities [1]. However, empirical studies specifically devoted to very small enterprises, the so-called micro enterprises, typical of the economic systems of several countries, including Italy, are missing. The published works usually suffer from some methodological limitations. In particular, they do not consider the multivariate nature of the response, they ignore the existence of confounding factors (firm size when testing the effect of age and vice-versa, economic sector, etc.), they use inferential methods not valid for small samples, they do not deepen the reliability of the results in case of unbalanced sample sizes.

The main goal of this work is to propose a two-sample permutation test for multivariate Bernoulli variables in order to investigate the relationship between propensity to CE and companies' age, overcoming the mentioned limitations of the current literature. The proposed method is valid also for directional alternatives and small samples. We investigate the power behaviour of the test through a simulation study. Furthermore, we contribute to the empirical literature on CE, applying the test to a case study concerning Italian micro enterprises in the metal sector.

Section 2 focuses on the presentation of the problem and on the methodological proposal. Section 3 is dedicated to the simulation study and Section 4 to the application problem. Section 5 includes results and concluding remarks.

2 Testing problem and methodological solution

Let us consider a multivariate test on proportions for two independent samples, where the multivariate response consists of q binary components, with $q \geq 2$. For the sake of simplicity, when a binary variable takes value 1 we symbolically denote the event a success. H_0 is the equality of the vectors of the q proportions of successes in population 1 and population 2 or, equivalently, the equality of the q marginal probabilities of success. For instance, the null hypothesis consists in the equality of

the proportions of young and old circular companies, according to q different points of view. We focus on the one-sided test because more challenging from the methodological point of view and more interesting in the applications. The adaptation of the procedure to the two-sided alternative is trivial. Hence, the alternative hypothesis H_1 is that the proportions of successes in population 2 are greater than those in population 1 (e.g. the propensity to be circular of young firms is less than that of old firms).

The general problem can be defined as follows, according to the union/intersection approach

$$\begin{cases} H_0: \bigcap_{k=1}^{q} (X_{1k} =^d X_{2k}) \\ H_1: \bigcup_{k=1}^{q} (X_{1k} <^d X_{2k}) \end{cases}$$

where X_{jk} is the Bernoulli random variable that takes 1 in case of success (e.g. circularity of a firm) in the j-th population according to the k-th partial aspect and 0 otherwise. The k-th partial null hypothesis is equivalent to $P(X_{1k} = 1) = P(X_{2k} = 1)$, whilst the k-th partial alternative hypothesis can be written as $P(X_{1k} = 1) < P(X_{2k} = 1)$. H_0 is true if all the partial null hypotheses are true. H_1 is true if at least one partial alternative hypotheses is true. When a partial alternative hypothesis does not hold, the corresponding partial null hypothesis is true. Each vector of q values observed on a statistical unit (e.g. firm) is supposed to be realization of the q-variate random variable $(X_{j1}, \dots, X_{jq})$ where $j = 1$ or $j = 2$ depending on whether the statistical unit belongs to sample 1 or sample 2 respectively. One crucial difficulty of such a problem is the need to take into account the dependence structure of the q marginal Bernoulli variables. This problem cannot be easily solved within a parametric approach.

The proposed solution is based the approach of Combined Permutation Tests (CPTs) [2]. CPTs consist of a family of nonparametric tests, suitable for problems that can be broken down into partial tests, robust and powerful in particular (but not only) for small sample sizes. For the application of this test, the independence with respect to units can be relaxed in favor of the exchangeability condition [5]. The solution is based on the application of q partial permutation tests on proportions. Each of them consists of testing the k-th partial null hypothesis versus the k-th partial alternative hypothesis with the difference between the proportion of sample 2 and sample 1 as a test statistic. The dependence between the partial tests in the null distribution is implicitly considered by permuting the rows of the dataset or equivalently by randomly assigning the observed row-vectors of q values to sample 1 and sample 2. For the general problem, a suitable combination of the q partial tests, based on the application of a function ψ on the partial p-values, can be used as a test statistic: $T_\psi = \psi(l_1, \dots, l_q)$, where l_k $(k = 1, \dots, q)$ is the p-value of the k-th partial test. We take into account the combining function of Fisher and Tippett, $T_F = -2\sum_{k=1}^{q} \log(l_k)$ and $T_T = \max_k(1 - l_k)$ respectively.

3 Simulation study

The power of the proposed test is assessed through Monte Carlo simulations. We consider different settings, according to the sample sizes n_1 and $n_2 = u \cdot n_1$ where $u = n_2/n_1$ is the unbalance ratio. For each setting, 1000 datasets with $q = 5$ response variables, are randomly generated. The vectors of success probabilities are $\theta_j \mathbf{1}_q$, $j = 1,2$, where θ_1 and $\theta_2 = \theta_1 + \delta$ are scalars taking values in the interval $(0,1)$ and $\mathbf{1}_q$ is a vector of q elements equal to 1. After the generation from the q-variate normal distributions $\mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, data are dichotomized according to θ_1 and θ_2. Let z_{ik} be the value of the i-th statistical unit and k-th variable generated from the normal distribution, the dichotomous value from the Bernoulli distribution of parameter θ_j is attained according to the following rule: $w_{ik} = 1$ if $z_{ik} \leq z_{1-\theta_j}$ and 0 otherwise, where $z_{1-\theta_j}$ represents the quantile of the standard normal distribution with cumulative probability equal to θ_j. The elements of $\boldsymbol{\Sigma}$ in the main diagonal, i.e. the variances σ_{kk}, $k = 1, \dots, q$, are set to 1, and the covariances (or equivalently the correlations) are assumed to be constant: $\sigma_{ks} = \rho$, $k \neq s$. We set $\theta_1 = 0.3$ and $\rho = 0.3$. In the null hypothesis $\delta = 0$ whilst in the alternative $\delta = 0.05$. For the computation of the p-values under H_0, instead of considering the exact permutation distribution, for computational convenience, we randomly generate 1000 permutations and compute approximate p-values according to the Conditional Monte Carlo approach [5]. The significance level of the simulations is α=0.05.

Table 1 reports the rejection rates of the permutation tests with the combinations of Fisher and Tippet. In both the cases the rates tend to be less than 0.05 and, when this is not true, they are very close to α. This conclusion holds for any sample size, it is valid for both balanced and unbalanced case and for any unbalance ratio. Hence, we can conclude that both the combined tests are well approximated.

Table 1: Rejection rate under H_0 with q=5, ρ=0.3, θ_1=0.3 and α=0.05

n_1	u	*Fisher*	*Tippett*	n_1	u	*Fisher*	*Tippett*
20	0.1	0.053	0.049	20	2	0.048	0.052
40		0.040	0.041	40		0.042	0.040
60		0.052	0.052	60		0.044	0.043
80		0.052	0.048	80		0.037	0.037
100		0.050	0.044	100		0.048	0.044
20	0.5	0.047	0.036	20	10	0.045	0.042
40		0.050	0.052	40		0.052	0.030
60		0.049	0.040	60		0.043	0.051
80		0.052	0.048	80		0.037	0.047
100		0.054	0.053	100		0.055	0.056
20	1	0.043	0.050				
40		0.047	0.046				
60		0.055	0.048				
80		0.046	0.052				
100		0.047	0.044				

Table 2 shows the rejection rates under the alternative hypothesis. It is evident that the power increases as the sample size increases and therefore the test is consistent. Since the rejection rates under H_1 are always greater than α, then the test is unbiased. The power increases with the unbalance ratio. Hence, the larger n_2 with respect to n_1, the higher the capacity of the test to recognize that the true hypothesis is $\theta_1 < \theta_2$. In all the considered settings Fisher's function seems to provide a combined test more powerful than Tippett's function.

Table 2: Power under H_1 with q=5, p=1, ρ=0.3, θ_1=0.3, α=0.05 and δ=0.05

n_1	u	*Fisher*	*Tippett*	n_1	u	*Fisher*	*Tippett*
20	0.1	0.076	0.061	20	2	0.148	0.113
40		0.093	0.068	40		0.226	0.177
60		0.102	0.095	60		0.318	0.220
80		0.123	0.089	80		0.350	0.248
100		0.152	0.107	100		0.385	0.258
20	0.5	0.121	0.094	20	10	0.161	0.115
40		0.161	0.129	40		0.270	0.173
60		0.171	0.134	60		0.360	0.247
80		0.250	0.168	80		0.461	0.327
100		0.263	0.177	100		0.501	0.341
20	1	0.141	0.115				
40		0.192	0.135				
60		0.207	0.158				
80		0.302	0.222				
100		0.349	0.243				

4 Application problem

The combined permutation test based on Fisher's function was applied to original data collected with the CATI method in January 2020. The survey involved Italian small and medium enterprises. We focused on the metal sector to avoid confounding effects due to sector heterogeneity and because this sector, more than others, lends itself to the CE. In fact metals are inherently recyclable, as they retain their utility for multiple product lifecycles and also for the versatility in favor of their reuse. For this reason, metal sector has always operated according to an approach of recovery, recycling and reuse [3]. Specifically, we analysed the interesting case, not much explored so far, of micro enterprises, i.e. firms with less than 9 employees [1]. We are in presence of small samples: 4 young and 32 old firms, with a balance ratio of 8.

The goal is to answer the following research question: "Is the propensity towards CE of younger firms lower, within the category of micro firms?". The responses are 6 binary variables that correspond to the following actions, chosen according to [6]: (1) innovations aimed at reducing the use of materials; (2) innovations aimed at

reducing the emitted waste; (3) innovations aimed at reuse of waste in its production cycle; (4) innovations aimed at transferring waste to other companies, which use them in their own production cycle; (5) innovations aimed at changing the design of the products in order to maximize their recyclability; (6) investments in R&D aimed at reducing the environmental impacts of production.

The p-value of the test is 0.266. It indicates no significance at α=0.05. Hence, there is not empirical evidence to reject the hypothesis of null effect of firm age.

5 Results and conclusions

The combined permutation test on the proportions of multivariate dichotomous variables is powerful, flexible because distribution-free, unbiased and consistent. This methodology is suitable also for small samples and its power depends also on the extent and the direction of the unbalance. The combination of Fisher is always preferable to that of Tippett. The application of the test to original sample data concerning Italian micro enterprises in the metal sector does not bring empirical evidence in favor of the hypothesis that the propensity towards CE is increasing function of age.

Acknowledgments

Authors thank the Italian Ministry of Education, University and Research that funded the departmental development program (DEM – University of Ferrara) for the period 2018-2022, to promote excellence in education and research (“Dipartimenti di Eccellenza”).

References

1. Bassi, F., Dias, J.G.: The use of circular economy practices in SMEs across the EU. Resources, Conservation & Recycling **146** (2019) 523–533. doi: 10.1016/j.resconrec.2019.03.019
2. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric hypothesis testing. Rank and permutation methods with applications in R. Wiley (2014)
3. Curcurutu, S., D’Amico, M., Merli, R., Preziosi, M., Acampora, A., Sandonnini, G.: EMAS ed Economia Circolare: Il caso studio del settore manifatturiero del metallo. ISPRA, Rapporti 299/2018 (2018) ISBN 978-88-448-0930-0
4. Murray, A., Skene, K., Haynes, K.: The circular economy: an interdisciplinary exploration of the concept and application in a global context. J Bus Ethics **140** (2017) 369–380. doi: 10.1007/s10551-015-2693-2
5. Pesarin, F., Salmaso, L.: Permutation tests for complex data: applications and software. Wiley series in probability and statistics (2010).
6. Saidani, M., Yannou, B., Leroy, Y., Cluzel, F., Kendall, A.: A taxonomy of circular economy indicators. Journal of Cleaner Production **207** (2019) 542-559. doi: 10.1016/j.jclepro.2018.10.014
7. Smol, M., Kulczycka, J., Henclik, A., Gorazda, K.: The possible use of sewage sludge ash (SSA) in the construction industry as a way towards a circular economy. Journal of Cleaner Production **95** (2015) 45-54. doi: 10.1016/j.jclepro.2015.02.051

Comparison of Two Different Approaches to Measure Economic Access to Food and Insecurity: an Application to Mexican data

Comparazione di due approcci per valutare l'insicurezza alimentare: un'applicazione sui dati ENIGH in Messico

Stefano Marchetti, Luca Secondi and Adrian Vargas-Lopez

Abstract The UN SDG 2 establishes the achievement of food security by 2030. The pandemic has exacerbated inequalities and what was primarily a problem of emerging countries now also characterizes developed countries. Measuring economic access to food properly is an important issue to be addressed to allow a constant official monitoring of the phenomenon. By using the Mexican National Survey of Household Income and Expenditure data, the aim of this paper is to assess the convergence between a qualitative assessment of food insecurity and an indicator of economic access to food obtained from Household Budget Surveys, usually carried out in several countries. As the first analyses suggest, when quality assessment of food insecurity is difficult to obtain, the food access economic indicator can be used instead to support data driven policy decisions.

Abstract *L'SDG 2 stabilisce il raggiungimento della sicurezza alimentare entro il 2030. La pandemia ha aumentato il livello di disuguaglianza anche nei Paesi sviluppati con situazioni di marginalità sociale che includono l'insicurezza alimentare. Si rendono necessarie misure appropriate per il monitoraggio della capacità di soddisfare i bisogni primari di alimentazione. In questo lavoro utilizziamo i dati di indagine messicani, per valutare la convergenza tra una misura di insicurezza alimentare qualitativa basata su domande specifiche e una quantitativa basata sulla spesa per alimenti. Come confermato dalle prime analisi, la misura quantitativa, più facilmente ottenibile dalle indagini correnti degli uffici nazionali di statistica in*

Stefano Marchetti
University of Pisa, Department of Economics and Management, Via Ridolfi 10 - 56124 Pisa, e-mail: stefano.marchetti@unipi.it

Luca Secondi
University of Tuscia, Department for Innovation in Biological, Agro-food and Forest Systems, Via S.C. De Lellis, snc 01100 Viterbo, e-mail: secondi@unitus.it

Adrian Vargas Lopez
Tecnologico de Monterrey, School of Government and Public Transformation, Mexico City, Mexico, e-mail: a.vargaslopez@tex.mx

moltissimi paesi, può rappresentare una preziosa fonte di informazioni per decisioni basate sui dati.

Key words: food expenditure, qualitative assessment, SDG 2

1 Introduction

The UN Sustainable Development Goals have set clear targets on global poverty, hunger and malnutrition to be achieved by 2030, which have prompted academics and policy-makers to identify and define useful strategies and measures as well as proper methods to assess drivers effectively. Specifically, SDG 2 foresees the reduction of food insecurity by reducing chronic hunger, defined as situation that exists when people lack access to sufficient amounts of nutritious food for an active and healthy life as measured by the Prevalence of Undernourishment (PoU), in accordance with SDG Indicator 2.1.1 ([1]).

Food insecurity is a complex problem, manifesting as obesity and malnutrition in addition to extreme hunger and starvation ([2]) and it is commonly defined as *"having at all times, physical, social and economic access to sufficient, safe and nutritious food that meets dietary needs and food preferences for an active and healthy life"* ([3]).

Food insecurity and the related notion of economic access to food may appear to refer exclusively to developing countries, but in actual fact it is a phenomenon that was also present in developed and affluent countries ([4]) including Europe ([5]) and more generally OECD countries, even before the outbreak of Covid-19 pandemic even if the pandemic has exacerbated inequality, with almost 811 million people faced hunger in 2020, according to figures by the Food and Agriculture Organization of the United Nations.

In this perspective a correct and proper measurement of this phenomenon is essential. Focussing on measuring economic access to food and (in)security both qualitative scales and indicators based on the household income expenditure have been proposed. However, these measures have been used as alternatives or as single dimensions within a multidimensional perspective capturing availability, access, utilization and stability of food.

However, since in developed countries it is the task of official statistics to conduct the survey on household consumption by collecting household expenditures for different categories of expenditure (i.e. the so-called Household Budget Surveys, HBSs) it would be good to investigate to what extent measures based on data of an economic nature (income or consumption) are aligned with qualitative measures and scales that are already well explored and widely developed, instead, in emerging countries.

The statistical robustness and correspondence of information and classification between different approaches may in fact allow for the extension of the use of HBSs (but also surveys on income) for further purposes that go beyond or, even better,

complement the analyses that have so far been conducted on poverty in general, expanding them towards the analysis of poverty and food insecurity.

In this study, we focus on Mexico and specifically on the National Survey of Household Income and Expenditure (ENIGH) carried out every two years by the National Institute of Statistics and Geography (INEGI). We aim to compare two measures of food insecurity. The first is based on the ENIGH set of questions to measure food insecurity from a qualitative perspective. The latter is based on household consumption expenditures - and specifically household consumption expenditure on food - that is also surveyed in ENIGH. Therefore, the possibility of referring to a double perspective within the same survey has motivated us to explore the statistical reliability of these two approaches in measuring a unique phenomenon.

2 Data and Method

The National Survey of Household Income and Expenditure (ENIGH) is a representative sample survey conducted in Mexico every two years with the general aim of providing stakeholders with a statistical overview of income and expenses behaviour of households in terms of their amount, origin and distribution.

For our analysis we use the ENIGH 2018 wave, which involves a national sample of 87,826 housing units. INEGI uses a stratified, two-stage design, where the secondary units are the households and the primary units are municipalities, while stratification is done according to geographic characteristics.

Within the Mexican ENIGH survey the food insecurity issue is assessed through the use of the Latin American and Caribbean Food Security Scale (ELCSA) by taking advantage of 6 questions where families signal if during the past three months they had access to a limited variety of food, whether they skipped a meal, if they had eaten less than they thought they should, if they ran out of food, if they felt hungry but did not eat, and if they had not eaten for a whole day. These questions are asked twice if in the household there are children (i.e., individuals younger than 18). The second time, respondents answer for the infants living in the dwelling ([6]).

The severity of food insecurity is constructed by the number of questions that people answer affirmatively. When households without children answer "Yes" to 5-6 questions, they are Severely Insecure. If they answer 3-4 questions affirmatively, they are Moderately Insecure; 1-2 questions, Mildly Insecure; and, 0 questions, Secure. Similarly, each threshold is built for households with children. In Mexico we estimated 60.3% of people is food secure, while 39.7% are not. In particular 19.3% are mildly insecure, 12.1% are moderately insecure and 8.3% are severely insecure. There are about $49,670,000$ persons who are not food secure. Figure 1 shows the proportions of persons by severity of food insecurity according to the ELCSA scale divided by households with and without children. We can observe that households with children show a higher level of food insecurity, highlighting the difficulties of such households to afford adequate nutrition.

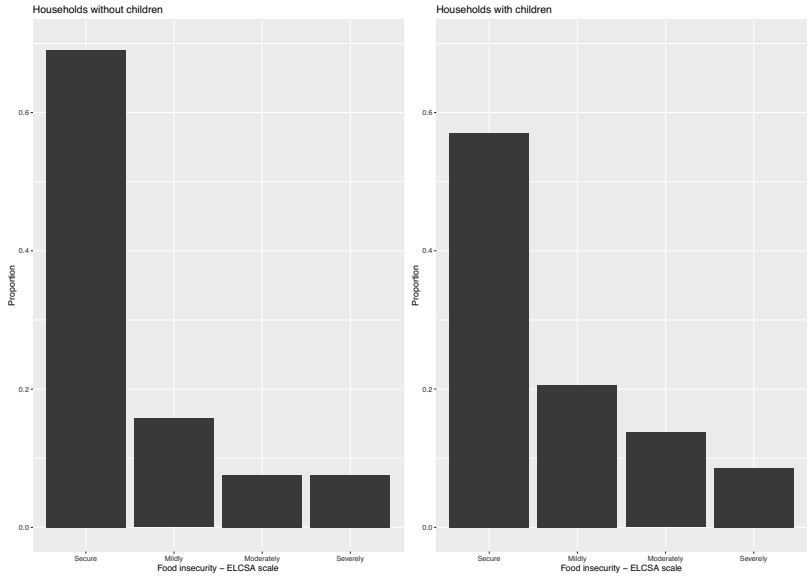


Fig. 1 Proportions of people living without (left) and with (right) children by ELCSA food insecurity scale. Source: ENIGH 2018 data.

The ENIGH surveys also information about household consumption expenditure, used to obtain the share of food expenditure and a proxy indicator of the food (in)security ([1]). Let x_i be the total expenditure and y_i the food expenditure for household i, then the share of food expenditure at household level is simply defined as $SF_i = y_i/x_i$. According to the Integrated Food Security Phase Classification (IPC, 2021), we use a four-group classification: i. no food insecurity if $SF < 40\%$, ii. mild food insecurity if $40\% \leq SF < 50\%$, iii. moderate food insecurity if $50\% \leq SF < 70\%$ and iv. severe food insecurity if $SF \geq 70\%$. We estimated at national level 48.3% of persons being food secure, 23.5% being mildly food insecure, 24.4% being moderately food insecure and 3.8% being severely food insecure. According to this measure, about $64,675,000$ persons are not food secure. The picture obtained is a little bit different from the previous one, and seems to emphasize more the food insecurity phenomenon. We show in figure 2 the proportions of persons living in the different food (in)security conditions, classified by persons living with or without children. Using this measure, the difference between households with and without children are minimal, with a bigger proportion of food severely insecure for person living without children.

Both the food insecurity measures are estimated using the Horwitz and Thompson expansion estimator. Standard error are obtained using Taylor approximation and are considered small at national level.

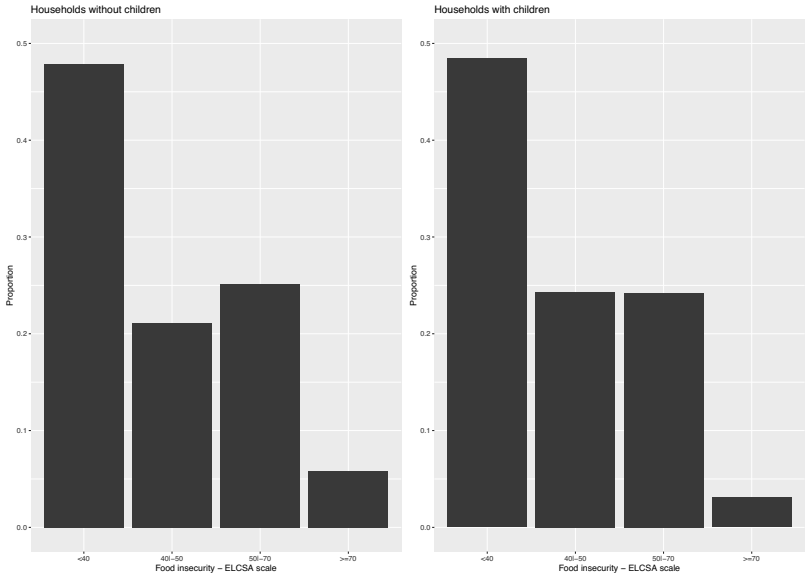


Fig. 2 Proportions of people living without (left) and with (right) children by IPC food insecurity scale. Source: ENIGH 2018 data.

3 Discussion on the comparison between food insecurity measures

In this section we compare the food insecurity measure obtained from the ELCSA scale and from the IPC scale. The first is based on specific questions aimed at investigating about properly nutrition, the latter is based on the consumption expenditure, usually surveyed in many countries.

The two measures are dependent according to the Pearson χ^2 test for contingency table with the first and second-order Rao-Scott corrections (p-value $< 2.2e-16$) ([7]). Even if we carry out the test within regions the result does not change. We also carry out a regression model to assess the relation between the *SF* and the food insecurity on the quality ELCSA scale and some controlling variables (presence of children, regions and urban/rural area). The average *SF* is significantly different among secure, mild insecure, moderately insecure and severe insecure categories.

Although the two measures are dependent there are some differences. Figure 3 visualizes the cross classification of food insecurity level between ELCSA and IPC scale. Firstly, it is reasonable that persons classified severely, moderately or mild insecure on the IPC scale can be secure according to the ELCSA scale, this mean that a very large amount of resources are used for nutrition. What need further investigations are those persons who are secure on IPC scale (i.e. $SF < 40\%$) while mildly, moderately or severely insecure according to ELCSA scale. Indeed, these

persons devote less than 40% to consumption expenditure for food, but they have difficult to have an adequate nutrition level.

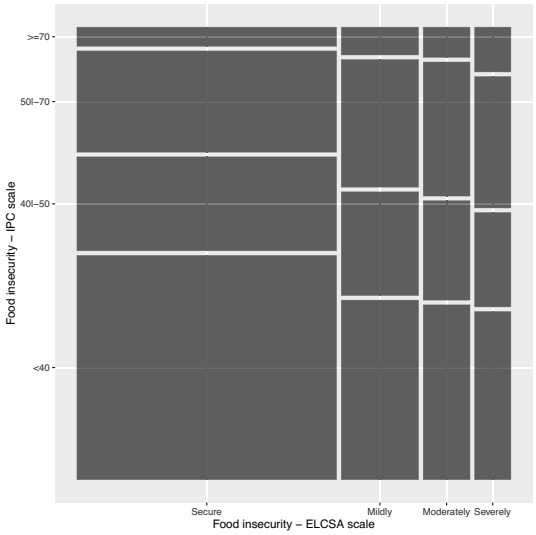


Fig. 3 Cross classification of food insecurity level between ELCSA and IPC scale. Source: ENIGH 2018 data.

We can conclude that the IPC based scale emphasizes more the problem than ELCSA scale, but exclude persons that have not access to sufficient nutrition. Future works will investigate on these issues.

References

1. Marchetti, S., Secondi, L. The Economic Perspective of Food Poverty and (In) security: An Analytical Approach to Measuring and Estimation in Italy. Social indicators research, 1-26. (2022)
2. Candel, J. J. Food security governance: a systematic literature review. Food Security, **6**(4), 585-601. (2014)
3. World Food Summit (1996). Rome Declaration on World Food Security.
4. Zaçe, D., Di Pietro, M.L., Reali, L., De Waure, C., Ricciardi, W. Prevalence, socio-economic predictors and health correlates of food insecurity among Italian children-findings from a cross-sectional study. Food Security, **13**(1), 13-24. (2021)
5. Penne, T., Goedemé, T. Can low-income households afford a healthy diet? Insufficient income as a driver of food insecurity in Europe. Food Policy, 99, 101978. (2021)
6. Villagómez-Ornelas, P., Hernández-López, P., Carrasco-Enríquez, B., Barrios-Sánchez, K. Statistical validity of the Mexican Food Security Scale and the Latin American and Caribbean Food Security Scale. Salud publica de Mexico, **56**1 5-11. (2014)
7. Rao, JNK, Scott, AJ On Chi-squared Tests For Multiway Contigency Tables with Proportions Estimated From Survey Data. Annals of Statistics **12** 46-60. (1984)

Image analysis and visual methods

Bias correction of the maximum likelihood estimator for Emax model at the interim analysis

Correzione per distorsione della stima di massima verosimiglianza del modello Emax nell'analisi ad interim

Caterina May and Chiara Tommasi

Abstract The Emax model is a dose-response model commonly applied in clinical trials, agriculture and environmental experiments. We consider a two-stage adaptive design for collecting "optimal" data for estimating the model parameters. At the first stage (interim analysis) a locally D-optimum design is computed to get a sample of independent observations and to produce a first-stage maximum likelihood estimate (MLE). At the second stage, the first-stage MLE is used as initial parameter-value to determine another D-optimum design and then to collect the second-stage observations.
The first-stage estimate influences the quality of the data gathered at the second stage, where a large number of observations can be collected. In real life problems, instead, the sample size of the interim analysis is usually small; therefore, the first-stage MLE should be precise enough even if based on few data. From this consideration, our guess is that if we improved the behaviour of the first-stage MLE through a bias correction, then the D-optimal design determined at the second stage would produce better experimental points. In this study we provide the analytic expression of the first-order bias correction of the MLE in the Emax model.
Abstract *Il modello Emax è un modello di risposta alla dose comunemente usato negli esperimenti clinici, agricoli e ambientali. In questo lavoro consideriamo un disegno adattivo a due stadi per raccogliere dati "ottimali" al fine di stimare i parametri del modello. Al primo stadio (analisi ad interim) viene calcolato un disegno localmente D-ottimo per raccogliere un campione di osservazioni indipendenti e ottenere una stima di massima verosimiglianza (SMV). Al secondo stadio la SMV di primo stadio viene usata come valore inziale del parametro per determinare un secondo disegno D-ottimo e raccogliere le osservazioni di secondo stadio a par-*

Caterina May
Università degli Studi del Piemonte Orientale, Department DiSEI, via Perrone, 18, 28100 Novara (Itay), e-mail: caterina.may@uniupo.it

Chiara Tommasi
Università degli Studi di Milano, Department DEMM, via Conservatorio, 7, 20122 Milano (Italy)
e-mail: chiara.tommasi@unimi.it

tire da quest'ultimo disegno sperimentale. Tali osservazioni dipendono quindi dalle risposte precedenti.
La stima di primo stadio influenza la qualità dei dati al secondo stadio. Nei problemi reali la dimensione campionaria dell'analisi ad interim è di solito bassa, quindi la SMV di primo stadio sarrebbe opportuno fosse precisa anche se basata su un campione piccolo. La nostra idea è quindi quella di migliorare la SMV di primo stadio correggendone la distorsione, così che il disegno D-ottimo al secondo stadio sia migliore. In questo lavoro determniamo l'espressione analitica al primo ordine della distorsione della SMV di primo stadio dei parametri del modello Emax.

Key words: maximum likelihood estimator, interim analysis, small sample, bias correction, D-optimality, two-stage adaptive design, Emax model

1 Introduction

The *Emax* model is well-characterized in the literature and it is frequently used for dose-response designs in clinical trials, as well as in agriculture and in environmental experiments (see, for instance, [3] and [2]). It has the form $y = \eta(x,\theta) + \varepsilon$ where y denotes a response at the dose $x \in \mathscr{X} = [a,b]$, $0 \le a < b$; $\theta = (\theta_0, \theta_1, \theta_2)^T$ is a vector of unknown parameters; ε is a Gaussian random error; the nonlinear mean response is

$$\eta(x,\theta) = \theta_0 + \theta_1 \frac{x}{x+\theta_2}. \tag{1}$$

In equation (1), θ_0 represents the response at the dose zero; θ_1 is the maximum effect attributable to the drug; and θ_2 is the dose which produces the half of the maximum effect.

In this study, to collect observations that provide a precise estimation of θ, we consider a two-stage adaptive design. By the fact, sequential adaptive designs are quite common in clinical trials. More specifically, assume that a guessed value $\tilde{\theta} = (\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2)^T$ for θ is available, for instance from an expert opinion. At the first stage (or interim analysis) we take $n_1 < n$ observations according to a locally D-optimal design

$$\xi_1^*(\tilde{\theta}) = \arg\max_{\xi \in \Xi} |M(\xi;\tilde{\theta})|,$$

where $\xi = \begin{Bmatrix} x_{11} & \cdots & x_{1M_1} \\ \omega_{11} & \cdots & \omega_{1M_1} \end{Bmatrix}$ denotes a design, which is defined as a finite discrete probability distribution over $\mathscr{X}$, Ξ is the set of all possible designs and

$$M(\xi;\theta) = \int_{\mathscr{X}} \nabla\eta(x,\theta)\nabla\eta(x,\theta)^T d\xi(x), \tag{2}$$

is the information matrix of ξ. Moreover $\nabla\eta(x,\theta)$ denotes the gradient of the mean response $\eta(x,\theta)$ with respect to θ (see for instance, [8] or [1] as references in op-

timal design of experiments). The design $\xi^*(\tilde{\theta})$ is said locally optimal because it depends on a guessed parameter value $\tilde{\theta}$ due to the non-linearity of $\eta(x,\theta)$. Since a D-optimal design minimizes the generalized asymptotic variance of the MLE for θ, it should improve the precision of the parameter estimates. Let $\hat{\theta}_{n_1}$ be the first-stage MLE based on the n_1 observations gathered during the interim analysis. At the second stage, the available data information can be used to improve the choice of the experimental points. Therefore, $n_2 = n - n_1$ additional responses are collected according to another locally D-optimal design, $\xi_2^*(\hat{\theta}_{n_1})$, where $\hat{\theta}_{n_1}$ is used in (2) instead of $\tilde{\theta}$.

The final MLE is computed employing the whole sample of $n = n_1 + n_2$ data, which are dependent because the second-stage data depend on the first-stage responses through $\hat{\theta}_{n_1}$. In [9] and in [10] the theoretical properties of this final MLE are described.

The sample size of the interim analysis is usually small and thus $\hat{\theta}_{n_1}$ might be affected by the bias which converges to zero as n increases to infinity. On the other hand, the second-stage D-optimal design depends on $\hat{\theta}_{n_1}$ and at this stage a larger number of observations are collected. If these n_2 responses are observed at bad design points (or in bad proportions), they might produce an unreliable final MLE. On the other hand, if we improve the behaviour of the first-stage MLE through a bias correction, then the D-optimal design determined at the second stage should produce better experimental points. In Section 2 we provide the analytic expression of the first-order bias correction of the MLE in the *Emax* model.

2 Simulations of MLEs efficiencies and first order bias correction

As explained in the introduction, we need to understand how the MLE at the first stage influences the D-optimal design at the second stage. From the analytical expression of the locally D-optimal design ξ_D^* for the *Emax* model (provided by [7]):

$$\xi_D^*(\theta) = \xi_D^*(\theta_2) = \begin{Bmatrix} a & x^*(\theta_2) & b \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}, \tag{3}$$

where the interior support point is

$$x^*(\theta_2) = \frac{b(a+\theta_2) + a(b+\theta_2)}{(a+\theta_2) + (b+\theta_2)}, \tag{4}$$

we have that the second stage D-optimal design depends only on θ_2.

The behaviour of $x^*(\theta_2)$ when $[a,b] = [0,150]$ is plotted in Figure 1. Let θ_2^t denotes the "true" value of θ_2. From Figure 1 we can note that the derivative of $x^*(\theta_2)$ is a positive decreasing function of θ_2 and thus the effect on $x^*(\theta_2)$ is larger for the values $\theta_2 < \theta_2^t$.

The following theorem provides the expression for the bias of the first stage MLE of θ_2, which is herein denoted by $\hat{\theta}_{n_1,2}$.

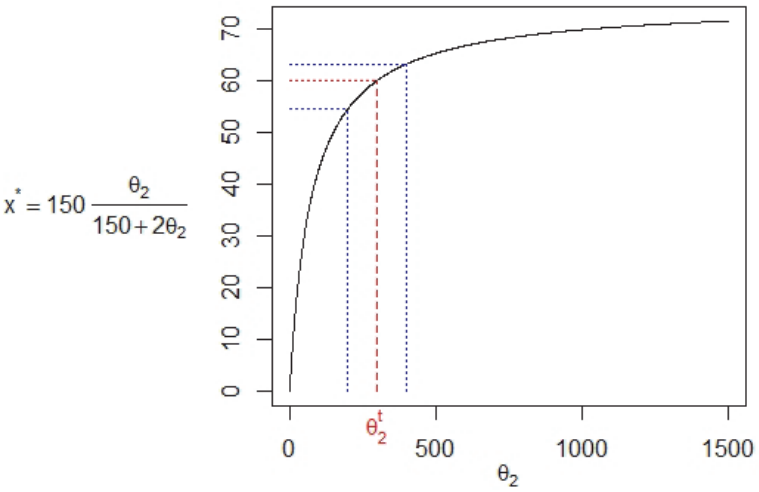


Fig. 1 D-optimum middle dose $x^*(\theta_2)$ for the Emax model

Proposition 1. *Let $\theta_{2,0}$ be a nominal value for θ_2. If n_1 first stage observations are taken according to the local D-optimal design (3), with equal numbers treated at the experimental points a, $x^*(\theta_{2,0})$ and b, then the bias of the first stage MLE of θ_2 is*

$$E(\hat{\theta}_{n_1,2} - \theta_2) = \frac{b_2(\theta)}{n_1} + O(n_1^{-2}),$$

where $b_2(\theta) > 0$ is given by

$$\begin{aligned} b_2(\theta) = & \frac{1}{(a-b)^4\theta_1^2\theta_2^2(a+\theta_{2,0})^2(b+\theta_{2,0})^2} \\ & \cdot\big\{3\sigma^2(a+\theta_2)^2(b+\theta_2)^2[2ab+(a+b)\theta_{2,0}+\theta_2(a+b+2\theta_{2,0})]^2 \\ & [3ab(a+b)+(a^2+10ab+b^2)\theta_{2,0}+3(a+b)\theta_{2,0}^2 \\ & +2\theta_2(a^2+ab+b^2+3(a+b)\theta_{2,0}+3\theta_{2,0}^2)]\big\}. \end{aligned} \tag{5}$$

Proof. Cox and Snell (1968) introduced the $O(n^{-1})$ formula for the bias of the MLE in the case of n observations not being identically distributed. Cordeiro and Klein (1994) proposed a matrix expression for this bias, which is herein specialized for the Emax model and the D-optimal design $\xi_D^*(\theta_{2,0})$. Calculations are available by the authors upon request.

This result justifies the fact that, when $\theta_{2,0} < \theta_2^t$ the fixed procedure (that consists in collecting all the n observations according to the initial (first-stage) D-optimal design) seems to have a worst performance than a two-stage design; in facts, $\hat{\theta}_{n_1,2}$

has a positive bias and thus it takes (on average) larger values, and thus we expect that $x^*(\hat{\theta}_{n_1,2})$ is closer to $x^*(\theta_2^t)$ than $x^*(\theta_{2,0})$ does.

3 Conclusions

In this paper we have presented the idea of improving the two-stage adaptive design proposed in [9] by introducing a bias correction. An analytic form of the first-order bias correction under the Emax model has been provided. We also justify the non-symmetric performance of the adaptive procedure in comparison with a fixed one, that results from the simulations in [9].

As a future work, we aim to apply the bias correction to the two-stage procedure in order to investigate its possible improvement. We aim at exploring also other types of bias corrections, in particular the ones obtained by modifying the score function (see, for instance, [11]).

References

1. A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum experimental designs, with SAS*, volume 34 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 2007.
2. M. Baker, J. L. Hobman, C. Dodd, S. J. Ramsden, D. J. Stekel. Mathematical modelling of antimicrobial resistance in agricultural waste highlights importance of gene transfer rate. *FEMS Microbiology Ecology*, 92(4), 2016.
3. F. Bretz, H. Dette, and J. Pinheiro. Practical considerations for optimal designs in clinical dose finding studies. *Statistics in medicine*, 29(7-8), 731–742, 2010.
4. B. Bornkamp, J. Pinheiro, and F. Bretz. *DoseFinding: Planning and Analyzing Dose Finding Experiments*, 2018. R package version 0.9-16.
5. D.R. Cox and E.J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society (B)*, 30 (2):248-275, 1968.
6. G.M. Cordairo and R. Klain. A general definition of residuals. *Statistics and Probability Letters*, 19:169-176, 1994.
7. H. Dette, C. Kiss, M. Bevanda, and F. Bretz. Optimal designs for the emax, log-linear and exponential models. *Biometrika*, 97(2):513–518, 2010.
8. V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.
9. N. Flournoy, C. May, and C. Tommasi. The effects of adaptation on maximum likelihood inference for nonlinear models with normal errors. *Journal of Statistical Planning and Inference*, 214:139-150, 2021.
10. C. May, and C. Tommasi. On the behaviour of the maximum likelihood estimator for exponential models under a fixed and a two-stage design. *Book of Short Papers SIS 2020*, Pearson, 2020.
11. L. Pronzato and A. Pázman. *Design of experiments in nonlinear models*, volume 212 of *Lecture Notes in Statistics*. Springer, New York, 2013. Asymptotic normality, optimality criteria and small-sample properties.

Visual and automated methods in digital microscopy to evaluate fungal colonisation on plant roots

Metodi visivo ed automatizzato per valutare la colonizzazione di funghi nelle radici di piante in microscopia digitale

Ivan Sciascia[1], Andrea Crosino[1] and Andrea Genre[1]

Abstract The vast majority of studies investigating arbuscular mycorrhizas (AM) and their applications in agriculture require a precise quantification of the intensity of root colonisation. Here we propose a novel semi-automated approach to quantify AM fungal root colonisation based on digital image analysis comparing three methods: i) visual quantification ii) image thresholding, iii) machine learning. We recognize machine learning as a very promising tool for accelerating, simplifying and standardizing this critical step in analyzing AM of interest to applied and basic science.

Abstract *La grande maggioranza degli studi sulle micorrize arbuscolari (AM) e la loro applicazione in agricoltura richiede una precisa quantificazione dell'intensità della coloniszazione della radice. Proponiamo un metodo semiautomatico per quantificare la coloniszazione da funghi AM basata sull'analisi delle immagini digitali confrontando tre metodi: i) quantificazione visiva, ii) sogliatura di immagini, iii) machine learning. Riconosciamo il machine learning come metodo molto promettente per accelerare, semplificare e standardizzare la quantificazione della coloniszazione delle radici con AM, di interesse nella scienza applicata e di base.*

Key words: agriculture**,** image segmentation, microscopy, arbuscular mycorrhiza, digital thresholding, machine learning

[1] Ivan Sciascia, Department of Life Sciences and Systems biology;
ivan.sciascia@unito.it
Andrea Crosino, Department of Life Sciences and Systems biology;
andrea.crosino@unito.it
Andrea Genre, Department of Life Sciences and Systems biology;
andrea.genre@unito.it

Introduction

Arbuscular mycorrhizas (AM) are widespread plant endosymbioses that develop between Glomeromycotina fungi and the roots of the majority of plant species, including most crops. The symbiosis benefits extend to both partners by improving plant mineral absorption, tolerance to biotic and abiotic stresses and fitness, while rewarding the fungal symbionts with carbohydrates compounds derived from photosynthesis process, such as sugars and lipids (Rich et al. 2017). The main applied benefit of AM could be in low-input farming. A critical step in all studies on AM is the precise quantification of root colonisation by AM fungi (Ferrol and Lanfranco 2020). However, molecular analyses based on the quantification of fungal sequences in total root DNA or arbuscule-specific markers in total root RNA extracts are still outnumbered by direct microscopic quantification of the intraradical fungal structures after histochemical staining (McGonigle et al. 1990; Novero et al. 2002). In an attempt to improve the spread, repeatability, and reliability of root colonisation measurements, we here present a semi-automated approach based on digital imaging and different types of image processing techniques and compared it to the commonly used visual method developed by Trouvelot et al (1986). We designed a semi-automated algorithm to generate quantitative indexes of root colonisation deriving from either image thresholding using ImageJ, or based on machine learning analyses (Arganda-Carreras et al. 2017) using the commercial software Zeiss Intellesis (Volkenandt et al. 2018). Our analyses identify machine learning as a promising alternative approach to visual quantification of AM fungal root colonisation.

Materials and methods

To quantify the extent of AM fungal root colonisation, we acquired two sets of images of mycorrhizal and not mycorrhizal roots stained with 0.1% methylene blue in lactic acid. Quantitative AM fungal colonisation was measured based on the frequency and abundance of fungal structures following Trouvelot et al. (1986) as described in Volpe et al. (2020) considering it as the i) visual method. The images were acquired with CI-L fitted with a 4x / 0.10 WD30 objective or Leica DMA500 fitted with a PLAN4x/0.10 objective. The fungal structures were stained blue while the plant tissues and cells remained transparent or light blue. As ii) image thresholding we use ImageJ framework: a color image can be represented by three components: red, green and blue (RGB) and the shades of each color can be represented by $2^8 \rightarrow [0,255]$ bit, thus using one byte (1byte = 8 bits) for each color. A threshold value can be applied to an image: if we have a gray-toned image or on a single RGB channel in the form $2^8 \rightarrow [0,255]$ we can collect the area size of the pixels which are in the range between minimum and maximum threshold according to the function:

$$S = \left\{ \begin{array}{c} a \quad if\ i_{xy} \leq s \\ b\ otherwise \end{array} \right\}$$

Considering the last approach as iii) machine learning, we tested an image analysis procedure, based on machine learning, using Zeiss Zen Intellesis application (Carl Zeiss Microscopy GmbH Jena, Germany) (Nesbit et al. 2021). Unlike image segmentation by intensity thresholding, image analysis by machine learning extracts features of images that are then used for classification. Machine learning is a branch of artificial intelligence that solves tasks using algorithms that are capable of learning from experience (training), without being explicitly programmed for a specific task. Differently from the intensity thresholding based approach, the method we used in this analysis was based on the neighborhood pixel characteristics that were subsequently classified based on description vectors. Zeiss Intellesis can apply a number of machine learning tools called basic features. In this study, we used basic feature 25, which achieves digital image segmentation by applying a number of filters based on 25 parameters derived from Gaussian, Gabor and Hessian filtering.

Results

After applying segmentation macro to all our images, a set of quantitative values was obtained corresponding to the supposed colonised area (darkest pixels) and the total root section area (as isolated from the image background).
The ratio between the colonised and total root area in an image was used as threshold (t) index and represents the intensity (I) of root colonisation:

$t = \frac{mycorrhized\ area}{total\ area} * 100$

In Figure 1, we summarize the fit of the t-index of each visually classified intensity class (1-6) describing x-y axis between visual classification (x) and the t index (y) and in Table 1 we perform ANOVA with pairwise comparisons between visually classified intensity class inferred by t-index.

Table 1. ANOVA Pairwise (I-J) correlation analyses of the root colonisation intensities as inferred by the t-index in the visual categories. Statistics with pairwise post hoc multiple comparisons (Bonferroni method).*Mean difference is significant at 0.05 level.

Visual categories	***1***	***2***	***3***	***4***	***5***	***6***
1		4.86	13.38*	14.16*	28.74*	28.95*
2			8.51*	9.29*	23.87*	24.09*
3				0.77	15.35*	15.57*
4					14.57*	14.79*
5						0.21
6						

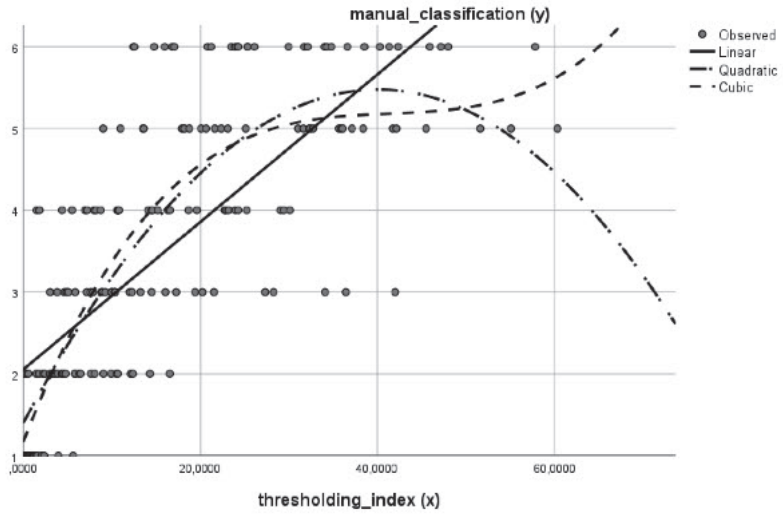


Figure 1: Root colonisation intensities as determined by contrast thresholding and as classified by manual image scoring. Cubic fit R^2=0.687

Considering now the machine learning perspective we used the software Zeiss Intellesis and after the procedure of training and segmentation performed by the tool we calculated the fractional total root area colonised as I = ml which can be considered as root colonisation intensity (I)

$$ml = \frac{colonized\ area}{colonized\ area + non\ colonized\ area} * 100$$

In the Table 2 and Figure 2 we show the fit statistics and ANOVA pairwise comparisons for the calculated ml-index, based on Zeiss Intellesis.

Table 2. ANOVA Pairwise (I-J) correlation analyses of the root colonisation intensities as inferred by the ml-index in the visual categories. Statistics with pairwise post hoc multiple comparisons (Bonferroni method).*Mean difference is significant at 0.05 level.

Visual categories	***1***	***2***	***3***	***4***	***5***	***6***
1		1.46	10.87*	20.36*	37.72*	42.00*
2			9.41*	18.90*	36.26*	40.54*
3				9.49*	26.84*	31.12*
4					17.35*	21.63*
5						4.27
6						

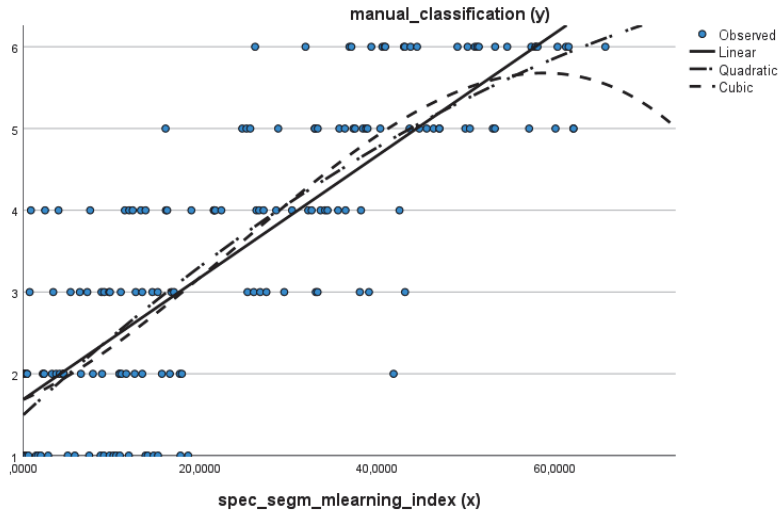


Figure 2. Root colonisation intensities as determined by simple image machine learning and as classified by manual image scoring. Cubic fit R^2=0.728

4 Discussion

The present study evaluated the reliability of two semi-automated image analysis methods in comparison to visual scoring (Trouvelot et al., 1986) to evaluate root AM quantification (Vierheilig et al. 2005). Contrast thresholding uses the gradient of pixel brightness (inversely related to cotton blue staining) as an indicator of fungal presence. Quantification of root colonisation intensity by contrast thresholding resolved the six root colonisation intensity classes, as used for visual scoring, and can, therefore, be considered reliable for rapidly screening root samples A few critical aspects should anyway be considered. One major limitation of the thresholding method is the variability of brightness range between images: different dyes, optical setups, root translucence and the presence of additional microorganisms (such as bacteria, algae, endophytic fungi, invertebrates) especially in field samples, that cannot be discriminated from fungal structures simply based on pixel brightness. In addition, the method is strongly affected by image background noise and the utilized magnification. Lastly, the segmentation process can only be set *ex ante*, by changing the setting of the macro without any possibility for subsequent intervention by the user. The machine learning-based procedure of the Zeiss Zen Intellesis turned out to be the most efficient. It allows for discrimination among the different fungal structures, such as hyphae, arbuscules and vesicles, during the manual training phase to generate a model that the software then applies to all analysed samples. This approach also resolved the six classes of intensity and achieved the best correlation with manual colonisation scoring. Importantly, training phase was relatively short (it requires 50 minutes overall) and turned out to be effective, even when using a limited number of images (10 images). Lastly, the use of machine-learning allowed reliable discrimination between fungal structures, such as intra- and extraradical hyphae, and arbuscules, something that contrast thresholding of images could not resolve. A critical factor for accurate machine learning-based root colonisation level assessments is that the software is trained by an expert. The analysed

images of the training phase are saved in a reference folder and can be shared for verification by other researchers and even re-use. This opens a new perspective for data reproducibility. Experts can share their expertise with the entire research community by simply sharing the images. Furthermore, the training file can become a common resource for reference and, thereby, make the detected intensity by AM fungi more uniform and repeatable. The current model of machine learning could be extended so that it cannot only assign samples to mycorrhization intensity classes, but also reliably discriminate between root colonisation by arbuscules, vesicles, hyphal coils and so forth. At present, the commercial nature of the software hampers the modification of the image analysis algorithm, which could, however, be overcome as shown by Evangelisti et al. (2021). A simple increase in the number of images used for training might be sufficient to refine the efficiency of machine learning to discriminate among different fungal structures. Indeed, considering the method recently published by Evangelisti et al. (2021) was described to reliably discriminate among fungal structures (arbuscules, hyphae, and vesicles). It would be extremely interesting to investigate the possibility to merge the two algorithms in an attempt to develop a more powerful tool for image analysis that could make quantification of root colonisation by AM fungi more reproducible and allow discrimination of root colonisation and a detailed identification of intraradical fungal structures.

References

1. Arganda-Carreras I, Kaynig V, Rueden C, et al (2017) Trainable Weka Segmentation: A machine learning tool for microscopy pixel classification. Bioinformatics 33:2424–2426. https://doi.org/10.1093/bioinformatics/btx180
2. Evangelisti, E., Turner, C., McDowell, A., Shenhav, L., Yunusov, T., Gavrin, A., Servante, E.K., Quan, C. and Schornack, S. (2021), Deep learning-based quantification of arbuscular mycorrhizal fungi in plant roots. New Phytol. https://doi.org/10.1111/nph.17697
3. Ferrol N, Lanfranco,L. (2020) Arbuscular mycorrhizal fungi: methods and protocols. New York, Springer US. https://doi.org/10.1007/978-1-0716-0603-2?nosfx=y
4. McGonigle TP, Miller MH, Evans DG, et al (1990) A new method which gives an objective measure of colonization of roots by vesicular-arbuscular mycorrhizal fungi. New Phytol 115:495–501. https://doi.org/10.1111/j.1469-8137.1990.tb00476.x
5. Nesbit M, Mamo, JC, Majimbi M, Lam, V, Takechi R. (2021). Automated quantitative analysis of *ex vivo* blood-brain barrier permeability using intellesis machine-learning. Front Neurosci, 15, 617221. https://doi.org/10.3389/fnins.2021.617221
6. Novero M, Faccio A, Genre A, et al (2002) Dual requirement of the LjSym4 gene for mycorrhizal development in epidermal and cortical cells of Lotus japonicus roots. New Phytol 154:741–749. https://doi.org/10.1046/j.1469-8137.2002.00424.x
7. Rich MK, Nouri E, Courty PE, Reinhardt D (2017) Diet of arbuscular mycorrhizal fungi: bread and butter? Trends Plant Sci 22:652–660 https://doi.org/10.1016/j.tplants.2017.05.008
8. Trouvelot,A, Kough JL, Gianinazzi-Pearson V (1986) Mesure du taux de mycorhization VA d'un système radiculaire. Recherche de méthodes d'estimation ayant une signification fonctionnelle. In: Gianinazzi-Pearson V, Gianinazzi S (Eds) Physiological and Genetical Aspects of Mycorrhizae INRA Press, Paris, 217-221.
9. Vierheilig H, Schweiger P, Brundrett M (2005) An overview of methods for the detection and observation of arbuscular mycorrhizal fungi in roots. Physiol Plantarum https://doi.org/10.1111/j.1399-3054.2005.00564.x
10. Volkenandt T, Freitag S, Rauscher M (2018) Machine learning powered image segmentation. Microsc and Microanal 24:520–521. https://doi.org/10.1017/s1431927618003094
11. Volpe V., Carotenuto G., Berzero C., Cagnina L., Puech-Pagès V., Genre A. (2020) Short chain chito-oligosaccharides promote arbuscular mycorrhizal colonisation in *Medicago truncatula*. Carbohydr. Polym. 229, 115505. https://doi.org/10.1016/j.carbpol.2019.115505

From satellite images to road pavement type: an object-oriented classification approach

Dalle immagini satellitari al tipo di pavimentazione stradale: un approccio di classificazione orientato agli oggetti

Arianna Burzacchi, Matteo Landrò and Simone Vantini

Abstract This research aims at developing an innovative supervised classification method for road pavement which recognizes surface type of road segments starting from satellite images. Roads of unknown surface are labelled as *paved* or *unpaved* by means of new algorithms developed in the field of Object-oriented Data Analysis and using open-source software and data. The proposed approach is proven to be accurate, low-cost, and can be straightforwardly extended from the single case study of Greater Maputo to other cities.

Abstract L'obiettivo di questa ricerca è sviluppare un metodo innovativo di classificazione supervisionata per la superficie di manto stradale a partire dalle immagini satellitari delle strade. La ricerca si inserisce nell'ambito della Object-oriented Data Analysis e fa uso di software e dati open-source per classificare il tipo di pavimentazione stradale come pavimentato (*paved*) o sterrato (*unpaved*). La metodologia proposta risulta essere accurata, a basso costo e scalabile, facilmente replicabile dal caso studio della Grande Maputo all'analisi di altre città.

Key words: Classification; k-nearest neighbours; pixel distribution; object-oriented; satellite images

---

Arianna Burzacchi
MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. e-mail: arianna.burzacchi@polimi.it

Matteo Landrò
MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. Now at SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, USA. e-mail: matteo.landro@polimi.it

Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy. e-mail: simone.vantini@polimi.it

1 Introduction

Road infrastructure conditions could be very informative for road maintenance, route planning and transport operations [4, 11]. However, road network databases rarely include the information on road pavement status, especially in developing countries. Within this context, the research aims at developing an automatic classification method for road pavement surfaces using an object-oriented approach.

While much research on image-based pavement classification aims at detecting pavement distress [e.g. 10], only few related studies have addressed surface type classification. Common data sources are panoramic images from private databases and Google Street View imagery, which however lack in most developing countries [2], while further analysis is performed by means of feature-based extraction and classification methods [e.g. 6, 12] and neural networks [e.g. 11].

The proposed method works with open-source and always available data, namely Google Earth satellite images [3] and map features exported from the OpenStreetMap (OSM) road network [8]. Among all OSM attributes, the classification algorithm considers the road surface type, indicating whether a road is *paved* or otherwise *unpaved* [9].

The object-oriented approach is preferred to pixel-based methods since it was proven to be more accurate and high-performing [5, 7, 13]. Indeed, object-oriented image-based algorithms take into account both spectral information of each pixel and features of pixel distribution, such as shape and size. In this study, "objects" to be considered are groups of street pixels in the RGB space, extracted from each image and used as sample units of the classification algorithm.

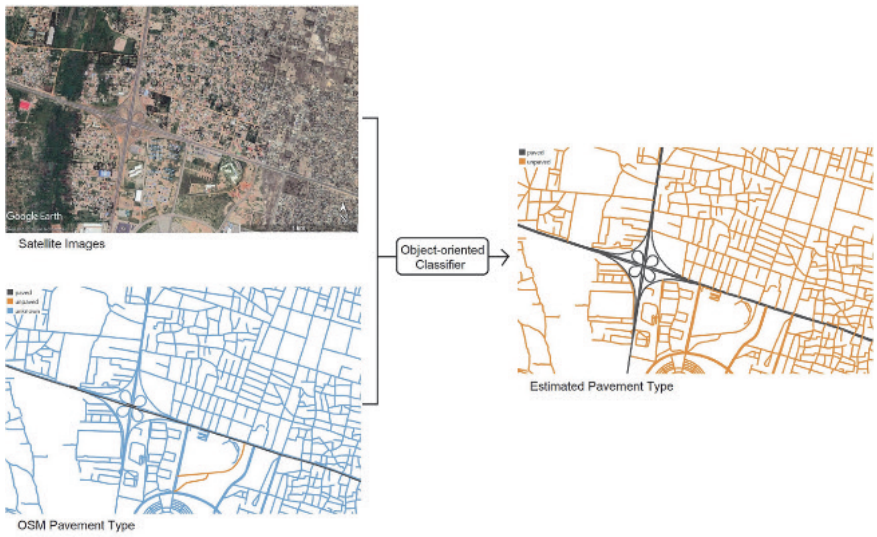


Fig. 1 Input data (left) and output pavement classification (right) of the object-oriented classifier.

2 Outline of the methodology

The first part of the study focuses on street pixel extraction from satellite images. Most images of the dataset portray both street surface and additional objects, such as vehicles, vegetation and buildings. For this reason, a two-step filtering phase is needed to remove uninformative pixels from each image. In the first step, the shade of covering vegetation is identified to set a threshold on pixel darkness. Then, dark pixels associated to vegetation are removed. During the second step, street pixels are identified by means of DBSCAN with properly-tuned parameters. The density-based method DBSCAN is chosen among other clustering algorithm due to the peculiar shape of pixel distributions in the RGB space, namely pixels of streets, buildings and vehicles form distinct clusters connected by background noise.

The following part of the research aims at building an object-oriented classification algorithm for road pavement. Street pixel clusters are input instances of a k-nearest neighbours algorithm which labels roads of unknown surface based on the k most similar pixel clusters of known surface. More formally, each street pixel object is considered as a random sample from a multivariate distribution and, given a distance among probability distributions, the k closest distributions are found and used for classification.

In this second part, the focus is on the choice of the best mathematical embedding for the object-oriented approach. Four metrics between multivariate distributions are compared by evaluating classification performance and computational time of k-NN. Moreover, data are transformed from RGB to alternative colour spaces to investigate which is the most suitable for this application.

The third part of the research focuses on the choice of proper frequency thresholds for k-NN classification. Considering the real-world context of application of the proposed algorithm, misclassification and uncertainty costs are assigned to possible outcomes of k-NN. Then, thresholds are chosen through the minimization of the total predicted cost.

Finally, road type information is exploited to improve the classification of road pavement surface. Road type is an additional attribute provided by OSM to measure the importance of roads in road network systems [9]. The proposed methodology is slightly modified to incorporate this information. The neighbour research, which was previously performed on the set of all streets, is now restricted to the set of streets which have the same typology of the one to be classified.

3 Application and main results

The proposed classification algorithm is applied to the case study of Greater Maputo, in Mozambique. The dataset comprises cropped and filtered raster images of 53 240 road segments and their OSM attributes, and accounts only 5% roads of known pavement surface. Starting from known surface roads, the algorithm predicts pavement surface of the remaining 95%.

The method shows an overall classification error of 7.8% on the test set, which decreases to 5.5% when the street type attribute is included in the model. A detailed description of the method and its application can be found in [1].

References

1. A. Burzacchi, M. Landrò, and S. Vantini. Object-oriented classification of road pavement type in Greater Maputo from satellite images. Technical report, 2022.
2. Google. Google street view. https://www.google.com/intl/it/streetview/, 2021. Accessed: 2021-09-01.
3. Google Earth. http://www.google.com/earth/index.html, 2020. Accessed: 2020-04-20.
4. R. Iles. *Public transport in developing countries*, chapter Transport Infrastructure. Emerald, 2005.
5. D. Liu and F. Xia. Assessing object-based classification: advantages and limitations. *Remote Sensing Letters*, 1(4):187–194, 2010.
6. S. Marianingsih, F. Utaminingrum, and F. A. Bachtiar. Road surface types classification using combination of k-nearest neighbor and naïve bayes based on glcm. *International Journal of Advances in Soft Computing and its Applications*, 11(2), 2019.
7. H. Matinfar, F. Sarmadian, S. Alavi Panah, and R. Heck. Comparisons of object-oriented and pixel-based classification of land use/land cover types based on Lansadsat7, Etm+ spectral bands (case study: arid region of Iran). *American-Eurasian Journal of Agricultural & Environmental Sciences*, 2(4):448–456, 2007.
8. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org , 2020.
9. OpenStreetMap Wiki. https://wiki.openstreetmap.org, 2021. Accessed: 2021-08-21.
10. A. Ragnoli, M. R. De Blasiis, and A. Di Benedetto. Pavement distress detection methods: A review. *Infrastructures*, 3(4), 2018.
11. A. Riid, D. L. Manna, and S. Astapov. Image-based pavement type classification with convolutional neural networks. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, pages 55–60. IEEE, 2020.
12. V. Slavkovikj, S. Verstockt, W. De Neve, S. Hoecke, and R. Van de Walle. Image-based road type classification. In *2014 22nd International Conference on Pattern Recognition*, pages 2359–2364. IEEE, 2014.
13. M. S. Tehrany, B. Pradhan, and M. N. Jebuv. A comparative assessment between object and pixel-based classification approaches for land use/land cover mapping using SPOT 5 imagery. *Geocarto International*, 29(4):351–369, 2014.

Valid inference for group analysis of functionally aligned fMRI images

Inferenza valida per l'analisi di gruppo di immagini fMRI allineate funzionalmente

Angela Andreella, Riccardo De Santis, and Livio Finos

Abstract Functional magnetic resonance imaging (fMRI) data require preprocessing steps before statistical analysis. Multi-subjects fMRI studies are complicated: the brain's anatomical and functional structure varies across subjects. Anatomical alignment does not capture the functional variability across subjects; the functional alignment is then applied. Generally, group analysis on functionally aligned fMRI data refers to between-subject classification. We propose an inference group analysis arguing that using functional aligned images based on Procrustes transformation does not affect type I error.

Abstract *I dati di risonanza magnetica funzionale (fMRI) necessitano di fasi di pre-elaborazione prima dell'analisi statistica. Gli studi fMRI multi-soggetto sono complessi: la struttura anatomica e funzionale del cervello varia tra i soggetti. L'allineamento anatomico non cattura la variabilità funzionale tra i soggetti, si applicano dunque metodi di allineamento funzionale. Generalmente, ci si riferisce ad analisi di classificazione quando si parla di studi multi-soggetto su dati allineati funzionalmente. In questo lavoro proponiamo un'analisi di gruppo inferenziale sostenendo che utilizzare immagini allineate funzionalmente tramite metodi basati sulla trasformazione di Procuste non influisce sull'errore di tipo I.*

Key words: Procrustes method, ProMises model, fMRI group analysis, fMRI data

---

Angela Andreella
Department of Economics, University Ca' Foscari Venezia, Italy
e-mail: angela.andreella@unive.it

Riccardo De Santis
Department of Statistics, University of Padova, Italy
e-mail: riccardo.desantis.1@phd.unipd.it

Livio Finos
Department of Developmental Psychology and Socialization, University of Padova, Italy
e-mail: livio.finos@unipd.it

1 Introduction

Functional Magnetic Resonance Image (fMRI) is the most used technique to study the neural underpinnings of human cognition. The brain activation is expressed in the correlation between the sequence of cognitive stimuli and the sequence of measured blood oxygenation levels (BOLD). In reaction to neural activity, changes in brain hemodynamics impact the local intensity of the magnetic resonance signal, i.e., the voxel intensity (single volume elements). The voxel expresses statistically significant neural activity if the correlation computed is statistically significant concerning an inactive region null hypothesis. However, when group analysis is performed, many problems arise.

First, brain activity is not functionally nor anatomically aligned across subjects since brains' anatomical and functional structures vary across subjects, even in response to identical sensory inputs. Anatomical normalization (e.g., [13]) fixes the anatomical misalignment; but it fails to capture the functional variability across subjects. Second, analyzing group activation means performing several statistical tests equal to the number of voxels (i.e., roughly $200,000$ statistical tests). It leads to multiple testing problems, and family-wise error rate must be controlled.

In this paper, we proposed a method that simultaneously resolves the functional misalignment and gains power in group analysis inference without affecting the type I error. The functional alignment is done by the ProMises model [2] which can be seen as a procedure that sorts the null hypotheses based on a priori information. The brain images' functional alignment can be described as a sorting criterion independent from the test statistics, making the inference on aligned data completely valid. This procedure is not new in the literature; some methods use, for example, previous findings in similar experiments or the total variance of the variables if central location tests are used [5].

The outline of the paper is as follows. Subsect. 2.1 introduces the ProMises model [2], while Subsect. 2.2 shows the group analysis procedure in fMRI data. The valid inference using functionally aligned data is described in Subsect. 2.3. Finally, the ProMises model is evaluated by performing a group-level inference, analyzing task-related fMRI data in Sect. 3. The entire code used in this work is available in [3].

2 Methods

2.1 ProMises model

Let $\{X_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$ a set of matrices to be aligned and $M \in \mathbb{R}^{n\times m}$ the shared unknown matrix. The ProMises model uses similarity transformation [7], i.e., scaling, rotation/reflection, and translation, to map $\{X_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$ into a common reference matrix M. The model is defined as follows:

$$X_i = \alpha_i(M + E_i)R_i^\top + 1_n^\top t_i \tag{1}$$

where E_i is the random error matrix with a normal matrix distribution [8] $E_i \sim \mathscr{MN}_{nm}(0, \Sigma_n, \Sigma_m)$, R_i is the orthogonal matrix parameter with a von Mises-Fisher distribution [4] (i.e, $f(R_i) \sim C(F,k)\exp(kFR_i)$ with $F \in \mathbb{R}^{m\times m}$ location parameter, $k \in \mathbb{R}^+$ regularization parameter and $C(F,k)$ normalizing constant), α_i is the scaling parameter and t_i is the translation parameter with 1_n a n-dimensional vector of ones. The maximum a posteriori estimates for the sets $\{R_i\}_{i=1,\dots,N}$ and $\{\alpha_i\}_{i=1,\dots,N}$ equal

$$\hat{R}_i = \{U_i V_i^\top\}_{i=1,\dots,N}; \quad \hat{\alpha}_{i\hat{R}_i} = \frac{||\Sigma_m^{-1/2}\hat{R}_i^\top X_i^\top \Sigma_n^{-1/2}||_{\mathrm{F}}^2}{\mathrm{tr}(D_i)} \quad \forall i \in \{1,\dots,N\}, \tag{2}$$

where $||\cdot||_{\mathrm{F}}$ is the Frobenius norm, and $U_i D_i V_i^\top$ is the singular value decomposition of $X_i^\top \Sigma_n^{-1} M \Sigma_m^{-1} + kF$. For further details about the estimation process, see [2].

2.2 Group-level analysis

Let consider the set $\{X_i \in \mathbb{R}^{n\times m}\}_{i=1,\dots,N}$. After the X_i matrices' functional alignment (e.g., using the ProMises model explained in Subsection 2.1), the first-level analysis (i.e., subject-level analysis) is performed to find significant activation in the set of m voxels for each subject under the null hypothesis of no activation. Therefore, we consider for each subject i the following model:

$$X_i = DB_i + ZG_i + E_i \tag{3}$$

where $D \in \mathbb{R}^{n\times p}$ and $Z \in \mathbb{R}^{n\times q}$ are fixed matrices, $B_i \in \mathbb{R}^{p\times m}$, $G_i \in \mathbb{R}^{q\times m}$ and $B_i = B + U_i$, $G_i = G + g_i$ with B is the true matrix of fixed effects of interest, G of fixed nuisance effects. $[U_i^\top | g_i^\top] \sim \mathscr{MN}(0, \Sigma_m, \Sigma_{pq})$ is the matrix of random effects.

We have now a set of N matrices $\{\hat{B}_1, \dots, \hat{B}_N; \hat{B}_i \in \mathbb{R}^{p\times m}\}$, which for example describes the difference between the neural activation during two stimuli recorded in the voxel $k \in \{1,\dots,m\}$ of the subject i. The one-sample t-test [11] is generally performed to study the group's mean activation due to the difference between the neural activation during two stimuli:

$$T = \frac{\hat{\mu}}{\hat{\sigma}/\sqrt{N}}, \tag{4}$$

where $\hat{\mu} = \sum_{i=1}^N \hat{B}_i/N$ is the sample mean between-subjects with $\hat{\mu} \in \mathbb{R}^{p\times m}$, and $\hat{\sigma} = \frac{1}{N-1}\sum_{i=1}^N (\hat{B}_i - \hat{\mu})^2$ is the sample variance between-subjects with $\hat{\sigma} \in \mathbb{R}^{p\times m}$. Therefore, we have one local test T_k for each voxel/null hypothesis $H_0^k : \mu_k = 0$ against the two-sided alternative hypothesis.

In the ProMises model (1) we assume M equals for all subjects, while in (3) the mean effect is subject-specific, that is $M_i = DB_i = DB + DU_i = M + DU_i$. The

additional random matrix DU_i must be then inserted in (1). However, $\mathbb{E}(U_i) = 0$ by definition, and $\mathbb{E}(M_i) = M$. We assume that under H_0 the random effect involved in B_i is null (i.e., these terms does not impact the estimation process of R_i).

2.3 Valid Inference

Here, we propose a theorem that ensures the validity of inference using functionally aligned data.

Theorem 1. *Let consider the p-values p_k related to the statistical test T_k, where $k \in M = \{1, \dots, m\}$. If the ProMises model defined in Subsection* (2.1) *is valid, then:*

$$\Pr(p_k \leq \alpha \mid \hat{R}_i, \hat{\alpha}_i) = \alpha, \quad \forall k \in S \tag{5}$$

where $S \subseteq M$ is the set of true null hypotheses, α is the significance level, and $\hat{R}_i$ defined as (2)*.*

Proof. Let assume that the the ProMises model defined in Equation (1) is valid, then we have:

$$X_i = \alpha_i(M_i + E_i)R_i^\top = \alpha_i(DB_i + ZG_i + E_i)R_i^\top. \tag{6}$$

W.l.o.g. we assume the set $\{X_i\}_{i=1,\dots,N}$ to be column-centered matrices. Write $H_Z = Z(Z^\top Z)^{-1}Z^\top$, so $\hat{B}_i = (D^\top(I - H_z)D)^{-1}D^\top(I - H_z)X_i$. The ProMises model becomes $X_i = \alpha_i(ZG_i + E_i)R_i^\top$ under H_0. It is now easy to see that the $\hat{B}_i \mid \hat{R}_i, \hat{\alpha}_i$ is normal zero-centered under H_0 like $\hat{B}_i$ with variance Σ. In addition, the information involved in the estimation of R_i and B_i are orthogonal (i.e., $\mathbb{E}((ZG_i + E_i)^\top(I - H_z)\hat{X}_i) = tr(\Sigma_n^\top(I - H_Z))\Sigma_m^\top$ which is independent of D and other subjects's information, and $X_{j \neq i}$ are independent of $(I - H_z)X_i$ by definition). This implies that also the statistical test $T \mid \hat{R}_i, \hat{\alpha}_i$ defined in (4) is normal zero-centered and finally $\Pr(p_k \leq \alpha \mid \hat{R}_i, \hat{\alpha}_i) = \alpha \; \forall k \in N$.

Theorem 1 assures the validity of the group analysis inference when data functionally aligned by the ProMises model are used instead of the raw data. In the next section, we apply the ProMises model to fMRI images and then we perform one-sample t-tests for each voxel to analyze the group mean activation. The gain in power is notable if the functionally aligned data are used instead of the raw ones.

3 Application

3.1 Data

The auditory data collected by [12] are analyzed, available at `https://openneuro.org/datasets/ds000158/versions/1.0.0`. The study consists

of neural activation of 218 subjects passively listening to vocal, i.e., speech, and nonvocal sounds. We randomly select 18 subjects.

We preprocess the data using the FMRIB Software Library (FSL) [10] using a standard processing procedure. For further details about the experimental design and data acquisition, please see [12]. We perform a higher-level analysis ROI, i.e., group-subject analysis [11], considering the superior temporal gyrus (STG) as ROI, which is well known to be involved in auditory processing [6]. The STG was extracted from the Harvard Oxford cortical structural atlas (`http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases`). The matrices are then composed of 310 rows (time points/stimuli) and 10233 columns (voxels). The ProMises model is implemented in [3] based on the programming language Python [14] according to the PyMVPA package [9].

3.2 Results

Using the ProMises model, we found notable results for group analysis of fMRI data. The left panel of Figure 1 illustrates the one-sample t-tests (4) computed using the images aligned by the anatomical alignment [13] and by the ProMises model. The right panel of Figure 1 shows the empirical cumulative distribution function of these one-sample t-tests. The ProMises model returns a set of t-tests 85.85% higher in mean (in absolute value) than the t-tests computed using anatomical alignment. In the case of equivalent t-tests, the percentage would equal 50%.

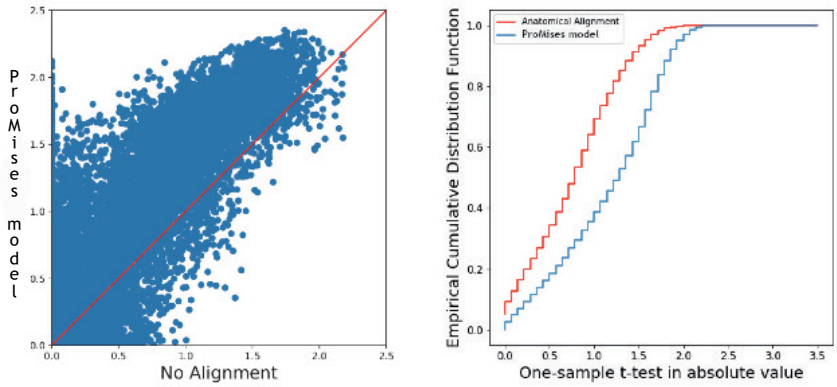


Fig. 1 Left panel: Scatter plot of the one-sample t-tests in absolute value considering the fMRI images of 18 subjects aligned by anatomical alignment (x-axis) and the ProMises model (y-axis). **Right panel**: Empirical Cumulative Distribution Function of one-sample t-tests in absolute value considering the fMRI images of 18 subjects aligned by anatomical alignment (red line) and the ProMises model (blue line).

4 Discussion

In this manuscript, we evaluated the ProMises model in the context of group-level inference analysis of fMRI data. We proved the validity of the inference if functionally aligned data are used instead of the raw ones (where only the anatomical registration is performed). The information used in the ProMises model to estimate the orthogonal matrix parameter can be seen as prior information used to sort the statistical tests in the group analysis. Finally, we showed an interesting application where the one-sample t-tests at the group level using data aligned by the ProMises model illustrate higher absolute values than the one-sample t-tests computed using data aligned by the anatomical alignment.

References

1. Basso, D. and Finos, L. (2012) Exact multivariate permutation tests for fixed effects in mixed-models. *Communications in Statistics-Theory and Methods* 41(16-17), :2991–3001.
2. Andreella, A. and Finos, L. (2022). Procrustes analysis for high-dimensional data. *Psychometrika (accepted).* DOI: 10.1007/s11336-022-09859-5.
3. Andreella, A. (2020). ProMises (Procrustes von Mises-Fisher) model, GitHub repository. DOI: 10.5281/zenodo.5095114.
4. Downs, D. T. (1972). Orientation statistics. *Biometrika*, 59(3):665.
5. Farcomeni, A. and Finos, L. (2013). FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. *Biometrics*, 69(3):606–613.
6. Galaburda, A. and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. *The Journal of Comparative Neurology*, 190(3):597–610.
7. Gower, J. C. and Dijksterhuis, B. G. (2004). *Procrustes problems*. Oxford University Press.
8. Gupta, A. K. Nagar, D. K. (2018). *Matrix variate distributions*. Chapman and Hall/CRC.
9. Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, (7):37–53.
10. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage*, 62:782–90.
11. Mumford, J. A. and Poldrack, R. (2007). Modeling group fMRI data. *Social Cognitive and Affective Neuroscience*, 62(2):251–257.
12. Pernet, C. R., McAleer, P., M., L., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., and Belin, P. (2015). The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage*, 119:164–74.
13. Talairach, J. J. and Tournoux, P. (1988). *Co-Planar stereotaxic stlas of the human brain.3-dimensional proportional system: an approach to cerebral imaging*. Atlante.
14. Van Rossum, G. and Drake Jr, F. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Topological persistence for astronomical image segmentation

Segmentazione di immagini astronomiche mediante strumenti topologici

Ceccaroni, R., Brutti, P., Castellano, M., Fontana, A. and Merlin, E.

Abstract The fast evolution of telescope technologies is making possible the collection of a massive wide variety of data. In our era, a tool that automates information extraction from astronomical images is essential. Within this broad task, image segmentation plays a key role and classical edge-based detection algorithms are not well-suited to deal with astronomical images because they typically lack a clear-cut boundary structure.
Thus, to effectively tackle this task, it is mandatory to develop dedicated tools.
The main goal of this work is to design and test a new, unsupervised segmentation method strongly based on Topological Data Analysis (TDA) techniques. Thanks to tools like persistent homology and persistence diagrams, in fact, it is possible to identify the connected components of abstract objects, like an image, and then put them to use in order to compute a sensible segmentation.
Abstract La rapida evoluzione delle tecnologie dei telescopi sta rendendo possibile la raccolta di un'ampia varietà di dati. Nella nostra era è essenziale uno strumento che automatizza l'estrazione di informazioni dalle immagini astronomiche. All'interno di questo ampio compito, la segmentazione dell'immagine gioca un ruolo chiave. Gli algoritmi classici sono basati sul rilevamento dei bordi e non sono adatti per le immagini astronomiche poiché in quest'ultime gli oggetti non hanno dei confini ben definiti.

Riccardo Ceccaroni
Sapienza University of Rome, e-mail: ceccaroni.1884368@studenti.uniroma1.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

Marco Castellano
Osservatorio Astronomico di Roma-INAF, e-mail: marco.castellano@inaf.it

Adriano Fontana
Osservatorio Astronomico di Roma-INAF, e-mail: adriano.fontana@inaf.it

Emiliano Merlin
Osservatorio Astronomico di Roma-INAF, e-mail: emiliano.merlin@inaf.it

Pertanto, per affrontare efficacemente questo compito, è stato necessario lo sviluppo di strumenti adeguati.
L'obiettivo principale di questo lavoro è progettare e testare un nuovo metodo di segmentazione non supervisionato fortemente basato sulle tecniche di analisi dei dati topologici (TDA). Grazie a strumenti come l'omologia persistente e ai diagrammi di persistenza, infatti, è possibile identificare le componenti connesse di oggetti astratti, come un'immagine, ed utilizzarle per calcolare una segmentazione apprezzabile.

Key words: Topological data analysis, Persistence diagram, Image segmentation, Astronomical imaging

1 Introduction

Modern astronomy has made relevant strides in the last 60 years with increasingly advanced technologies. Since 1965, a growing number of space telescopes have been launched into Earth orbit. In the same years, Willard Boyle and George E. Smith invented the Charge-Coupled Device (CCD) [7]. This electronic detector is still used today and marks the transition from analog to digital technology. Thanks to this device, it is possible to capture digital images and store precise and rigorous information about astronomical objects.

Not so long ago, the James Webb Space Telescope (JWST) was successfully launched. It is intended to succeed the Hubble Space Telescope. Other space telescopes will be launched into orbit in the coming years. Among them, there is Euclid: a visible to near-infrared space telescope currently under development by the European Space Agency (ESA).

Like observation, the analysis of astronomical images has also undergone improvements. Initially, objects were identified and classified by eye. Today, in the Big Data era, the need for time-saving and accurate tools to detect and analyze objects in images dominated by observational noise has significantly increased.

In this work, we present a new tool aimed at the so-called *image segmentation* which is part of the detection step: in the segmentation process, the significative (i.e., above the noise level) pixels of the digital image are univocally assigned to the sources, defining their extension. This new tool is based on TDA. In particular, we use persistent homology. [9, 5, 3, 12].

2 Materials and Methods

In the following part we firstly provide details of used data. Then, we describe our segmentation algorithm (TIS). Finally the metrics used to compare the outputs are analysed.

2.1 Data

We test the source-detection algorithms on simulated data. The simulated image (Figure 1) has a size of $25,000 \times 25,000$ pixels. It was produced using the software Galsim [11] and it is a simulation for Euclid. The image simulates a realistic scenario in which there are overlapping galaxies: a deep wide-field observation like the one obtained with Euclid's VIS filter. Of this image is known the ground truth.

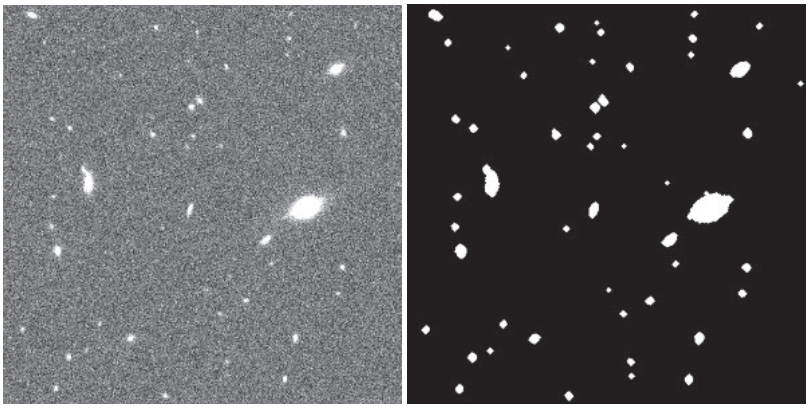


Fig. 1: On the left is a patch of size 500×500 of of the simulated image, on the right the segmentation using TIS.

2.2 Topological Image Segmentation

In this section, we present the segmentation algorithm for astronomical images based on persistent homology. The astronomical image I together with its *RMS*-map are the inputs. As output, we would get the segmented image S_I, in which each object is identified with an integer (*id*). A catalog containing all the relevant information about each object we found is also generated.

2.2.1 Pre-Processing

Before calculating persistent homology, the image is cleaned. Denoising is done to reduce the noise in the image. The thresholds are calculated and, thanks to the map *RMS*, the damaged pixels are masked.

As a first step, the noise in the image I is reduced by using one of the denoising methods described in [10]. By default, Gaussian Smoothing is used. The window size depends on σ and is $\lfloor 8\sigma + 1 \rfloor$. If $\sigma < 1/4$ then the window has fixed size 3×3. From the convolution, we get $I_{denoise} \coloneqq (G_\sigma \star I)$.

The second step involves estimating the background (bkg) of I that can be executed by taking the median of the values of I. By default $bkg \coloneqq 2.5\ median - 1.5\ mean$.

At this point, a $mask \coloneqq (I_{denoise} - bkg) > t \cdot RMS$ is generated, where t is a parameter, default set at $t = 2$. Then $mask$ is dilated with a $d = 5$ default parameter. The values of the $I_{denoise}$ image are normalized in $[0,1]$. The 0 value is assigned to the damaged pixels using $mask$. Then, we get $I_{pre-process} \coloneqq \text{MinMaxScaler}(I_{denoise}) \cdot mask$.

2.2.2 Persistent Diagram

The next step is the calculation of persistent homology [4] from $I_{pre-process}$. The 0-dim persistence diagram is calculated. The implementation used is called *Cubical Ripser* and is described in [8]. The result is stored in a $(n,6)-$array, where n is the number of objects. Each row consists of

$$\begin{array}{|cccccc|} \hline birth & death & x_1 & y_1 & x_2 & y_2 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline \end{array}$$

where (x_1, y_1) is the location of the birth of the object and (x_2, y_2) is the location of the death of the object in the image $I_{pre-process}$. For further computational details see [9, 5, 1, 8].

At this point, the persistence diagram is filtered to separate topological features from noise using clustering or persistent entropy. The values of $lifetime = birth - death$ are calculated. By default, a k-means is applied with $k = 2$ on the values $\log(lifetime)$. Filtering returns only the indices ($idxs$) of the objects to be analyzed during segmentation.

2.2.3 Segmentation

For each index returned by the filtering, a temporary image $I_{temp} \coloneqq (death \leqslant I_{pre-process} \leqslant birth)$ is generated. In this binary image, we identify the connected

components and select the component that contains (x_1, y_1). This is our segmentation of the object. We have to repeat these steps for all indexes to gain the segmentation of the entire image S_I.

2.3 Metrics

For the comparison of segmentations, we use the metrics proposed in [6]. For *matching detections* we define *detection recall (completeness)* as the proportion of detected objects and *detection precision (purity)* as the proportion of segments matchable to real objects. We use *F-score* as the harmonic mean of precision and recall. For *evaluating areas*, assume that the ground truth segmentation consists of N objects, $\{R_1, \ldots, R_N\}$, with their corresponding areas $\{A_1, \ldots, A_N\}$, whereas the test segmentation is composed by M objects, $\{T_1, \ldots, T_M\}$, with areas $\{a_1, \ldots, a_M\}$. We then define the Under-Merging error (UM) as

$$UM = \sum_{j=1}^{M} \frac{\big(A_k - (T_j \cap R_k)\big)(T_j \cap R_k)}{A_k},$$

where k is the index such that R_k maximise $T_j \cap R_k$ for each test object T_j, and we define the Over-Merging error (OM) as

$$OM = \sum_{k=1}^{N} \frac{\big(a_j - (T_j \cap R_k)\big)(T_j \cap R_k)}{a_j},$$

where j is the index such that T_j maximise $T_j \cap R_k$ for each test object R_k. In the end, we defined an Area score $= 1 - \sqrt{OM^2 + UM^2}$, which combines the two to give a single, overall score.

3 Results

In this section, we report the results obtained in comparing the segmentation output by TIS and SExtractor [2], a well-established and reliable segmentation tool in astronomy. The results were obtained using patches of 500×500 pixels of the simulated image as samples.

Figure 2 shows the range of F-scores produced by TIS and SExtractor using their default settings. For SExtractor, it is notable that the scores have a smaller interquartile range. Overall, focusing on median scores, we see that TIS is performing better than SExtractor with a median score of 0.58, while SExtractor achieves a median score of 0.42. TIS showed a higher variability also in Area scores. However, we see the strongest performance from TIS, with a median score of 0.76. The weakest performance was produced by SExtractor, with a median score of 0.37.

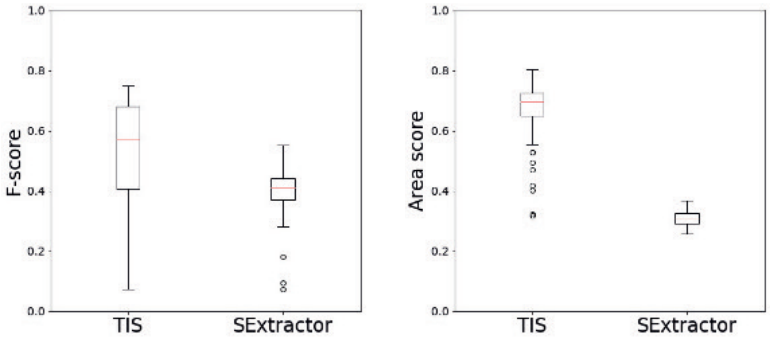


Fig. 2: On the left the F-scores comparison, on the right the Area scores comparison

In conclusion, the proposed TDA based segmentation methods is relatively simple and effective, showing performance on simulated images that are comparable to (if not better than) those obtained by a broadly adopted tool like SExtractor. Going forward, a broader and more in depth comparison is surely needed.

References

1. Bauer, U.: Ripser: efficient computation of vietoris-rips persistence barcodes. Journal of Applied and Computational Topology (2021). DOI 10.1007/s41468-021-00071-5
2. Bertin, E., Arnouts, S.: Sextractor: Software for source extraction. Astronomy and Astrophysics Supplement Series **117** (1999). DOI 10.1051/aas:1996164
3. Boissonnat, J.D., Chazal, F., Yvinec, M.: Geometric and Topological Inference (2017). DOI 10.1017/9781108297806
4. Edelsbrunner, H., Harer, J.: Persistent homology—a survey. Discrete & Computational Geometry - DCG **453** (2008). DOI 10.1090/conm/453/08802
5. Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction (2010). DOI 10.1007/978-3-540-33259-6_7
6. Haigh, C., Chamba, N., Venhola, A., Peletier, R., Doorenbos, L., Watkins, M., Wilkinson, M.: Optimising and comparing source-extraction tools using objective segmentation quality criteria. Astronomy and Astrophysics **645** (2020). DOI 10.1051/0004-6361/201936561
7. HOWELL, S., Tavackolimehr, A.: Handbook of CCD Astronomy (2019)
8. Kaji, S., Sudo, T., Ahara, K.: Cubical ripser: Software for computing persistent homology of image and volume data (2020)
9. Rojas Moraleda, R., Valous, N., Xiong, W., Halama, N.: Computational Topology for Biomedical Image and Data Analysis: Theory and Applications (2019). DOI 10.1201/9780429443077
10. Roscani, V., Tozza, S., Castellano, M., Merlin, E., Ottaviani, D., Falcone, M., Fontana, A.: A comparative analysis of denoising algorithms for extragalactic imaging surveys. Astronomy and Astrophysics **643** (2020). DOI 10.1051/0004-6361/201936278
11. Rowe, B., Jarvis, M., Mandelbaum, R., Bernstein, G., Bosch, J., Simet, M., Meyers, J., Kacprzak, T., Nakajima, R., Zuntz, J., Miyatake, H., Dietrich, J., Armstrong, R., Melchior, P., Gill, M.: Galsim: The modular galaxy image simulation toolkit. Astronomy and Computing **10** (2014). DOI 10.1016/j.ascom.2015.02.002
12. Zomorodian, A.J.: Topology for Computing. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2005). DOI 10.1017/CBO9780511546945

Statistical assessment and empirical estimation

Confidence regions for optimal sensitivity and specificity of a diagnostic test

Regioni di confidenza per sensibilità e specificità ottimali di un test diagnostico

Gianfranco Adimari, Duc-Khanh To and Monica Chiogna

Abstract We propose new methods that provide approximate joint confidence regions for the optimal sensitivity and specificity of a diagnostic test, fixed by the Youden index criterion. Such methods are semiparametric and overcome limitations of alternative approaches available in the literature. Our proposal is based on empirical likelihood pivots and covers two situations: binormal model and binormal model after the use of Box-Cox transformations. In the last case, we show how to use two different transformations, for the healthy and the diseased subjects.

Key words: Empirical likelihood, Box-Cox transformation, Youden index, ROC analysis, bootstrap calibration

1 Introduction and background

The accuracy of a diagnostic test can be assessed by its receiver operating characteristic (ROC) curve. The test result can be dichotomized at a specified cutpoint. Given the cutpoint, the sensitivity is the probability of a true positive, i.e., the probability that the test correctly identifies a diseased subject. The specificity is the probability of a true negative, i.e., the probability that the test correctly identifies a non-diseased subject. When one varies the cutpoint throughout the entire real line, the resulting pairs (1 − specificity, sensitivity) form the ROC curve (see Pepe[1], as general reference). Let X denote the result of a continuous diagnostic test for a non-diseaded subject and Y the result of the test for a diseaded subject. We as-

Gianfranco Adimari, Duc-Khanh To
Department of Statistical Sciences, University of Padova, Via C. Battisti, 241; I-35121 Padova, Italy. e-mail: gianfranco.adimari@unipd.it; e-mail: duckhanh.to@unipd.it

Monica Chiogna
Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via Belle Arti, 41; 40126 Bologna, Italy. e-mail: monica.chiogna2@unibo.it

sume, without loss of generality, that high test values indicate high likelihood of disease. Then, for a given cutpoint τ, the sensitivity and the specificity of the test are $\theta(\tau) = \Pr(Y > \tau) = 1 - F_Y(\tau)$ and $\eta(\tau) = \Pr(X \leq \tau) = F_X(\tau)$, respectively, where $F_X(\cdot)$ denotes the cumulative distribution function of X and $F_Y(\cdot)$ the cumulative distribution function of Y. In practice, for the purpose of making diagnosis, that is, classifying a subject as either diseased or healthy, a diagnostic threshold is required. In order to select an "optimal" diagnostic cutpoint, there exists a variety of approaches. Among them, the one based on the Youden index [2], J, is certainly the most popular. The Youden index is the maximum of the sum of sensitivity and specificity minus one, i.e., $J = \max_\tau J(\tau)$, with $J(\tau) = \theta(\tau) + \eta(\tau) - 1$. The corresponding optimal cutpoint, τ^+ say, is the value that maximizes $J(\tau)$, i.e., the cutpoint that has the largest value in the associated sum of sensitivity and specificity. Clearly, sensitivity and specificity at the optimal cutpoint τ^+ are relevant measures for the diagnostic ability of the test.

Actually, the ROC curve of a new diagnostic test is unknown, being unknown (at least partially) the distributions of the test results, for both diseased and non-diseased populations. Hence, the statistical evaluation of the discriminatory ability of the test is obtained by making inference about its ROC curve and other quantities of interest, such as optimal thresholds and associated sensitivities and specificities. Generally, inference is based on data from some suitable sample of patients for which the disease status can be exactly assessed by means of a so-called gold standard test (GS). When an optimal threshold is estimated from data of both diseased and healthy samples, the corresponding estimated sensitivity and specificity are correlated, and joint inference is necessary to take into account such a correlation. Methods to built joint confidence regions for sensitivity and specificity at the optimal cutpoint based on the Youden index are proposed by Bantis et al.[3] and Yin and Tian[4]. Both works deal, in particular, with parametric (or semi-parametric) methods for the binormal model. When the hypothesis of normality is not adequate, in both articles it is suggested to resort, when possible, to a single Box-Cox transformation. The main limits of such methodologies are: (i) the related confidence regions have (or derive from regions that have) elliptical shape, because they are based on the asymptotic normality of appropriate estimators or pivots; (ii) when the normality of the biomarker does not meet, application of Box-Cox transformations is considered, but limited to a single transformation for both populations.

We propose new methods that provide approximate joint confidence regions for the sensitivity and specificity of a test, corresponding to the optimal cutpoint fixed by the Youden index criterion. They are based on empirical likelihood pivots, so give rise to likelihood-type regions that have no predetermined constraints on the shape and are automatically range respecting. The proposal covers binormal model and binormal model after the use of Box-Cox transformations. Importantly, in the second case, we show how to use two different transformations, for the healthy and the diseased subjects.

Background Let $x_1, \ldots, x_m$ be a random sample from X, i.e., the test results from m non-diseased patients, and $y_1, \ldots, y_n$ a random sample from Y, i.e., the test results from n diseased patients. The true disease status of each patient is assumed to

be ascertained by a GS test. Let $\hat{F}_X$ denote the empirical distribution function based on $x_1,\ldots,x_m$ and $\hat{F}_Y$ the empirical distribution function based on $y_1,\ldots,y_n$. Hence, for a fixed value t, $\hat{F}_X(t) = (1/m)\sum_{i=1}^{m} I(x_i \le t)$, where $I(\cdot)$ denotes the indicator function. Recall that we consider a continuous diagnostic test and that, for a fixed threshold τ, $\theta = \theta(\tau) = 1 - F_Y(\tau)$ and $\eta = \eta(\tau) = F_X(\tau)$. Adimari and Chiogna[5] define the empirical likelihood [6] statistic

$$\ell(\theta,\eta,\tau) = 2m\left\{\hat{F}_X(\tau)\log\frac{\hat{F}_X(\tau)}{\eta} + \left[1-\hat{F}_X(\tau)\right]\log\frac{1-\hat{F}_X(\tau)}{1-\eta}\right\} + 2n\left\{\hat{F}_Y(\tau)\log\frac{\hat{F}_Y(\tau)}{1-\theta} + \left[1-\hat{F}_Y(\tau)\right]\log\frac{1-\hat{F}_Y(\tau)}{\theta}\right\},$$

for $\theta \in (0,1)$, $\eta \in (0,1)$, $\tau \in \mathscr{T}$, where $\mathscr{T} = \left[\max\{x_{(1)},y_{(1)}\}, \min\{x_{(m)},y_{(n)}\}\right)$, and $x_{(i)}$, $i = 1,\ldots,m$, and $y_{(j)}$, $j = 1,\ldots,n$, denote the order statistics from the samples. When $\tau \notin \mathscr{T}$, then $\ell(\theta,\eta,\tau) = +\infty$. The Authors prove that, under very weak conditions, for each triplet τ_0, $\theta_0 = 1 - F_Y(\tau_0)$ and $\eta_0 = F_X(\tau_0)$ of true parameter values, when $\min\{m,n\} \to +\infty$ and $\lim m/n$ is finite and not zero, $\ell(\theta_0,\eta_0,\tau_0) \xrightarrow{d} \chi^2_2$, where χ^2_2 indicates the chi-squared distribution with 2 df. Then, such result can be used to construct non-parametric confidence regions for a pair of parameters, for example (θ_0,η_0), when the third remaining parameter is fixed: the set $\mathscr{R}_\alpha = \{(\theta,\eta) : \ell(\theta,\eta,\tau_0) \le c_\alpha\}$, where $\alpha \in (0,1)$ and c_α is such that $\Pr(\chi^2_2 > c_\alpha) = \alpha$, is a confidence region with nominal coverage probability $1-\alpha$ for the pair (sensitivity, specificity), at a fixed cutt-off τ_0. Regions obtained by $\ell(\theta,\eta,\tau_0)$ retain all good features of empirical likelihood confidence regions, whose shape and orientation is determined only by the data, and are range respecting.

2 The proposed method

Let τ^+ be the true optimal cutpoint based on the Youden index approach, and let θ^+ and η^+ be the corresponding values of sensitivity and specificity. If τ^+ was known, the pivot $\ell(\theta^+,\eta^+,\tau^+)$ could be used directly to build approximate confidence regions for (θ^+,η^+). In practice, τ^+ is unknown and must be estimated from the available data. Let $\hat{\tau}^+$ be a suitable estimator for τ^+. By resorting to the plug-in method, we can consider the quantity $\ell(\theta^+,\eta^+,\hat{\tau}^+)$, where an estimate replaces an unknown value. Unfortunately, such a replacement is not untroublesome. Indeed, the standard χ^2 approximation is no longer applicable in such a case. It is well known in the literature that when estimated entities enter in an empirical likelihood pivot, this has an asymptotic distribution which is a linear combination of independent χ^2_1 random variables (see, for example, Hjort et al.[9]), whose coefficients are (unknown) eigenvalues of a suitable 2×2 matrix. Since accurate estimation of such coefficients can be somewhat challenging, in the following we resort to the bootstrap calibration and apply it to the distribution of $\ell_*(\theta^+,\eta^+) = \ell(\theta^+,\eta^+,\hat{\tau}^+)$. This allows us to obtain estimates of the quantiles of the distribution of $\ell_*(\theta^+,\eta^+)$, which we will use to define the desired confidence regions.

Binormal model In some situations it is reasonable to assume that the diagnostic test has normal distribution, in both populations of healthy and diseased subjects. Let $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. To formalize the assumption that the values of the test are positively associated with the disease status, we impose hereafter that $\mu_y > \mu_x$. For the binormal model, the optimal threshold provided by the criterion based on the Youden index can be obtained analytically and results to be

$$\tau^+ = \frac{\mu_x((\sigma_y/\sigma_x)^2-1) - (\mu_y-\mu_x) + (\sigma_y/\sigma_x)\sqrt{(\mu_y-\mu_x) + ((\sigma_y/\sigma_x)^2-1)\sigma_x^2 \log((\sigma_y/\sigma_x)^2)}}{(\sigma_y/\sigma_x)^2-1}.$$

When variances are equal, i.e. $\sigma_x^2 = \sigma_y^2$, then $\tau^+ = (\mu_x+\mu_y)/2$. Then, by the plug-in principle, a consistent estimator of τ^+ can be obtained by substituting in the above expressions the empirical counterparts $\bar{X} = (1/m)\sum_{i=1}^m X_i$, $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$, $S_x = \sqrt{(1/(m-1))\sum_{i=1}^m (X_i-\bar{X})^2}$ and $S_y = \sqrt{(1/(n-1))\sum_{i=1}^n (Y_i-\bar{Y})^2}$ of μ_x, μ_y, σ_x and σ_y, respectively. Observe that the resulting $\hat{\tau}^+$ estimator is, essentially, the maximum likelihood estimator of τ^+.

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we compute the estimate $\hat{\tau}^+$ and the corresponding optimal estimated sensitivity $\hat{\theta}^+ = 1 - \Phi((\hat{\tau}^+ - \bar{y})/S_y)$ and specificity $\hat{\eta}^+ = \Phi((\hat{\tau}^+ - \bar{x})/S_x)$. To conveniently calibrate $\ell_*(\theta^+, \eta^+)$ we resort to a simple bootstrap procedure:

1. use parametric bootstrap to get B bootstrap samples $\{x\}_b$ and $\{y\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original sample, so as to obtain "enlarged" bootstrap samples of size $m+2$ and $n+2$, respectively;
3. compute $(\bar{x}_b, \bar{y}_b, S_{xb}, S_{yb})$, $\hat{\tau}_b^+$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ from the b-th pair of (enlarged) bootstrap samples, where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimates obtained from the original data and the empirical distributions are taken from bootstrap samples;
4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, $b = 1, \ldots, B$.

Then, the set $\mathscr{R}_\alpha = \{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\}$, is a confidence region, with nominal coverage probability $1-\alpha$, for the optimal pair (sensitivity, specificity), corresponding to the Youden index criterion. In the above presented bootstrap procedure, at step 2, we "enlarged" bootstrap samples in order to reduce effects of the so-called convex hull problem, i.e., to reduce the probability that $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ may be not finite. Moreover, we process only (enlarged) bootstrap samples whose averages respect the fixed order ($\mu_y > \mu_x$).

Normal models after Box-Cox transformations Suppose now that distributions of X and Y cannot reasonably be considered normal. Moreover, assume that $X > 0$ and $Y > 0$, with probability 1. In such a situation, Box-Cox transformations can help to achieve normality and are frequently used in ROC studies [3, 7]. However, to the best of our knowledge, papers available in the literature discuss, in this regard, a single transformation for the test results, both for healthy and diseased subjects. Although this approach has the advantage of leaving the ROC curve un-

changed (due to its invariance property with respect to increasing monotonic transformations), it can be inappropriate in cases where the two distributions of X and Y are very different from each other. In the following, we show how a more general approach involving two different transformations may be used.

Let

$$X^{(\lambda_1)} = \begin{cases} \frac{X^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(X) & \lambda_1 = 0 \end{cases} \qquad Y^{(\lambda_2)} = \begin{cases} \frac{Y^{\lambda_2}-1}{\lambda_2} & \lambda_2 \neq 0 \\ \log(Y) & \lambda_2 = 0, \end{cases}$$

the transformed test results. The parameters defining the transformations are denoted by λ_1 and λ_2. Our proposal starts from the following observation. Since the optimal threshold τ^+ meets the relation

$$\begin{aligned} \tau^+ &= \arg\max_{\tau} [F_X(\tau) - F_Y(\tau)] = \arg\max_{\tau} [\Pr(X \leq \tau) - \Pr(Y \leq \tau)] \\ &= \arg\max_{\tau} \left[\Pr\left(X^{(\lambda_1)} \leq \tau^{(\lambda_1)}\right) - \Pr\left(Y^{(\lambda_2)} \leq \tau^{(\lambda_2)}\right) \right], \end{aligned}$$

where $\tau^{(\lambda_1)}$ and $\tau^{(\lambda_2)}$ are the transformed values of a same generic threshold τ, the double transformation leaves the optimal values of sensitivity and specificity unchanged. Moreover, $\theta^+ = 1 - \Pr\left(Y^{(\lambda_2)} \leq \tau^{+(\lambda_2)}\right)$ and $\eta^+ = \Pr\left(X^{(\lambda_1)} \leq \tau^{+(\lambda_1)}\right)$. Clearly, in the transformed scales, such optimal values match with two different threshold values.

$$\begin{aligned} \ell(\theta^+, \eta^+, \hat{\tau}^+) = {} & 2m \left\{ \hat{F}_X^{(1)}(\hat{\tau}_1^+) \log \frac{\hat{F}_X^{(1)}(\hat{\tau}_1^+)}{\eta^+} + \left[1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)\right] \log \frac{1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)}{1 - \eta^+} \right\} \\ & + 2n \left\{ \hat{F}_Y^{(2)}(\hat{\tau}_2^+) \log \frac{\hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{1 - \theta^+} + \left[1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)\right] \log \frac{1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{\theta^+} \right\}, \end{aligned} \tag{1}$$

which is infinite when $\hat{\tau}_1^+$ is out of the range of the sample from $X^{(\hat{\lambda}_1)}$, or $\hat{\tau}_2^+$ is out of the range of the sample from $Y^{(\hat{\lambda}_2)}$. Observe that now the pivot in (1) depends on three estimated nuisance parameters: $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\tau}^+$.

In order to obtain confidence regions for the pair (θ^+, η^+), we propose the following algorithm. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we firstly compute the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Then, using transformed data $x'_1, \ldots, x'_m$ and $y'_1, \ldots, y'_n$, say, we obtain sample means and standard deviations $\bar{x}'$, $\bar{y}'$, S'_x, S'_y, and maximize, with respect to τ, $\Phi\left(\frac{\tau^{(\hat{\lambda}_1)} - \bar{x}'}{S'_x}\right) - \Phi\left(\frac{\tau^{(\hat{\lambda}_2)} - \bar{y}'}{S'_y}\right)$ to get $\hat{\tau}^+$. Next, we can get the estimates $\hat{\theta}^+ = 1 - \Phi\left(\frac{\hat{\tau}_2^+ - \bar{y}'}{S'_y}\right)$ and $\hat{\eta}^+ = \Phi\left(\frac{\hat{\tau}_1^+ - \bar{x}'}{S'_x}\right)$. Finally, we use bootstrap calibration:

1. get B parametric bootstrap samples $\{x'\}_b$ and $\{y'\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original (transformed) sample, so as to obtain "enlarged" bootstrap samples;
3. from the b-th pair of (enlarged) bootstrap samples, compute $(\bar{x}'_b, \bar{y}'_b, S'_{xb}, S'_{yb})$, $\hat{\tau}_b^+$, $\hat{\tau}_{1b}^+$, $\hat{\tau}_{2b}^+$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, using (1) where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimate obtained from the original (transformed) data;

4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, $b = 1, \ldots, B$.

The set $\mathscr{R}_\alpha = \{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\}$, is the searched confidence region with nominal coverage $1-\alpha$. In the bootstrap procedure, we process only (enlarged) bootstrap samples whose averages respect the order between the averages of the original (transformed) samples. As an example, Table 1 gives some results of a large simulation study conducted to evaluate the finite sample behaviour of our confidence regions.

Table 1 Simulation results for Log Normal distributions: empirical coverages (over 10000 Monte Carlo replications) for some values of m, n, θ^+ and η^+. $B = 1000$ and $\alpha = 0.1, 0.05, 0.01$.

θ^+	η^+	m	n	0.90	0.95	0.99
0.696	0.861	50	50	0.869	0.929	0.983
		50	75	0.866	0.927	0.984
		75	50	0.862	0.932	0.983
		75	75	0.865	0.932	0.984
		100	100	0.866	0.927	0.983
0.790	0.895	50	50	0.874	0.936	0.986
		50	75	0.877	0.936	0.988
		75	50	0.870	0.935	0.986
		75	75	0.877	0.934	0.986
		100	100	0.867	0.935	0.987
0.888	0.940	50	50	0.903	0.944	0.974
		50	75	0.900	0.946	0.977
		75	50	0.887	0.948	0.991
		75	75	0.891	0.950	0.991
		100	100	0.886	0.947	0.989

References

1. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press. 2003.
2. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3(1): 32–35.
3. Bantis LE, Nakas CT, Reiser B. Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biometrics* 2014; 70(1): 212–223.
4. Yin J, Tian L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput Stat Data An* 2014; 77: 1–13.
5. Adimari G, Chiogna M. Simple nonparametric confidence regions for the evaluation of continuous-scale diagnostic tests. *Int J Biostat* 2010; 6(1).
6. Owen AB. *Empirical likelihood*. Chapman and Hall/CRC . 2001.
7. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biometrical J* 2005; 47(4): 458–472.
8. Box GE, Cox DR. An analysis of transformations. *J R Stat Soc B* 1964; 26(2): 211–243.
9. Hjort NL, McKeague IW, Van Keilegom I. Extending the scope of empirical likelihood. *Ann Stat* 2009; 37(3): 1079–1111.

On the sensitiveness to the memory parameter in the network of tennis

Sensitività del parametro di memoria in una rete dei giocatori di tennis

Alberto Arcagni and Vincenzo Candila and Rosanna Grassi

Abstract In the recent paper of Arcagni et al. (2022), the use of network theory has been introduced in tennis to forecast the winning probabilities of male and female players. The main advantage of the proposed approach is the ability to capture the evolution of each players' ability, even when a set of players is not playing at the time of the matches under consideration. The approach of Arcagni et al. (2022) makes use of the Bonacich centrality computed on the network of players. Such a network is weighted with the number of wins-losses and it has a (decreasing) weight for the matches played far in the past depending on a fixed "memory" parameter. In the present paper, we aim at verifying the sensitiveness of the memory parameter to a grid of values. Surprisingly, the good performances of the approach proposed by Arcagni et al. (2022) are largely preserved, independently of the chosen memory parameter.

Abstract *Nel recente lavoro di Arcagni et al. (2022), l'uso della teoria delle reti è stato introdotto nel tennis per prevedere le probabilità di vittoria di giocatori maschili e femminili. Il maggiore vantaggio dell'approccio proposto è l'abilità di catturare l'evoluzione delle abilità di ciascun giocatore, anche quando un insieme di giocatori non sta giocando al tempo delle partite in esame. L'approccio di Arcagni et al. (2022) utilizza la Bonacich centraility ottenuta sulla rete di giocatori. Tale rete è pesata con il numero di vittorie-scofitte e ha un peso (decrescente) per le partite giocate nel passato, peso che dipende da un parametro di memoria fisso. Nel presente lavoro, si mira a verificare la sensitività del parametro di memoria*

Alberto Arcagni
MEMOTEF Dept., Sapienza University of Rome, Italy, e-mail: alberto.arcagni@uniroma1.it

Vincenzo Candila
MEMOTEF Dept., Sapienza University of Rome, Italy e-mail: vincenzo.candila@uniroma1.it

Rosanna Grassi
Dept. of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy, e-mail: rosanna.grassi@unimib.it

nei confronti di una griglia di valori. Sorprendentemente, le buone performance dell'approccio proposto da Arcagni et al. (2022) sono consistentemente preservate, indipendentemente dalla scelta del parametro di memoria.

Key words: Networks, eigenvector centrality, tennis, sensitiveness, memory parameter, forecasting.

1 Introduction

Recently, Arcagni et al. (2022) introduced the use of the network theory in the sport forecasting literature, with the aim of predicting more accurate and precise rating scores in tennis. In their contribution, Arcagni et al. (2022) take advantage of the network theory to calculate (and daily update) the ratings of the entire set of players, and not limited to the players involved in the matches of the day under investigation. The ratings are obtained through the Bonacich centrality (Bonacich and Lloyd, 2001), which comes out from the eigenvector decomposition of the weighted adjacency matrix, that is the cumulative matrix of wins-losses of the whole network of players. The rates for players at day t depend on the past wins-losses, under a fixed temporal aggregation, which attributes a decreasing weight to the matches played further in the past. In their original contribution, Arcagni et al. (2022) define α as the "memory" parameter, and adopt a fixed configuration for such a parameter. The aim of the present contribution is to verify the sensitiveness of the memory parameter to different values, by means of a calibration procedure across a grid of values.

The rest of paper is structured as follows. In Section 2, the calibration procedure for the memory parameter is described. Section 3 is devoted to the empirical analysis. Conclusions follow.

2 Network approach for players ratings

As in Arcagni et al. (2022), let i and j denote two players among the n different available players. Moreover, let t denote the time at which the matches between the players are disputed, where t_0 is the time of the first match available. Let $\mathbf{W}_t$ be the weighted adjacency matrix, defined as follows:

$$\mathbf{W}_t = \sum_{t^* \in [t_0, t]} f(t^*, t, \alpha) \cdot \mathbf{L}_{t^*}, \tag{1}$$

where $\mathbf{L}_{t^*}$ is the $n \times n$ binary matrix consisting of wins-losses at time $t^* \in [t_0, t]$. The weighting function $f(\cdot)$ in Eq. (1) depends on the memory parameter α and it is expressed as:

$$f(t^*,t,\alpha) = \left(1+\frac{t-t^*}{\alpha}\right)^{-1} \text{ with } t^* \in [t_0,t], \alpha > 0. \quad (2)$$

Note that:

$$\lim_{\alpha\to+\infty} f(t^*,t,\alpha) = 1 \,\forall t \text{ and } t^* \in [t_0,t],$$

meaning that larger values of α assign the same weight to all the matches. Starting from the weighted adjacency matrix in (1), the measure of the strength of all the players at time t is obtained through the Bonacich centrality (Bonacich and Lloyd, 2001). More in detail, it is the principal eigenvector (Horn and Johnson, 2012) of the matrix $\mathbf{W}_t$. The advantage of such approach is that if a player does not play, his/her score depends on the memory of function f, but it is also captured by the network context. In fact, if the player i stops playing, the network approach also captures the variations of the whole set of players.

The main contribution of the present work is to verify the sensitiveness of the scores to different values of α. In the original paper, Arcagni et al. (2022) have set the memory parameter α equal to 365 days, that is, the matches played one year before have a weight of one half. In the cited article, this solution provided excellent results: the scores have been able to forecast consistently better the winning probabilities compared to several popular approaches. However, setting $\alpha = 365$ may be too restrictive. Therefore, we allow the parameter α to vary from 1 to 1460 days. In the case of the limit of α to infinity, all the matches, even those played many years ago, have the same weight.

3 Empirical analysis

The data used in this work were downloaded from the site `www.tennis-data.co.uk`, which provides a large number of statistics concerning the match results of the professional tennis players. This archive has recently been used also by Angelini et al. (2022) and Baker and McHale (2017), among others. The period under investigation spans from July 4, 2005, to November 22, 2020. Only matches involved in the Masters 1000 and Grand Slams are considered. In the first step of the analysis, the *clean* function of the R package 'welo' (Candila, 2021), has been executed. The remaining matches are 14,170, for a total of 470 male players.

The full sample was divided into two non-overlapping periods: a training sample (from 2005 to 2015, consisting of 10,029 matches) and a testing sample (from 2016 to 2020, that is, 4,141 matches). The training sample was used to calculate the centrality-based scores. Once obtained the scores, for all the different memory parameter values adopted, we use the logit regressions as in Arcagni et al. (2022) to compute the winning probabilities. These logit regressions consider as covariates the centrality-based scores of the players. Finally, the Mean-squared-errors (MSE) and Log-loss scores are calculated. The averages of the MSE and Log-loss scores are reported in Figures 1 and 2, respectively. Interestingly, the pattern of the MSE

and Log-loss scores are very similar, reaching a minimum for $\alpha = 91.25$. But, it is worth noting that all the losses here used appear to be quite stable to the grid of α values considered. Only when the the memory parameter is very high, then the losses consistently increase.

4 Conclusions

The choice of the memory parameter in the weighting function (2) does not significantly alter the resulting MSE and Log-loss averages, except when the memory parameter is set to be extremely large. In this case, the weighting function attributes great importance to old matches and, consequently, the network is not flexible enough to catch changes in players abilities.

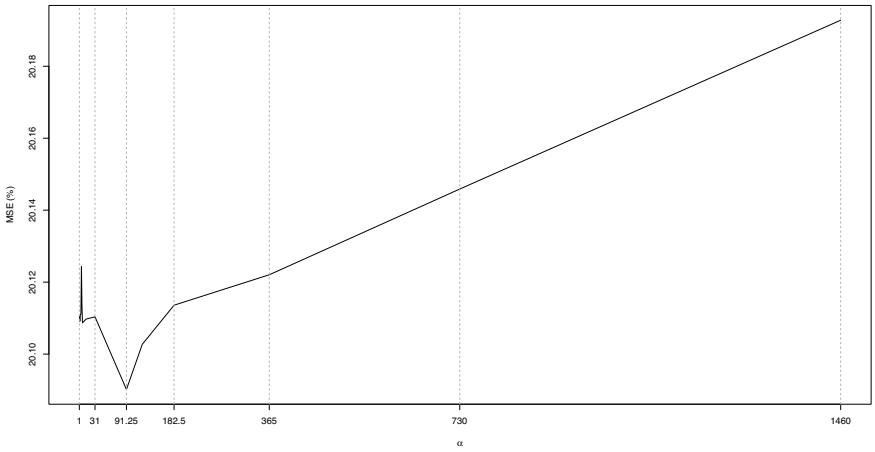


Fig. 1: MSE

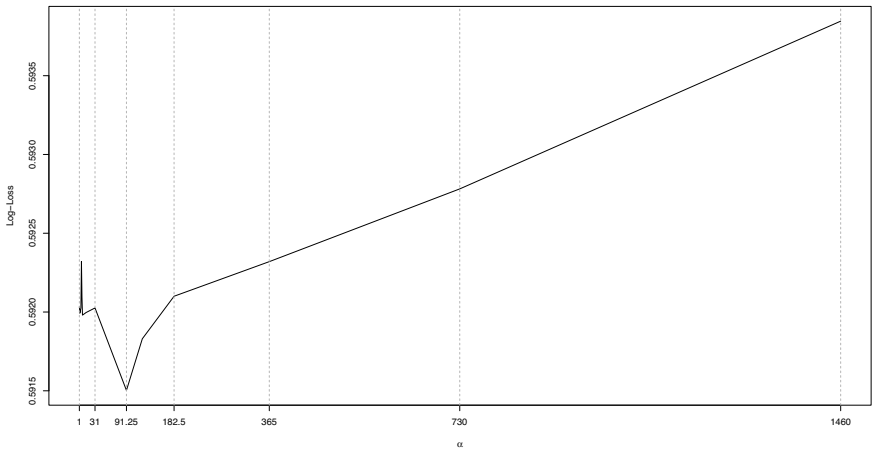


Fig. 2: Log-Loss

References

Angelini G, Candila V, De Angelis L (2022) Weighted Elo rating for tennis match predictions. European Journal of Operational Research 297(1):120–132

Arcagni A, Candila V, Grassi R (2022) A new model for predicting the winner in tennis matches based on the eigenvector centrality. Annals of Operations Research (Forthcoming)

Baker RD, McHale IG (2017) An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player? European Journal of Operational Research 258(1):328–333

Bonacich P, Lloyd P (2001) Eigenvector-like measures of centrality for asymmetric relations. Social networks 23(3):191–201

Candila V (2021) welo: Weighted and Standard Elo Rates. R package version 0.1.1

Horn RA, Johnson CR (2012) Matrix analysis. Cambridge university press

Two-part model with measurement error
Modello a due parti con errore di misura

Maria Felice Arezzo, Serena Arima, and Giuseppina Guagnano

Abstract In many applications, there are positive-valued phenomena which show a very high frequency at zero. One major difficulty with this type of data is that the existence of a point-mass at zero makes common distributions unsuited for modeling the data. To cope with these difficulties, some models have been developed. A popular example is the two-part model in which two stochastic models are assumed: the first governs whether the response variable is zero or positive an the second, conditional on its being positive, models the level. We extend the two-part model to cope with measurement error on the dependent variables of both stochastic parts. This situation is common in many applied works.

Abstract *In molte applicazioni la variabile di interesse assume valori positivi con una frequenza molto alta di valori nulli. Una delle principali difficoltà con questo tipo di dati è che l'esistenza di una massa a zero rende le distribuzioni comuni inadatte per la modellazione dei dati. Per far fronte a queste difficoltà, sono stati sviluppati alcuni modelli. Un esempio è il modello in due parti in cui vengono assunti due modelli stocastici: il primo determina se la variabile di risposta è zero o positiva e il secondo, condizionatamente al primo, ne modella il livello. Estendiamo il modello a due parti tenendo conto dell'errore di misura sulle variabili dipendenti di entrambi i modelli stocastici. Questa situazione è comune in molti lavori applicati.*

Key words: Two-part model, Measurement error in the dependent variable

Maria Felice Arezzo
Department MEMOTEF, Sapienza University of Rome e-mail: mariafelice.arezzo@uniroma1.it

Serena Arima
Department DSSSU University of Salento, Lecce e-mail: name@email.address

Giuseppina Guagnano
Department MEMOTEF, Sapienza University of Rome e-mail: giuseppina.guagnano@uniroma1.it

1 Introduction

In many fields, real phenomena with positive values often show a very high frequency at zero. This kind of data can be represented by semi-continuous variables, which are combination of a point-mass at zero and a positive skewed distribution.

One major difficulty with this type of data is that the existence of a point-mass at zero makes common distributions, such as the gamma, unsuited for modeling the data. To cope with these difficulties, the two-part model has been proposed. In it, two stochastic models are assumed: the first, by means of an additional binary variable, governs whether the response variable is zero or positive and the second, conditional on its being positive, models the level.

Also in many applied works, variables are flawed with measurement error. This could easily happen, for example, during an interview if the respondent misunderstands the question.

We extended the two part model to consider two types of measurement errors: the first affects the binary variable that governs whether the response variable is zero or positive, and the second is on the positive part of the response variable.

In the literature, a mismeasurement on a continuous variable is called measurement error while it is called misclassification when it affects a categorical variable. When the fallible variable is continuous, the two dominant error models are the Berkson's [1] and the classical [3]. In the first one, the error-prone observed value is fixed while the true unobservable variable is random and its random structure is specified conditionally on the former. In the classical approach the error-prone variable is specified as a function of the true one with the error component inserted in a multiplicative or additive form and independent from the true variable.

Let us introduce some notation on the measurement error models used in our work. Let Y^O be the fallible/error-prone binary variable and y^O be the observed value. The misclassification model, which specifies the behaviour of Y^O given the true unobserved value $Y^T = y^T$, is characterized by the misclassification probability:

$$P(Y^O = y^O | Y^T = y^T). \tag{1}$$

Following [4], we set the misclassification probabilities $\alpha_1 = P(Y^O = 0|Y^T = 1)$ (the probability of false negative) and $\alpha_0 = P(Y^O = 1|Y^T = 0)$. (the probability of false positive). Since Y^T is random, if we specify the distribution of $Y^O|Y^T$, it follows that:

$$P(Y_i^O = 1) = (1-\alpha_1)\pi_i + \alpha_0(1-\pi_i) = \pi_i(1-\alpha_0-\alpha_1) + \alpha_0 \tag{2}$$

where $\pi_i = P(Y_i^T = 1)$. Such a probability can be estimated as a function of covariates through a generalized linear model.
When we deal with a continuous variables W, the classical error model in the multiplicative and additive form respectively is:

$$W_i^O = \begin{cases} W_i^T \cdot \xi_i, & \xi_i \sim \log N(\mu, \sigma_\xi^2) \\ W_i^T + \xi_i, & \xi_i \sim N(\mu, \sigma_\xi^2) \end{cases} \qquad \text{with } \mu = 0 \tag{3}$$

where μ is usually null and W^T (or its logarithm) can be specified as a linear function of some predictors. In the proposed model, see section (2), we generalize the distribution of ξ_i, admitting $\mu \neq 0$ and specific for each unit.

2 The proposed model

Let us consider a semi-continuous random variable W, whose observability depends on a binary variable Y: when $Y_i = 1$, we observe a positive value for W_i; otherwise, when $Y_i = 0$, we have $W_i = 0$. Referring to the two-part model: in the first part, we have to specify a binary choice model for the probability of observing a positive-versus-zero outcome and then, in the second part, a regression model is fit for the positive outcome *conditional* on a positive outcome. Under this framework, let us initially assume that there is no measurement error in any of the response variables (the true W and Y coincide with the observable ones). Let us denote themwith W^T and Y^T, the probability of a positive response as $P(W_i^T > 0|\boldsymbol{Z}_i) = P(Y_i^T = 1|\boldsymbol{Z}_i) = \pi_i$, and the conditional distribution of the positive responses as $g(W_i^T|W_i^T > 0, \boldsymbol{X}_i)$, where Z and X are two sets of possibly overlapping explanatory variables.

The two-part model has the following mixture p.d.f. [2] and likelihood:

$$f\left(W_i^T\right) = (1-\pi_i)I(Y_i^T = 0) + \pi_i g(W_i^T|Y_i^T = 1, \boldsymbol{X}_i) \tag{4}$$

$$L(\beta,\theta) = \underbrace{\left[\prod_{Y^T=0}(1-\pi_i)\cdot\prod_{Y^T=1}\pi_i\right]}_{L_1(\boldsymbol{\beta})}\cdot\underbrace{\left[\prod_{Y^T=1}g(W_i^T|Y_i^T=1,\boldsymbol{X}_i)\right]}_{L_2(\boldsymbol{\theta})} \tag{5}$$

where β and θ are vectors of parameters that govern the binary and the continuous part respectively, and $I(Y_i^T = 0)$ is an indicator function such that it equals 1 if $Y_i^T = 0$ and 0 otherwise; it is motivated by the fact that when $Y_i^T = 0$, the density of W^T collapses to a unit probability mass.

More precisely, $L_1(\beta)$ is the likelihood of a standard binary regression model and the corresponding link function is usually specified as probit or logit. $L_2(\theta)$ refers to the regression model for the continuous variable W^T, usually involving gamma or log-Normal distributions. We model L_1 with a probit link and L_2 as a log-Normal regression dealing with the following two-part model:

$$\text{Part one:} \qquad P(W_i^T > 0|\boldsymbol{Z}_i) = P(Y_i^T = 1|\boldsymbol{Z}_i) = \pi_i = \Phi(\boldsymbol{Z}_i\boldsymbol{\beta}) \tag{6}$$

$$\text{Part two:} \qquad \log(W_i^T) = \boldsymbol{X}_i\boldsymbol{\theta} + u_i, \quad u_i \sim N(0;\sigma_u^2) \tag{7}$$

where $\Phi(\cdot)$ is the standard normal c.d.f.

Suppose, now, that both variables W and Y may be affected by measurement errors, so we have the observable Y_i^O and W_i^O and the true Y_i^T and W_i^T. We assume that the measurement error that affects the observable W^O acts in a multiplicative way as in the first line of equation (3), so that $\log(W^O) = \log(W^T) + \varepsilon$, with $\varepsilon = \log(\xi)$. When we admit the possibility of measurement error for W and Y, we can no longer refer only to the p.d.f. of the true W as in (4), but we need to consider the observability of W^O and define its p.d.f.:

$$f\left(\log(W_i^O)\right) = \sum_{j=1}^{4} \psi_{ji} \cdot g_j\left(\log(W_i^O)\right) \tag{8}$$

where the weights are defined as:

$$\begin{aligned}
\psi_{1i} &= P(Y_i^O = 0, Y_i^T = 0) = (1-\alpha_0)\cdot(1-\pi_i) \\
\psi_{2i} &= P(Y_i^O = 1, Y_i^T = 1) = (1-\alpha_1)\cdot\pi_i \\
\psi_{3i} &= P(Y_i^O = 1, Y_i^T = 0) = \alpha_0\cdot(1-\pi_i) \\
\psi_{4i} &= P(Y_i^O = 0, Y_i^T = 1) = \alpha_1\cdot\pi_i
\end{aligned} \tag{9}$$

and the conditional densities in equation (8) are:

$$\begin{aligned}
g_1\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 0, Y_i^T = 0\right) = 1 \\
g_2\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 1, Y_i^T = 1, \boldsymbol{X}_i\right) \\
g_3\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 1, Y_i^T = 0\right) \\
g_4\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 0, Y_i^T = 1\right) = 1
\end{aligned} \tag{10}$$

The density g_2 represents the main contribution in explaining $\log(W^O)$, but its weight ψ_2 tends to zero as α_1 goes to 1.

The density g_3 does not depend on the covariates $\boldsymbol{X}$ because $W_i^T = 0$ when $Y_i^T = 0$; hence it only refers to the erratic component ε. Its weight ψ_3 increases as α_0 gets higher. As a last step, we model Y_i^O and $\log(W_i^O)$ as:

$$P\left(Y_i^O = 1|Z_i\right) = \alpha_0 + (1-\alpha_0-\alpha_1)P\left(Y_i^T|Z_i\right) \tag{11}$$

$$\log(W_i^O) = \boldsymbol{X}_i\boldsymbol{\theta} + (u_i + \varepsilon_i) \tag{12}$$

where $\boldsymbol{X}_i$ is the row vector containing all information for the i-th individual. The first part of the model (i.e. equation 11) is consistent with equation (2). For the second part, coherently with (7), we assume a normal distribution for ε, $\varepsilon_i \sim N(\mu_i, \sigma_\varepsilon^2)$, and consequently for the global error component $u_i + \varepsilon_i = v_i \sim N(\mu_i, \sigma_v^2)$. Furthermore, we assume that u_i and ε_i are uncorrelated. It's important to stress that the above specification extends the classical measurement error model, allowing each unit to

have a different expected value μ_i. In other words, the measurement error may act with a different intensity for each population unit. We model the expected value as a function of individual characteristics: $\mu_i = h(\boldsymbol{X}_i^* \boldsymbol{\gamma})$, with $\boldsymbol{X}_i^*$ row vector. For the sake of simplicity, we just consider a linear function $\mu_i = \boldsymbol{X}_i^* \boldsymbol{\gamma}$. Admitting a varying μ_i implies that the conditional densities g_2 and g_3 must be conditioned to $\boldsymbol{X}_i^*$. Since $\log(W_i^O)$ and μ_i are both specified as linear functions of the predictors, to avoid any problem of identifiability of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we assume that the sets of covariates X and X^* do not overlap.

The contribution of the $i-th$ unit to the likelihood is:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \alpha_0, \alpha_1) = \{(1-\alpha_0) \cdot (1-\pi_i) + \alpha_1 \cdot \pi_i\}^{(1-y_i^O)} \tag{13}$$
$$\cdot \left\{(1-\alpha_1) \cdot \pi_i \cdot N(\boldsymbol{X}_i \boldsymbol{\theta} + \boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_v^2) + \alpha_0 \cdot (1-\pi_i) \cdot N(\boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_\varepsilon^2)\right\}^{y_i^O}$$

where $N(\boldsymbol{X}_i \boldsymbol{\theta} + \boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_v^2)$ and $N(\boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_\varepsilon^2)$ are the densities g_2 and g_3 in equation (10) respectively. In the following, for the sake of brevity and to highlight the dependency on the parameters, we denote them as $g_2(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $g_3(\boldsymbol{\gamma})$.

3 Simulation study

We present the finite sample performances of the proposed model and compare them to the classical probit/ols two part model via Monte Carlo simulations. We assumed the following generating model for the error-free dependent variables:

$$\Pr\left(Y^T | Z_1, Z_2, Z_3\right) = \Phi\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\right) \tag{14}$$
$$\log W^T = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + u \tag{15}$$

where $\Phi(\cdot)$ is the c.d.f of a standard normal. The covariates are generated as follows: Z_1 is log-normal with zero mean and unit variance, X_2 and Z_2 are binomial with $p = 1/3$, X_1 and Z_3 are uniformly distributed over the unit interval. To generate the observed (i.e. error-prone) binary variable, Y^O, we define the misclassification matrix based on equation (1) and we sample accordingly. Finally, the mismeasured continuous part is generated as in equation 12 allowing $\mu_i = \gamma X_4$, with $X_4 \sim Mult_4(p_1 = 0.01; p_2 = 0.06; p_3 = 0.33, p_4 = 0.60)$.

Across simulations we fixed: $\theta^T = (10, 0.8, -0.5)$, $\beta^T = (-1, 0.2, 1.5, -0.6)$, $\sigma_u^2 = 2$ and set the remaining parameters according three scenarios: 1) $\alpha_0 = \alpha_1 = 0.05; \sigma_\varepsilon^2 = 3; \gamma = -0.2$, 2) $\alpha_0 = 0.05; \alpha_1 = 0.20; \sigma_\varepsilon^2 = 3; \gamma = -0.2$ and 3) $\alpha_0 = \alpha_1 = 0.20; \sigma_\varepsilon^2 = 3; \gamma = -0.2$. We repeat each simulation scenario 100 times with samples of size $n = 5,000$. In table 1 we report the results.

For the binary part of the model, even in the case of a small amount of misclassification ($\alpha_0 = \alpha_1 = 0.05$), ordinary probit produces estimates that are biased by 14-22%. As expected, the problem worsens as the amount of misclassification

Table 1 Empirical mean and standard errors, over 100 simulated data sets, of the parameter estimates based on the proposed model and the two-part model.

	Scenario 1			Scenario 2			Scenario 3		
	True value	Proposed model	Two-part model	True value	Proposed model	Two-part model	True value	Proposed model	Two-part model
θ_0	10	9.837	0.881	10	9.854	0.738	10	9.937	0.643
		0.020	*0.004*		*0.021*	*0.004*		*0.014*	*0.004*
θ_1	0.8	0.805	11.138	0.8	0.795	9.394	0.8	0.781	9.521
		0.019	*0.028*		*0.021*	*0.030*		*0.021*	*0.031*
θ_2	-0.5	-0.493	0.155	-0.5	-0.494	0.128	-0.5	-0.486	0.136
		0.003	*0.011*		*0.003*	*0.010*		*0.003*	*0.010*
γ	-0.2	-0.151		-0.2	-0.153		-0.2	-0.172	
		0.005			*0.005*			*0.002*	
σ_v^2	5	4.731	8.371	5	4.708	9.707	5	4.468	10.138
		0.017	*0.032*		*0.019*	*0.030*		*0.018*	*0.030*
β_0	-1	-1.023	-0.843	-1	-1.028	-0.914	-1	-1.003	-0.515
		0.014	*0.005*		*0.018*	*0.005*		*0.035*	*0.004*
β_1	0.2	0.201	0.157	0.2	0.201	0.119	0.2	0.196	0.086
		0.003	*0.002*		*0.004*	*0.001*		*0.006*	*0.001*
β_2	1.5	1.529	1.293	1.5	1.516	1.093	1.5	1.464	0.795
		0.016	*0.004*		*0.023*	*0.004*		*0.035*	*0.004*
β_3	-0.6	-0.626	-0.495	-0.6	-0.619	-0.411	-0.6	-0.578	-0.272
		0.012	*0.007*		*0.015*	*0.007*		*0.019*	*0.006*
α_0	0.05	0.054		0.05	0.050		0.2	0.178	
		0.003			*0.004*			*0.007*	
α_1	0.05	0.046		0.2	0.185		0.2	0.173	
		0.003			*0.006*			*0.006*	

Note: The standard error of the simulation results are reported in italic.

grows. Conversely, the proposed model provides more accurate estimates, in terms of mean squared errors, for all levels of misclassification.

For the continuous part, the results of the proposed model are very encouraging since the estimates of all parameters are trustworthy. These results hold for all simulations scenarios (all tables are available upon request). Although satisfactory, the estimates of γ and σ_v^2 showed some variability in the accuracy when $\sigma_\varepsilon^2 = 3$.

References

1. Berkson, J.:Are there Two Regressions? Journal of the American Statistical Association. **45**,164–180 (1950).
2. Cragg, J. G.: Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica. **39**, 829–844 (1971).
3. Fuller, W.: Measurement Error Model. Wiley (1988).
4. Hausman, J., Abrevaya, J., Scott-Morton, F. M.: Misclassification of the dependent variable in a discrete-response setting. Journal of Econometrics. **87**, 239–269 (1998).

Statistical assessment of practical significance

Valutazione statistica della significatività pratica

Andrea Ongaro, Sonia Migliorati, and Enrico Ripamonti[1]

Abstract Traditionally, common testing problems are formalized as precise null hypotheses representing an idealized situation such as absence (vs. presence) of a treatment effect. However, in most applications the real purpose of the analysis is to assess evidence in favor of a practically relevant effect more than simply determine its presence. We have proposed a sufficiently general approach to the problem of establishing the practical significance of obtained results. This entails the derivation of *ad hoc* procedures for modifying commonly employed tests to make them suitable for appropriate interval null hypotheses. Herein we describe how the problem of the assessment of practical significance could be dealt with in our proposed framework, in case of a sample from a Normal population.

Abstract *Tradizionalmente i problemi di verifica di ipotesi sono formalizzati in termini di ipotesi nulle precise che rappresentano una situazione idealizzata indicante l'assenza (vs. presenza) di un effetto del trattamento. Tuttavia, nella maggior parte delle applicazioni il vero scopo dell'analisi è valutare la significatività pratica di un effetto più che semplicemente determinarne la presenza. Abbiamo proposto un approccio sufficientemente generale al problema di stabilire la significatività pratica. Ciò comporta la derivazione di procedure ad hoc per modificare i test di uso comune al fine di renderli idonei ad appropriate ipotesi di tipo intervallare. In questo contributo descriviamo come affrontare il problema della significatività pratica secondo la nostra prospettiva, considerando il caso studio di un campione da una popolazione Normale.*

Keywords: precise hypotheses, interval hypotheses, *p*-value, practical significance, significance curve

[1] **Andrea Ongaro**, Department of Economics, Management, and Statistics, University of Milan-Bicocca, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: andrea.ongaro@unimib.it. **Sonia Migliorati**, Department of Economics, Management, and Statistics, University of Milan-Bicocca, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: sonia.migliorati@unimib.it **Enrico Ripamonti**, Department of Economics and Management, University of Brescia, Brescia, Italy, and Milan Center for Neuroscience, University of Milan-Bicocca, Milan, Italy; email: enrico.ripamonti@unibs.it

1. Introduction

In applications, most common statistical testing problems are formulated as precise null hypotheses. This is the case, for example, of widespread null hypotheses such as "a certain treatment or explanatory variable is ineffective" or "two or more treatment effects are identical". Very often, however, the real objective of the analysis is to assess whether the effect is large enough to be of practical or substantive significance for the problem at hand rather than simply establishing its presence/absence. For instance, two different drugs can be equivalent from a clinical point of view without having an identical effect.

It follows that the null hypothesis should be more properly formulated as an interval null hypothesis, including all practically irrelevant departures from the precise null. The use of a precise null hypothesis seems justified only if it can be considered as an accurate approximation of the real interval one (see [1] for discussion). However, this is not the case in the commonly encountered context of moderate or large data sets, where even a very small departure from the precise null can be detected (leading to rejection) due to the high power of the test. To appreciate how severely misleading results based on precise null formulation can be, consider the following simple but representative example.

Let X be a vector of i.i.d. observations from a Normal r.v. with unknown mean μ and known variance σ^2. Suppose to test the hypothesis that μ is not substantially different from μ_0. The standard practice is to test the precise null hypothesis $H_0^*: \mu = \mu_0$ vs. $H_0^*: \mu \neq \mu_0$. Let us indicate with $\bar{X}$ the sample mean, t the observed value of the test statistic $T = \sqrt{n}|\bar{X} - \mu_0|/\sigma$, and Φ the standard Normal distribution function. One would then reject H_0^* when the standard p-value $p^* = P_{\mu_0}(T \geq t) = 2(1 - \Phi(t))$ is sufficiently low.

Consider now the more realistic interval hypothesis $H_0: |\mu - \mu_0| \leq \delta$. The corresponding p-value can be shown to be

$$p = p_\delta(t) = sup_{|\mu-\mu_0|\leq\delta} P_\mu(T \geq t) = 2 - [\Phi(t - \delta_{st}) + \Phi(t + \delta_{st})],$$

where $\delta_{st} = \delta\sqrt{n}/\sigma$ expresses the departure in terms of sample mean standard deviation. The validity of the precise hypothesis approximation can be established by choosing t so that the real evidence $p_\delta(t)$ is fixed at a given level (e.g. 0.05) and then comparing this level with p^* as δ_{st} varies (see Figure 1). Remarkably, p^* can be considered approximately correct (less than 10% error) only if δ_{st} is very small (less than 0.1 - 0.2), drastic differences between p and p^* (p^*being about one half p) occurring for δ_{st} ranging from 0.4 to 0.6. Finally, values of δ_{st} in the interval (1, 1.1) lead to a p^* about 1/10 p. In fact, p^* can be proved to decay exponentially fast as δ_{st} diverges, for fixed $p_\delta(t)$.

The implication is that p^* becomes substantially smaller than p, even for moderate n and small δ values. For example, for a departure as small as $\delta/\sigma = 0.05$, the approximation is accurate only for $n \leq 25$; for a less extreme departure of $\delta/\sigma = 0.5$ a 50% error is obtained even when $n = 1$, n =4 already producing a 90% error.

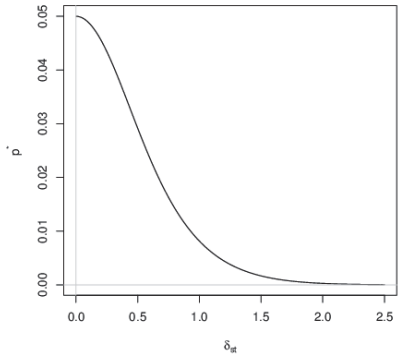


Figure 1. p^* as a function of δ_{st} with $p_\delta(t) = 0.05$.

The distinction between statistical and practical significance is well known in the literature, and it has been, and still is, the subject of animated debates on the usefulness of significance testing in many applied fields [2,3]. More recently an issue of the *American Statistician* devoted to the *p*-value debate has also touched upon this topic (e.g. see [4] and [5]).

Herein we adopt an interval null hypothesis formulation of the problem, which entails two types of inherent difficulties. From a methodological standpoint, the construction of tests for interval null hypotheses in general contexts is not trivial as only tests for precise null hypotheses (in the following standard tests, ST) are implemented and available. Moreover, from a practical point of view, the choice of the threshold δ is often a difficult task in some applied contexts. In [6] a general and unifying method addressing these problems has been put forward. Here we illustrate the method focusing on its application to the simple but very representative Normal case. In the next section we briefly review the general approach.

2. A general framework to test interval null hypotheses

The basic idea of our methodology is to properly modify commonly used precise null hypotheses ST tests. Let $\{T \geq k\}$ be the rejection region of the ST for a given statistic T. In presence of nuisance parameters ψ, its level under the interval null hypothesis can be excessively large (equal to 1) for any fixed k. This is because, typically, the distribution of T does not depend on ψ only under the precise null hypothesis, and T may become very large or even diverge under the interval null hypothesis.

To solve this problem, we let the threshold of the rejection region depend on data, so that it can account for changes in the distribution of T due to changes in ψ. The new threshold is obtained by suitably estimating $k_\alpha(\delta, \psi)$, which represents the α level ST threshold for known ψ. Specifically, we propose two solutions:

1. *Plug-in Test* (PIT): ψ is replaced in $k_\alpha(\delta, \psi)$ with some point estimate $\hat{\psi}$;

2. *Confidence Interval Test* (CIT): $k_\alpha(\delta, \psi)$ is maximized over an arbitrary confidence interval (at level $1-\gamma$) for ψ.

The main advantage of the PIT is its simplicity and good expected behavior for large sample size. However, the attained level is not guaranteed to coincide with the nominal level. The advantage of the CIT is that it can be shown to have a completely general and simple level, namely $\alpha+\gamma$. Consequently, the desired level α is guaranteed in this case by simply choosing an $(\alpha-\gamma)$ level threshold (with $\alpha > \gamma$).

As for the problem of determining the threshold δ, no general automatic rules can be given, since this problem typically relies on a careful context-dependent evaluation. In specific fields reference values may be identified, e.g. see [7]. In other, more difficult cases, we suggest focusing on those departure values which are crucial to assess practical significance. Therefore, we look for the minimum information on δ needed to reach a conclusion with a certain level of evidence. To this end we propose to derive the value δ'_α, defined as *the largest departure δ to be rejected at level α*. In this way, to decide about practical significance (at level α) it is only required to decide whether δ'_α is a relevant or not.

To allow an easy implementation of all proposed procedures, we suggest using a graphical tool, called *significance curve*. This curve is made up of *p*-values for the interval null hypothesis as a function of δ. Thus, it directly provides the degree of evidence against all possible specifications of the interval null hypothesis. Any value δ'_α is readily available by finding the δ value making the curve equal to α.

3. Practical significance in the Normal case

We now illustrate our procedure in the context of a Normal population. When the variance is known, the ST reject region $\{T \geq k\}$ is suitable for the interval null hypothesis. In the notation of Section 1, the test statistic $T = \sqrt{n}|\bar{X} - \mu_0|/\sigma$ is stochastically increasing in $|\mu - \mu_0|$. Therefore a size α threshold $k_\alpha(\delta)$ for the hypothesis $H_0: |\mu - \mu_0| \leq \delta$ can be shown to exist unique and can be obtained by numerically solving the equation

$$p_\delta(k) = sup_{|\mu-\mu_0|\leq\delta} P_\mu(T \geq k) = 2 - [\Phi(k - \delta_{st}) + \Phi(k + \delta_{st})] = \alpha.$$

The value δ'_α can be determined considering that the null hypothesis is rejected iff the *p*-value $p_\delta(t) \leq \alpha$. As $p_\delta(t)$ can be shown to be increasing and continuous in δ, then δ'_α can be computed by solving the equation $p_\delta(t) = \alpha$. Thus δ'_α can be graphically evaluated by plotting the significance curve, i.e. the *p*-value $p_\delta(t)$ as a function of δ or, more conveniently, of δ_{st} (see Figure 2).

This curve promptly provides a clear and complete picture of the information contained in the data: roughly speaking, a strong evidence is present against departures smaller than 0.67 ($p \leq 0.01$), a moderate one against departures between 0.67 and 1.36 ($0.01 \leq p \leq 0.05$), and so on. Notice that, even for moderate *n*, $p^* =$

0.0027 is likely not to be evidence of practical significance: there is evidence at, say, 0.01 level iff a departure $|\mu - \mu_0|$ equal to $0.067\sigma/\sqrt{n}$ is thought to be relevant.

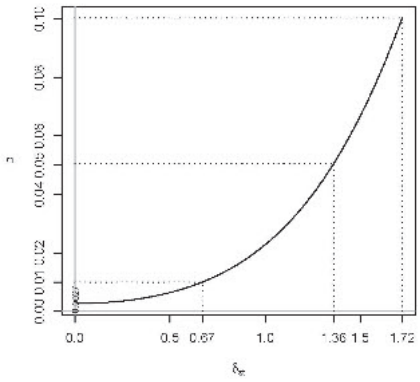


Figure 2. The significance curve with t=3 ($p^* = 0.0027$)

The situation is more involved in case of unknown variance σ^2. For the precise hypothesis $H_0^*: \mu = \mu_0$ the ST is the usual t-test with size α^* rejection region: $R^* = \{|T| \geq t_{n-1;1-\alpha*/2}\}$, where $T = \sqrt{n}(\bar{X} - \mu_0)/S$, S^2 is the sample (unbiased) variance, and $t_{\upsilon,\beta}$ is the β quantile of a Student's t distribution with υ degrees of freedom. In this case the ST $\{|T| \geq k\}$ cannot be used for the interval null hypothesis, as its size can be shown to be equal to 1 for any $k \geq 0$. This is because the distribution of T can be shown to be continuous and stochastically increasing in $\sqrt{n}|\mu - \mu_0|/\sigma$, diverging to infinity when $\sqrt{n}|\mu - \mu_0|/\sigma \to \infty$. Therefore, the probability of $\{|T| \geq k\}$ tends to 1 when $\sigma^2 \to 0$ (if $\mu \neq \mu_0$), for any given k. However, were σ^2 known, a size α threshold $k_\alpha(\delta, \sigma^2)$ can be determined. Specifically, this threshold can be obtained as solution of the equation:

$$p_{\delta,\sigma^2}(k) = P_{\delta,\sigma^2}(|T| \geq k) = 1 - V_{n-1,\frac{\sqrt{n}\delta}{\sigma}}(k) + V_{n-1,\frac{\sqrt{n}\delta}{\sigma}}(-k) = \alpha$$

where $V_{\upsilon,ncp}$ is the non-central Student's t distribution function with υ d.f. and non-centrality parameter *ncp*.

The PIT is obtained by replacing σ^2 with its estimate S^2, which leads to the rejection region $\{|T| \geq k_\alpha(\delta, S^2)\}$. Here the significance curve is given by $p_{\delta,S^2}(t)$ viewed as a function of δ, from which the value δ'_α can be deducted.

Let us now focus on the CIT. The usual two-tailed $(1-\gamma)$ level CI for σ^2 is $[(n-1)S^2/c_{n-1;1-\gamma/2}, (n-1)S^2/c_{n-1;\gamma/2}]$, where $c_{\upsilon;\beta}$ is the β quantile of a Chi-square distribution with υ d.f.. *As* $k_\alpha(\delta, \sigma^2)$ can be shown to be decreasing in σ^2, an α level CIT is given by

$$\{|T| \geq k_{\alpha-\gamma}(\delta, S^2_{inf}\},$$

where $S_{inf}^2 = (n-1)S^2/c_{n-1;1-\frac{\gamma}{2}}$, and the corresponding significance curve takes form $p_{\delta,S_{inf}^2}(t) + \gamma$.

A peculiar feature of the present context is that large values of $|T|$ are associated with large values of $k_\alpha(\delta, S^2)$, being S^2 negatively correlated with $|T|$ and $k_\alpha(\delta, S^2)$ decreasing in S^2. Thus, both tests can be expected to be conservative. Indeed, when $|\mu - \mu_0| = \delta$ the simulated probability of rejection of the PIT is always less than α, being increasing in σ^2. Moreover, the limit of this probability when $\sigma^2 \to \infty$, can be proved to be α, which can thus be considered the exact size of the test. It is shown by simulation that values close to α are already obtained for moderate σ^2 values ($\delta/\sigma \leq 1/4$ for *n*=30) with improving performances as *n* increases.

The CIT exhibits a similar behavior, but with size $\alpha - \gamma$ and a (sometimes relevant) lower first kind error probability.

Interestingly, in this context it is possible to compare the two test performances. Given that in this example both tests have level α, the PIT is preferable as, being by construction more liberal, it can be expected to attain higher power than the CIT. Notice that in general this comparison is not possible, as there is no guarantee that the PIT has level α. Indeed, the PIT was found to be rather liberal in other settings.

4. Conclusion

We have illustrated how our general approach can be rather easily implemented in the Normal case producing sensible results even when nuisance parameters are present. Many other noteworthy cases are covered by our methodology, encompassing the Student's *t*, Chi-square, and Snedecor's *F* type of tests, as well as null hypotheses concerning multidimensional parameters. Our proposal can be applied in many scientific areas involving use of tests and precise hypotheses testing, like medicine, psychology, ecology [8].

References

1. Berger JO, Delampady M. Testing precise hypotheses. Statistical Science. 1987; 317–335.
2. Ziliak S, McCloskey DN. The cult of statistical significance: How the standard error costs us jobs, justice, and lives. Chicago: University of Michigan Press; 2008.
3. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5(2):241–301.
4. Betensky R. The p-value requires context, not a threshold. The American Statistician. 2019; 73(sup1): 115–117.
5. Blume JD, Greevy RA, Welty VF et al. An introduction to second-generation p-values. The American Statistician. 2019; 73(sup1): 157–167.
6. Ongaro A, Migliorati S, Ripamonti E. Testing practical relevance of treatment effects. Milano; 2022. (Mimeo).
7. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Elrbaum Associates, 1988.
8. Fidler F, Geoff C, Mark B, Neil, T. Statistical reform in medicine, psychology and ecology. The Journal of Socio-Economics. 2004;33(5), 615-630.

Autoregressive and mixed effects models

Asymptotic Properties of the Nonlinear Least Squares Estimator in HE-HAR Models

Proprietà asintotiche dello stimatore dei minimi quadrati non lineari in modelli HE-HAR

Emilija Dzuverovic, Edoardo Otranto

Abstract We examine the properties of nonlinear least squares (NLS) estimator for a nonlinear extension of the class of heterogeneous autoregressive (HAR) models for realized covariance (RC) matrices. The Monte Carlo (MC) experiments verify the asymptotic properties of the OLS for multivariate HAR specifications, used as a benchmark. Then we replicate the experiment to verify the same properties for the Hadamard exponential (HE)-based HAR extensions, establishing the convergence of regular HAR coefficients in all cases, while the asymptotic normality of the NLS estimates is uniquely confirmed for the "HE_vech-HAR" specification with log-transformed RC series. The only persistent but relatively narrow asymptotic bias is evident for the "HE" parameter estimate. We submit models in MC exercises to several sensitivity checks and show the robustness of corresponding results.

Abstract *In questo lavoro esaminiamo le proprietà dello stimatore dei minimi quadrati non lineari (NLS) per una estensione della classe di modelli autoregressivi eterogenei (HAR) per matrici di covarianza realizzata (RC). Tramite esperimenti Monte Carlo (MC) verifichiamo le proprietà asintotiche dello stimatore dei minimi quadrati per il modello base HAR multivariato. Ripetiamo poi lo stesso esperimento per verificare le stesse proprietà per il modello Hadamard Exponential (HE)-HAR, stabilendo la convergenza dei coefficienti tipici dell'HAR in tutti i casi, mentre la normalità asintotica è ottenuta in modo univoco solo per la specifica "HE_vech-HAR" con le serie RC trasformate logaritmicamente. E' evidente la presenza di una (leggera) distorsione per lo stimatore del parametro "HE". Vengono realizzati diversi esperimenti MC per verificare la robustezza dei risultati sotto varie ipotesi distribuzionali.*

Keywords: HAR, Hadamard exponential matrix, Monte Carlo, nonlinear least squares, realized covariances.

1 Introduction

The covariance matrix of asset returns is a central input to asset pricing, portfolio allocation and risk management decisions. Over recent decades, considerable research by statisticians and econometricians has been devoted to modelling, estimating and forecasting variance-covariance matrices. The majority of introduced models assume that variances and covariances of asset returns are assessable conditional upon the past information, i.e., multivariate GARCH (MGARCH) models (Bauwens et al. (2006)), or treat them as inherently latent quantities, i.e., multivariate stochastic volatility (MSV) models (Asai et al. (2006)).

Emilija Dzuverovic, University of Messina, Italy; email: edzuverovic@unime.it.

Edoardo Otranto, University of Messina, Italy; email: edoardo.otranto@unime.it.

In contrast, the recent preferable realized covariance method adopts high-frequency data to allow for accurate estimation of the variances and covariances of low-frequency returns (e.g., Andersen et al. (2003), Barndorff-Nielsen and Shephard (2004), Barndorff-Nielsen et al. (2008, 2011)). As volatility becomes essentially "observable", it can be modelled directly, rather than being treated as a latent variable. However, realized covariance (RC) models generally face several challenges, and the most prominent contain ensuring the positive definiteness of the covariance forecasts coupled with the dimensionality issue. The latter implies the rapid increase in the number of parameters with the cross-sectional dimension that has a potential pronounced impact on the statistical efficiency of the estimation and could even render the model estimation unfeasible. Hence, the RC modelling framework calls for parsimonious yet flexible parameterizations, capable of capturing complex serial dependencies observed in realized variances and covariances.

Among the high-frequency data (HFD)-based models, the multivariate extensions of the Heterogeneous Autoregressive (HAR) Model (Corsi, 2009) are widely employed (e.g., vech-HAR Model (Chiriac and Voev (2011), HAR-DRD Model (Oh and Patton (2016)). This class of models originally inspired by the Heterogenous Market Hypothesis of Müller et al. (1993) and the asymmetric propagation of volatility between long and short horizons appears capable of capturing the commonly recognized long memory property of realized covariances via simple linear regression structures.

Nevertheless, in the face of the emphasized curse of dimensionality issue, an easy-to-implement framework contains the adoption of the scalar versions of HAR-based models. This implies that all variances and covariances, or correlations obey the same dynamics, an assumption that is obviously highly restrictive. In this regard, using the element-by-element Hadamard exponential (HE) function of the matrix, the asset pair-specific and time-varying impact coefficients of the lagged realized variances and covariances, or correlations, could be defined, adding great flexibility in the dynamics via a single parameter (Bauwens and Otranto (2020, 2022)).

A practical advantage of the HAR family of models is the straightforward estimation via OLS. By incorporating the so-called "HE" extension (Bauwens and Otranto (2020, 2022)), the nonlinear LS (NLS) could be adopted. Thus, great computational advantages in terms of simplicity, stability and cost potentially arise on behalf of flexible and parsimonious HE_HAR models compared to all other maximum likelihood (ML)-based RC models. It has therefore been natural to ask if the convincing NLS method would generate (at least asymptotically) unbiased, efficient and normally distributed estimators. The approach taken to provide a reliable and empirically valid answer to this question contains the extensive Monte Carlo (MC) simulation experiment.

The rest of the paper is organized as follows. Section 2 introduces the HAR-based framework adopted for MC experiments; Section 3 provides a brief discussion on empirical results and gives some concluding comments.

2 Framework

2.1 vech-HAR Model

The study began with the aim to validate the theoretically established asymptotic properties of the HAR OLS estimates (e.g., McAleer and Medeiros (2008), Hwang and Shin (2014)). We consider the vech-HAR Model (Chiriac and Voev (2011)):

$$\text{vech}(C_t) = (1 - \alpha_D - \alpha_W - \alpha_M)\text{vech}(\overline{C}) + \alpha_D\text{vech}(C_{t-1}) + \alpha_W\text{vech}(\overline{C}_{t-1:t-5}) + \alpha_M\text{vech}(\overline{C}_{t-1:t-22}) + \varepsilon_t,$$

where vech(.) is the operator that stacks the lower triangular portion of its matrix argument as a vector, regressors are the averages of the past RC matrices, $\overline{C}$ is the sample mean of the RC series and α's are scalar parameters. Finally, ε_t is a vector of disturbances.

We consider two alternative specifications for the distribution of ε_t, i.e., Normal, applying the log transformation to realized covariances (Bauer and Vorkink (2011)), and the Wishart distribution (e.g., Golosnoy et al. (2012), Bauwens et al. (2012)). These hypotheses ensure the positive definiteness of RC series. For ease of notation, we indicate the resulting models by "log vech-HAR" and "vech-HAR Wishart".

2.2 HE_vech-HAR Model

However, the adoption of the scalar vech-HAR Model requires imposing a very strong assumption, i.e., all the variances and covariances obey the same dynamics. Bauwens and Otranto (2022) suggest a way of adding great flexibility in the model dynamics while preserving the parsimonious parameterization, i.e., via a single additional parameter, using the Hadamard exponential (HE) function of the matrix.

In practical terms, let us consider again the simple scalar vech-HAR Model (Chiriac and Voev (2011)), where each daily RC matrix is specified as a linear function of lagged daily, weekly and monthly covariances. Bauwens and Otranto (2022) suggest replacing, e.g., the scalar impact coefficient α_D of lagged daily RC on the current RC matrix by $\alpha_D\exp(\varphi\text{vech}(P_{t-1} - J_m)$, where φ is the parameter and P_{t-1} is the lagged realized correlation matrix. Thus, if φ differs from zero, the impact coefficient becomes both asset pair-specific and time-varying.

The adoption of lagged correlations in the impact parameter can be justified by the strong evidence that volatilities and correlations move together (e.g., Andersen et al. (2001)). As such, we expect the transfer of volatility clustering phenomenon to correlations as well. During turbulent periods, the correlations increase, but the level of persistence of such an increase can be distinct across asset pairs. Hence, by adding a dependence of the impact coefficient on lagged correlations, we are able to account for the volatility clustering effect on realized covariances that differs across asset pairs and changes through time.

The vech-HAR Model with the "HE" extension could be estimated by NLS. In this regard, we investigate the asymptotic properties of the HE_vech-HAR NLS estimates generated via the Gauss-Newton algorithm. Again, we consider both the realized covariances in log form (Bauer and Vorkink (2011)) and Wishart generated ones (e.g., Golosnoy et al. (2012), Bauwens et al. (2012)). Moreover, we examine the NLS properties under the recently proposed conditional Matrix F distribution for RC matrices (e.g., Opschoor et al. (2018), Vassallo et al. (2021)). For ease of notation, we indicate resulting models by "log HE_vech-HAR", "HE_vech-HAR Wishart" and "HE_vech-HAR Matrix F".

2.3 HE_HAR-DRD Model

Given that the benchmark scalar HAR-DRD Model (Oh and Patton (2016)) imposes the sharp restriction that all the correlations follow the same dynamics, we introduce the "HE" extension to the specification of the realized correlation series R_t, such that:

$$\text{vech}(R_t) = (1 - \overline{\alpha}_{D,t} - \alpha_W - \alpha_M)\text{vech}(\overline{R}) + \alpha_D \exp(\varphi \text{vech}(R_{t-1} - J_m)) \odot \text{vech}(R_{t-1}) + \alpha_W \text{vech}(\overline{R}_{t-1:t-5}) + \alpha_M \text{vech}(\overline{R}_{t-1:t-22}) + \varepsilon_t,$$

where $\odot$ denotes the Hadamard product, φ is the "HE" parameter, $\overline{R}$ is the sample mean of the realized correlation series, J_m is the m x m matrix of ones with respect to the cross-sectional dimension m and $\overline{\alpha}_{D,t}$ is the average of the entries of $\alpha_D \exp(\varphi \text{vech}(R_{t-1} - J_m))$. Clearly, if φ differs from zero, the impact coefficient becomes asset pair-specific and time-varying.

Within this framework, we keep adopting the benchmark HAR Model (Corsi (2009)) in the first step for each individual realized variance.

Similarly to HE_vech-HAR, HAR-DRD model with the "HE" extension could be estimated by NLS. As follows, we examine the asymptotic properties of HE_HAR-DRD NLS estimates generated via the Gauss-Newton algorithm. Based on the outcomes of initial asymptotic estimations, we simulate only "HE_HAR-DRD Wishart" Model.

3 Discussion

Both the linear and nonlinear HAR processes discussed above are simulated for the cross-sectional dimension m = 3 with sample sizes T = 1000, 3000, 5000, 10000, 50000. Then, the OLS/NLS estimates are computed for the parameters with the replication number N = 1000, coupled with the sample means of 1000 estimates and respective standard deviations. Finally, we check the asymptotic normality via the Jarque-Bera test (Jarque and Bera (1981)).

Based on the MC replication outcomes, we verify the consistency and asymptotic normality of the OLS estimates for the benchmark multivariate HAR specifications, i.e., "log vech-HAR" with(out) covariance targeting and "vech-HAR Wishart", with the full convergence of parameters attained at the sample size T = 10000 and steadily reducing standard deviations as the sample size grows. Furthermore, the asymptotic normal distribution of all estimated coefficients is confirmed simultaneously with the consistency evidence. E.g., Table 1 reports the results for the "log vech-HAR" Model.

Considering the Gauss-Newton NLS estimates, we establish the convergence of the regular HAR parameters in each HE_vech-HAR specification, i.e., "log HE_vech-HAR", "HE_vech-HAR Wishart" and "HE_vech-HAR Matrix F", with the consistency of the coefficients on log-transformed lagged RC validated even at the sample size T = 5000 coupled with the confirmation of the "normality" hypothesis on all parameters. Conversely, in each model, the remaining "HE" parameter estimate is featured by the relatively narrow asymptotic bias around the absolute 2%. E.g., Table 2 reports the results for the "log HE_vech-HAR" Model.

On the other hand, the NLS for the nonlinear HAR-DRD specification, i.e., "HE_HAR-DRD Wishart", generates inconsistent and non-Gaussian estimates, with the average absolute 1.5% asymptotic bias on conventional HAR parameters and rather pronounced "HE" term bias up to -10%.

Finally, we submit the models in the MC exercises to several sensitivity checks and show the robustness of corresponding results.

Table 1: $N = 1000$ "log vech-HAR" with the covariance targeting estimates.

Parameter/Sample Size	T = 1000	T = 3000	T = 5000	T = 10000	T = 50000
$\alpha_D = 0.45$					
Sample Mean	**0.45**	**0.45**	**0.45**	**0.45**	**0.45**
SD	(0.00114)	(0.00065)	(0.0005)	(0.00035)	(0.00016)
Jarque-Bera p-value	**0.8037**	**0.5434**	**0.6151**	**0.3466**	**0.3127**
$\alpha_W = 0.25$					
Sample Mean	0.24	**0.25**	**0.25**	**0.25**	**0.25**
SD	(0.00173)	(0.00101)	(0.00079)	(0.00055)	(0.00025)
Jarque-Bera p-value	**0.1051**	**0.9129**	**0.2670**	**0.9632**	**0.1026**
$\alpha_M = 0.15$					
Sample Mean	0.14	0.14	**0.15**	**0.15**	**0.15**
SD	(0.00174)	(0.00103)	(0.0008)	(0.00055)	(0.00024)
Jarque-Bera p-value	**0.9688**	**0.8918**	**0.8063**	**0.4520**	**0.8113**

Note: The table reports the sample mean of the 1000 OLS estimates of each "log vech-HAR" parameter (with covariance targeting) with the corresponding standard deviation in parenthesis (at sample sizes $T = 1000, 3000, 5000, 10000, 50000$). It also shows the p-value of the Jarque-Bera test for each coefficient and respective sample size, where the entries in boldface indicate the failure to reject the null hypothesis of "normality" at the 5% significance level.

Table 2: $N = 1000$ "log HE_vech-HAR" estimates.

Parameter/Sample Size	T = 1000	T = 3000	T = 5000	T = 10000	T = 50000
$\alpha_D = 0.21$					
Sample Mean	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**
SD	(0.00060)	(0.00034)	(0.00026)	(0.00019)	(0.00009)
Percentage AB	-0.274%	-0.335%	-0.301%	-0.275%	-0.266%
Jarque-Bera p-value	**0.0579**	**0.4740**	**0.1157**	0.0050	**0.1787**
$\alpha_W = 0.33$					
Sample Mean	**0.33**	**0.33**	**0.33**	**0.33**	**0.33**
SD	(0.00099)	(0.00057)	(0.00043)	(0.00031)	(0.00014)
Percentage AB	-0.041%	0.047%	0.015%	-0.005%	-0.004%
Jarque-Bera p-value	**0.8834**	**0.9225**	**0.7298**	**0.3233**	**0.1471**
$\alpha_M = 0.22$					
Sample Mean	0.21	**0.22**	**0.22**	**0.22**	**0.22**
SD	(0.00115)	(0.00062)	(0.00049)	(0.00035)	(0.00016)
Percentage AB	-1.327%	-0.431%	-0.249%	-0.089%	-0.007%
Jarque-Bera p-value	**0.8069**	**0.1961**	**0.3642**	**0.8022**	**0.2799**
$\varphi = 0.15$					
Sample Mean	0.14	0.13	0.13	0.13	0.13
SD	(0.00327)	(0.00185)	(0.00143)	(0.00100)	(0.00047)
Percentage AB	-1.005%	-1.621%	-1.757%	-1.836%	-2.082%
Jarque-Bera p-value	1.96E-10	0.0379	**0.5345**	**0.7134**	**0.7335**

Note: The table reports the sample mean of the 1000 NLS estimates of each "log HE_vech-HAR" parameter with the corresponding standard deviation in parenthesis, coupled with the percentage absolute bias (at sample sizes $T = 1000, 3000, 5000, 10000, 50000$). It also shows the p-value of the Jarque-Bera test for each coefficient and respective sample size, where the entries in boldface indicate the failure to reject the null hypothesis of "normality" at the 5% significance level.

Given the breadth of the HAR modelling framework and their widespread adoption in the context of accurate RC matrix forecasting, these results provide a valuable extension to existing ones in the literature and immediate practical implications. However, an avenue we have not explored in this study is building up the alternative flexible but still parsimonious multivariate HAR extensions subject to the NLS estimation. E.g., incorporating the smooth transition structure within the MHAR framework to explicitly account for the asymmetric behaviour of realized variances and covariances (see e.g., McAleer and Medeiros (2008), Qu et al. (2016)). Hence, it remains completely an open empirical question whether the consistency and asymptotic normality of the NLS could be confirmed for the upcoming nonlinear HAR models for RC matrix, which arguably reflect the interesting topics for further research.

The full set of MC experiments with the corresponding results is available upon request.

References

1. Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H.: The distribution of realized stock return volatility. J. Financ. Econom. (2001) doi: 10.1016/S0304-405X(01)00055-1
2. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: Modeling and forecasting realised volatility. Econometrica (2003) doi: 10.1111/j.1468-0262.2004.00515.x
3. Asai, M., McAleer, M., Yu, J.: Multivariate stochastic volatility: a review. Econometric Rev. (2006) doi: 10.1080/07474930600713564
4. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica (2008) doi: 10.3982/ECTA6495
5. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and nonsynchronous trading. J. Econometrics (2011) doi: 10.1016/j.jeconom.2010.07.009
6. Barndorff-Nielsen, O.E., Shephard, N.: Econometric analysis of realized covariation: high frequency based covariance, regression, and correlation in financial economics. Econometrica (2004) doi: 10.1111/j.1468-0262.2004.00515.x
7. Bauer, G.H., Vorkink, K.: Forecasting multivariate realized stock market volatility. J. Econometrics (2011) doi: 10.1016/j.jeconom.2010.03.021
8. Bauwens, L., Laurent, S., Rombouts, J.V.K.: Multivariate GARCH models: a survey. J. Appl. Econometrics (2006) doi: 10.1002/jae.842
9. Bauwens, L., Otranto, E.: Nonlinearities and regimes in conditional correlations with different dynamics. J. Econometrics (2020) doi: 10.1016/j.jeconom.2019.12.014
10. Bauwens, L., Otranto, E.: Modelling realized covariance matrices: a class of Hadamard exponential models. J. Financ. Econom. (2022) doi: 10.1093/jjfinec/nbac007
11. Bauwens, L., Storti, G., Violante, F.: Dynamic conditional correlation models for realized covariance matrices. CORE DP 60, 104-108 (2012)
12. Chiriac, R., Voev, V.: Modelling and forecasting multivariate realized volatility. J. Appl. Econometrics (2011) doi: 10.1002/jae.1152
13. Corsi, F.: A simple long memory model of realized volatility. J. Financ. Econom. (2009) doi: 10.1093/jjfinec/nbp001
14. Golosnoy, V., Gribisch, B., Liesenfeld, R.: The conditional autoregressive Wishart model for multivariate stock market volatility. J. Econometrics (2012) doi: 10.1016/j.jeconom.2011.11.004
15. Hwang, E., Shin, D.W.: Infinite-order, long-memory heterogeneous autoregressive models. Comput. Stat. Data Anal. (2014) doi: 10.1016/j.csda.2013.08.009
16. Jarque, C.M., Bera, A.K.: Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. Econ. Lett. (1981) doi: 10.1016/0165-1765(81)90035-5
17. McAleer, M., Medeiros, M.C.: A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. J. Econometrics (2008) doi: 10.1016/j.jeconom.2008.09.032
18. Müller, U., Dacorogna, M., Dav, R., Olsen, R., Pictet, O., Ward, J.: Fractals and intrinsic time - a challenge to econometricians. In: Proceedings of the XXXIXth International AEA Conference on Real Time Econometrics, Luxembourg (1993)
19. Oh, D.H., Patton, A.J.: High-dimensional copula-based distributions with mixed frequency data. J. Econometrics (2016) doi: 10.1016/j.jeconom.2016.04.011
20. Opschoor, A., Janus, P., Lucas, A., Dijk, D.V.: New heavy models for fat-tailed realized covariances and returns. J. Bus. Econom. Statist. (2018) doi: 10.1080/07350015.2016.1245622
21. Qu, H., Chen, W., Niu, M.Y., Li, X.D.: Forecasting realized volatility in electricity markets using logistic smooth transition heterogeneous autoregressive models. Energy Econ. (2016) doi: 10.1016%2Fj.eneco.2015.12.001
22. Vassallo, D., Buccheri, G., Corsi, F.: A DCC-type approach for realized covariance modeling with score-driven dynamics. Int. J. Forecast. (2021) doi: 10.1016/j.ijforecast.2020.07.006

A note on testing for threshold non-linearity in presence of heteroskedasticity in time series

Una nota sui test per non-linearità a soglia in presenza di eteroschedasticità

Simone Giannerini and Greta Goracci

Abstract In this note we study the behaviour of an asymptotic test for linearity against the threshold autoregressive moving average model in presence of conditional heteroskedasticity. We recall the relevant asymptotic theory both under the null and the alternative hypotheses for a supremum Lagrange Multiplier test statistic in presence of *i.i.d.* errors. By means of a small simulation study we show that when the innovations follow a GARCH process the size of the test can be severely biased and we discuss possible solutions to the problem.

In questa nota studiamo il comportamento di un test per la non linearità di tipo autoregressivo-media mobile con soglia, in presenza di eteroschedasticità. Introduciamo la teoria asintotica di riferimento, sia sotto l'ipotesi nulla sia sotto l'ipotesi alternativa, per la statistica test basata sui moltiplicatori di Lagrange quando gli errori sono i.i.d.. Per mezzo di uno studio di simulazione mostriamo poi che, quando gli errori seguono un processo GARCH, allora l'ampiezza del test può risultare fortemente distorta e discutiamo possibili soluzioni al problema.

Key words: Threshold autoregressive moving-average models, Lagrange multiplier test, Linearity testing, heteroskedasticity, GARCH models

Acknowledgements Greta Goracci acknowledges the support of Libera Università di Bolzano, Grant WW202G (TARMAECON).

Simone Giannerini
Dipartimento di Scienze Statistiche, Università di Bologna, via delle belle arti 41, Bologna, e-mail: simone.giannerini@unibo.it

Greta Goracci
Facoltà di Economia e Management, Libera Università di Bolzano, piazza Università 1, 39100, Bozen-Bolzano, e-mail: greta.goracci@unibz.it

1 Introduction

The problem of testing for threshold non-linearity in time series has received much attention since the seminal works of Chan [5] [6]. Among the contributions, in [10] a stochastic permutation scheme is used to derive the null distribution of the test statistics, whereas in [8] the validity of the bootstrap has been proved. Also, in [13] and [9] an extension of the test to threshold autoregressive moving average models (TARMA) is presented. Tests for unit-root against the alternative hypothesis of a stationary threshold process have been discussed in [3],[11], [4], [7].

In all the abovementioned works, the innovation process is assumed to be *i.i.d.*. To some extent, the behaviour of such tests in presence of conditional heteroskedasticity has been studied in [14], that discuss testing for a TAR process with ARCH-type innovations, whereas [12] introduce tests for TMA with GARCH-type innovations. To the best of our knowledge, there are no studies on the asymptotic behaviour of the general ARMA versus TARMA test when the innovation process presents conditional heteroskedasticity.

In this paper we revisit the general framework of asymptotic testing for threshold ARMA effects introduced in [9] and present a Monte Carlo study where we show that such test can be severely biased in presence of heteroskedastic innovations. The paper is structured as follows: in Section 2, we describe the sup Lagrange Multiplier test statistic (sLM*), recall its asymptotic distribution and the consistency of the associated test. In Section 3, we present a simulation study on the size of the sLM* test in presence of conditional heteroskedasticity, whereas in Section 4 we offer some discussion and possible solutions to the problem.

2 The test and its asymptotics

Let the time series $\{X_t : t = 1,\dots,n\}$ follow the threshold autoregressive moving-average model (TARMA(p,q)) defined by the difference equation:

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t + \left(\varphi_0 + \sum_{i=1}^{p} \varphi_i X_{t-i} - \sum_{j=1}^{q} \vartheta_j \varepsilon_{t-j} \right) I(X_{t-d} \leq r).$$

The innovations $\{\varepsilon_t\}$ follow a *i.i.d.* process with zero mean and finite variance σ^2; p, q and d are known positive integers and indicate the autoregressive order, the moving-average order and the delay parameter, respectively. Moreover, $I(\cdot)$ is the indicator function and $r \in \mathbb{R}$ is the threshold parameter. For notational convenience, we abbreviate $I(X_t \leq r)$ by $I_r(X_t)$. Let $\phi = (\phi_0, \phi_1, \dots, \phi_p)^\intercal$, $\theta = (\theta_1, \dots, \theta_q)^\intercal$, $\varphi = (\varphi_0, \varphi_1, \dots, \varphi_p)^\intercal$ and $\vartheta = (\vartheta_1, \dots, \vartheta_q)^\intercal$. Moreover we define:

$$\eta = (\eta_1^\intercal, \eta_2^\intercal, \sigma^2)^\intercal \quad \text{with} \quad \eta_1 = (\phi^\intercal, \theta^\intercal)^\intercal \quad \text{and} \quad \eta_2 = (\varphi^\intercal, \vartheta^\intercal)^\intercal.$$

The true parameters have a subscript zero, e.g. η_0, $\eta_{0,1}$, θ_0, ... Also, we assume η_0 to be an interior point of the parameter space.

We test whether a TARMA(p,q) model provides a significantly better fit than the linear ARMA(p,q) model by developing a Lagrange multiplier test statistic. Let $\mathbf{0}$ be the vector with all zeroes, hence the system of hypotheses becomes:

$$\begin{cases} H_0 & : \eta_{0,2} = \mathbf{0}, \\ H_1 & : \eta_{0,2} \neq \mathbf{0}, \end{cases}$$

Under H_0, the time series follows a linear ARMA(p,q) model:

$$X_t = \phi_{0,0} + \sum_{i=1}^{p} \phi_{0,i} X_{t-i} - \sum_{j=1}^{q} \theta_{0,j} \varepsilon_{t-j} + \varepsilon_t. \tag{1}$$

We rely upon the following assumptions:

A1. Let $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ and $\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q$, then $\phi(z) \neq 0$ and $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$ and they do not share common roots. Moreover, $\{X_t\}$ is ergodic and invertible both under H_0 and under H_1.

A2. ε_t has a continuous and strictly positive density f on the real line and $E\left[\varepsilon_t^4\right]$ is finite.

A3. The density f of ε_t is absolutely continuous with derivative f' and $\int \left(f'(x)/f(x)\right)^2 f(x)\,dx < \infty$.

The Gaussian log-likelihood, conditional on the initial values $X_0, X_{-1}, \ldots$, is given by

$$\ell_n(\eta, r) = -\frac{n}{2}\ln(\sigma^2 2\pi) - \frac{1}{2\sigma^2}\sum_{t=1}^{n} \varepsilon_t^2(\eta, r)$$

$$\varepsilon_t(\eta, r) = X_t - \left\{ \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}(\eta, r) \right\} - \left\{ \varphi_0 + \sum_{i=1}^{p} \varphi_i X_{t-i} - \sum_{j=1}^{q} \vartheta_j \varepsilon_{t-j}(\eta, r) \right\} I_r\left(X_{t-d}\right)$$

and the q initial values are set to zero. Consider the following first and second partial derivatives of $\ell_n(\eta, r)$:

$$\frac{\partial \ell_n(\eta, r)}{\partial \eta_2} = -\frac{1}{\sigma^2}\sum_{t=1}^{n} \varepsilon_t(\eta, r)\frac{\partial \varepsilon_t(\eta, r)}{\partial \eta_2};$$

$$\mathscr{I}_n(\eta, r) = \begin{pmatrix} \mathscr{I}_{n,11}(\eta) & \mathscr{I}_{n,12}(\eta, r) \\ \mathscr{I}_{n,21}(\eta, r) & \mathscr{I}_{n,22}(\eta, r) \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 \ell_n(\eta, r)}{\partial \eta_1 \partial \eta_1^\intercal} & -\frac{\partial^2 \ell_n(\eta, r)}{\partial \eta_1 \partial \eta_2^\intercal} \\ -\frac{\partial^2 \ell_n(\eta, r)}{\partial \eta_2 \partial \eta_1^\intercal} & -\frac{\partial^2 \ell_n(\eta, r)}{\partial \eta_2 \partial \eta_2^\intercal} \end{pmatrix}.$$

We write $\partial\hat{\ell}_n(r)/\partial\eta_2$ and $\hat{\mathscr{I}}_n(r)$ to refer to $\partial\ell_n(\eta,r)/\partial\eta_2$ and $\mathscr{I}_n(\eta,r)$ evaluated at the restricted maximum likelihood estimator of η. The test statistic is:

$$T_n = \sup_{r\in[r_L,r_U]} T_n(r), \tag{2}$$

where

$$T_n(r) = \left(\frac{\partial\hat{\ell}_n(r)}{\partial\eta_2}\right)^{\intercal}\left(\hat{\mathscr{I}}_{n,22}(r) - \hat{\mathscr{I}}_{n,21}(r)\hat{\mathscr{I}}_{n,11}^{-1}\hat{\mathscr{I}}_{n,12}(r)\right)^{-1}\frac{\partial\hat{\ell}_n(r)}{\partial\eta_2}.$$

Both under H_0 and H_1, the asymptotic distribution of T_n is a functional of a Gaussian process. Hence, define $\{\xi(r),\ r\in\mathbb{R}\}$ to be a centered Gaussian process with covariance kernel

$$\Sigma(r_1,r_2) = \Lambda_{22}(r_1\wedge r_2) - \Lambda_{21}(r_1)\Lambda_{11}^{-1}\Lambda_{12}(r_2),$$
$$\text{where } \Lambda_{ij}(r) = E\left[\frac{1}{\sigma_0^2}\left(\frac{\partial\varepsilon_t(\eta_0,r)}{\partial\eta_i}\right)\left(\frac{\partial\varepsilon_t(\eta_0,r)}{\partial\eta_j}\right)^{\intercal}\right].$$

In order to assess the behaviour of the test under the alternative hypothesis, let $\mathbf{h} = \left(h_{10},h_{11},\cdots,h_{1p},h_{21},\cdots,h_{2q}\right)^{\intercal}\in\mathbb{R}^{p+q+1}$ be a vector of constants and r_0 a fixed scalar; we consider the sequence of alternatives $H_{1,n}$ specifying that X_t satisfies the following difference equation:

$$X_t = \phi_{0,0} + \sum_{i=1}^{p}\phi_{0,i}X_{t-i} - \sum_{j=1}^{q}\theta_{0,j}\varepsilon_{t-j} + \varepsilon_t + \left[\frac{h_{10}}{\sqrt{n}} + \sum_{i=1}^{p}\frac{h_{1i}}{\sqrt{n}}X_{t-i} - \sum_{j=1}^{q}\frac{h_{2j}}{\sqrt{n}}\varepsilon_{t-j}\right]I_{r_0}(X_{t-d}).$$

The next theorem contains the main results.

Theorem 1. *Let $\{\xi(r),\ r\in\mathbb{R}\}$ be defined as above. Then it holds that:*

(i) under H_0 and Assumptions A1 and A2, T_n converges weakly to

$$\sup_{r\in[r_L,r_U]}\xi(r)^{\intercal}\left(\Lambda_{22}(r) - \Lambda_{21}(r)\Lambda_{11}^{-1}\Lambda_{12}(r)\right)^{-1}\xi(r).$$

(ii) under $H_{1,n}$ and Assumptions A1 – A3, T_n converges weakly to

$$\sup_{r\in[r_L,r_U]}(\xi_r+\gamma_r)^{\intercal}\left(\Lambda_{22}(r) - \Lambda_{21}(r)\Lambda_{11}^{-1}\Lambda_{12}(r)\right)^{-1}(\xi_r+\gamma_r),$$

where $\gamma_r = \left\{\Lambda_{22}(\min\{r,r_0\}) - \Lambda_{21}(r)\Lambda_{11}^{-1}\Lambda_{12}(r_0)\right\}\mathbf{h}$.

(iii) under $H_{1,n}$ and Assumptions A1 – A3, as $\|\mathbf{h}\|\to\infty$, the test statistic T_n has power approaching 100%.

3 Finite sample behaviour in presence of heteroskedasticity

In this section, we study the finite sample performance of the sLM* test in presence of GARCH innovations. The length of the series is $n = 100, 200, 500$ and z_t, $t = 1, \ldots, n$ is generated from a standard Gaussian white noise. The nominal size of the test is $\alpha = 5\%$ and the number of Monte Carlo replications is 1000. The critical values of the test are taken from [1] and the threshold is searched from percentile 25th to 75th of the sample distribution. We simulate from the following ARMA$(1,1)$-GARCH$(1,1)$ model:

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t. \\ \varepsilon_t &= (a_0 + a_1 \varepsilon_{t-1}^2 + b_1 h_{t-1})^{1/2} z_t. \end{aligned} \tag{3}$$

where $\phi_1 = (-0.8, -0.4, 0, 0.4, 0.8)$ and $\theta_1 = (-0.5, 0.0, 0.5)$. We combine these with the following parameters for the GARCH specification: $(1,0,0)$ (case A), $(a_0, a_1, b_1) = (1, 0.1, 0.8)$ (case B), $(1, 0.8, 0.1)$ (case C), so as to obtain 15 different parameterizations. Note that Case A corresponds to the case of *i.i.d.* errors (no GARCH effect) and is taken as a reference. The results are presented in Table 1. Clearly, when the errors are *i.i.d.* (Case A), the size of the test converges to the correct nominal 5% size as the sample size increases. However, in presence of GARCH-type errors this is not the case. In particular, for Case C, the size bias gets worse with increasing sample sizes.

Table 1 Empirical size at nominal level $\alpha = 5\%$ for the ARMA(1,1)-GARCH(1,1) process of Eq. (3) for sample sizes 100, 200, 500. The Case A corresponds to a ARMA(1,1) with *i.i.d.* errors.

ϕ_1	θ_1	Case A			Case B			Case C		
		100	200	500	100	200	500	100	200	500
-0.8	-0.5	8.0	7.5	6.6	8.0	10.2	10.3	29.3	41.5	54.3
-0.4	-0.5	6.2	5.3	5.1	8.6	9.1	11.4	31.5	39.3	58.6
0.0	-0.5	6.1	6.5	5.9	9.4	10.8	11.7	31.8	39.5	59.8
0.4	-0.5	12.1	7.9	5.4	12.7	9.7	11.7	29.2	38.9	53.7
0.8	-0.5	27.6	17.3	8.4	28.5	19.7	12.6	36.4	38.3	46.1
-0.8	0.0	6.7	4.6	4.8	9.1	7.8	9.2	28.5	40.4	55.2
-0.4	0.0	4.8	4.0	5.3	7.9	7.9	10.0	26.9	40.0	55.5
0.0	0.0	10.9	11.4	8.8	11.9	12.3	11.5	26.9	38.5	52.1
0.4	0.0	7.2	6.6	5.6	8.3	8.2	10.3	24.6	33.3	51.0
0.8	0.0	16.5	10.3	8.0	17.8	11.1	9.4	33.7	39.2	52.9
-0.8	0.5	8.8	7.1	4.9	9.8	9.4	9.8	26.2	35.0	48.9
-0.4	0.5	11.3	7.2	5.8	12.5	9.5	8.8	26.9	37.5	55.5
0.0	0.5	4.5	5.6	5.3	7.0	7.9	9.1	21.3	30.9	47.8
0.4	0.5	6.9	5.4	6.6	8.0	8.2	10.0	23.6	37.0	53.3
0.8	0.5	13.4	9.0	8.4	14.9	13.0	10.0	31.5	38.9	50.8

4 Conclusions

The results of the simulation study clearly show that disregarding the heteroskedasticity can produce a substantial bias in the size of tests for threshold effects in time series. Different remedies to this problem can be adopted. A first, naive, option would be to perform a preliminary test for the presence of (conditional) heteroskedasticity prior to testing for threshold effects. Obviously, this is a suboptimal solution since the overall significance level of the tests should be adjusted, which is most likely unfeasible since there are no available results on the dependence between the relevant statistics. A second approach would be to adopt a bootstrap scheme that is robust against heteroskedasticity, such as the wild bootstrap. Unfortunately, as also discussed in [8], there are major technical difficulties with proving the validity of the wild bootstrap in this framework. A third, more natural, approach would be to include the GARCH terms directly in the specification of the auxiliary model associated to the supLM statistic of Eq. (2). This latter approach is the subject of ongoing investigations [2].

References

1. D.W.K. Andrews. Tests for parameter instability and structural change with unknown change point: A corrigendum. *Econometrica*, 71(1):395–397, 2003.
2. F. Angelini, M. Castellani, S. Giannerini, and G. Goracci. Detecting and modelling threshold effects in presence of volatility and measurement error: The case of italian strikes. Technical report, 2022.
3. F. Bec, M. Ben Salem, and M. Carrasco. Tests for unit-root versus threshold specification with an application to the purchasing power parity relationship. *J. Bus. Econom. Statist.*, 22(4):382–395, 2004.
4. F. Bec, A. Guay, and E. Guerre. Adaptive consistent unit-root tests based on autoregressive threshold model. *J. Econometrics*, 142(1):94–133, 2008.
5. K.-S. Chan. Testing for threshold autoregression. *Ann. Statist.*, 18(4):1886–1894, 12 1990.
6. K. S. Chan. Percentage points of likelihood ratio tests for threshold autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 53(3):691–696, 1991.
7. K.-S. Chan, S. Giannerini, G. Goracci, and H. Tong. Testing for threshold regulation in presence of measurement error with an application to the PPP hypothesis. Technical report, 2021.
8. S. Giannerini, G. Goracci, and A. Rahbek. The validity of bootstrap testing in the threshold framework. Technical report, 2021.
9. G. Goracci, S. Giannerini, K.-S. Chan, and H. Tong. Testing for threshold effects in the TARMA framework. *Statistica Sinica*, 33(3), 2023.
10. B. E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430, 1996.
11. G. Kapetanios and Y. Shin. Unit root tests in three-regime SETAR models. *The Econometrics Journal*, 9(2):252–278, 2006.
12. G. Li and W.K. Li. Testing for threshold moving average with conditional heteroscedasticity. *Statist. Sinica*, 18:647–665, 04 2008.
13. G. Li and W.K. Li. Testing a linear time series model against its threshold extension. *Biometrika*, 98(1):243–250, 02 2011.
14. C.S. Wong and W.K. Li. Testing for threshold autoregression with conditional heteroscedasticity. *Biometrika*, 84(2):407–418, 1997.

The Conditional Autoregressive Wishart-$\mathscr{G}$ Model

Il Modello Wishart-$\mathscr{G}$ Autoregressivo Condizionale

Massimiliano Caporin and Marco Girardi

Abstract We propose a new distribution for modelling realised covariances of financial assets, obtained as the product of a scalar component distributed as a unit-mean inverse gamma and a matrix component following a Wishart distribution. We adopt an autoregressive moving average structure for the expected value of the resulting distribution. The aim of the model is to capture a common factor within the assets and to separate it from the *idiosyncratic* conditional covariance dynamic.

Abstract *Proponiamo una distribuzione innovativa per modellare la covarianza realizzata di strumenti finanziari, ottenuta come prodotto di una componente scalare con distribuzione gamma inversa a media unitaria e una componente matriciale che segue una distribuzione Wishart. Utilizziamo una struttuta autoregressiva a media mobile per il valore atteso della distribuzione risultante. Il fine del modello è di identificare un fattore comune tra gli asset e di separarlo dalla dinamica della covarianza condizionata idiosincratica.*

Key words: conditional autoregressive model, realised covariance, common factor, portfolio allocation, index tracking

1 Introduction

Conventional approaches for multivariate modelling and forecasting of covariances of asset returns are based on low-frequency data and model matrices either given past observations, such as multivariate GARCH models [1], or treat them as an un-

Massimiliano Caporin
University of Padova - Department of Statistical Sciences, Via Cesare Battisti, 241 - 35121 Padova - Italy e-mail: massimiliano.caporin@unipd.it

Marco Girardi
University of Padova - Department of Statistical Sciences, Via Cesare Battisti, 241 - 35121 Padova - Italy e-mail: marco.girardi.6@phd.unipd.it

observed stochastic process, such as multivariate stochastic volatility models [6]. More recently, realised measures from high-frequency data have proven to be a valuable tool in this field. A renowned approach is the Conditional Autoregressive Wishart (CAW) model by [3], generalising the Wishart Autoregressive (WAR) model by [4] with an autoregressive moving average structure for the mean of the Wishart distribution, allowing both lagged covariance matrices and lagged predictions of the covariance to contribute to the prediction.

In this work we propose a novel multivariate distribution for realised covariance modelling and forecasting, obtained as the product of a Wishart distribution and a scalar component following a unit-mean inverse gamma. The resulting distribution allows to identify a common factor of risk amongst the assets that can be removed to focus the attention on the residuals, which we refer to as the *idiosyncratic covariance dynamic*. The common factor is initially modelled as the conditional mean of a Multiplicative Error Model (MEM), and then realised covariances are filtered from this mutual element and endowed with a CAW-like structure for prediction. The model accounts for symmetry and positive definiteness of the covariance matrices and is estimated by maximum likelihood.

2 The $\mathscr{G}$ density

Let X be a unitary-mean inverse gamma random variable with shape parameter $-\alpha$ and $\mathbf{Y}$ be a Wishart random variable with n degrees of freedom and scale matrix C/n, that is

$$f_X(x) = \frac{x^{\alpha-1}}{(-\alpha-1)^{\alpha}\Gamma(-\alpha)} \exp\left(\frac{\alpha+1}{x}\right), -\alpha, x > 0, \tag{1}$$

and

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{(n/2)^{nm/2} \mid \mathbf{y} \mid^{(n-m-1)/2} \exp(-\frac{n}{2} Tr(C^{-1}\mathbf{y}))}{h(n,m) \mid C \mid^{n/2}}, \tag{2}$$

where m is the dimension of $\mathbf{Y}$ and $h(n,m) = \pi^{m(m-1)/4} \prod_{j=1}^{m} \Gamma(n/2 + (1-j)/2)$. The densities of $\mathbf{\Omega} = X\mathbf{Y}$ and $\mathbf{Z} = \delta\mathbf{\Omega}$, where δ is a scalar, are provided by Theorem 1 and Corollary 1, respectively.

Theorem 1. *Let X be a unitary-mean inverse gamma random variable and $\mathbf{Y}$ be a Wishart random variable with n degrees of freedom and scale matrix C/n, then $\mathbf{\Omega} = X\mathbf{Y}$ follows a $\mathscr{G}(\alpha, C, n)$ density, that is,*

$$f_{\mathbf{\Omega}}(\boldsymbol{\omega}) = \frac{(n/2)^{nm/2} \mid \boldsymbol{\omega} \mid^{(n-m-1)/2} \Gamma(nm/2 - \alpha)}{h(n,m) \mid C \mid^{n/2} (-\alpha-1)^{\alpha}\Gamma(-\alpha)} \times \\ \times \left(\frac{n}{2} Tr(C^{-1}\boldsymbol{\omega}) - (\alpha+1)\right)^{\alpha - nm/2}. \tag{3}$$

Proof. The density of $\mathbf{\Omega}$ is given by

$$f_{\mathbf{\Omega}}(\boldsymbol{\omega}) = \int_{R_+} f_{x\mathbf{Y}}(\boldsymbol{\omega}) f_X(x)\,\mathrm{d}x, \tag{4}$$

where the scale transformation $x\mathbf{Y}$ follows a Wishart with expected value xC. Thus, Eq. 4 becomes

$$\begin{aligned} f_{\mathbf{\Omega}}(\boldsymbol{\omega}) = & \frac{(n/2)^{nm/2} \mid \boldsymbol{\omega} \mid^{(n-m-1)/2}}{h(n,m) \mid C \mid^{n/2} (-\alpha-1)^{\alpha}\Gamma(-\alpha)} \times \\ & \times \int_{R_+} x^{\alpha-1-nm/2} \exp\Big(x^{-1}\Big(-\frac{n}{2}Tr(C^{-1}\boldsymbol{\omega}) + (\alpha+1)\Big)\Big)\,\mathrm{d}x. \end{aligned} \tag{5}$$

The solution of this kind of integral is provided by [5] in section 3.326 at page 337. Integrating Eq. 5 by substitution using $x = y^{-1}$, we get

$$\begin{aligned} f_{\mathbf{\Omega}}(\boldsymbol{\omega}) = & \frac{(n/2)^{nm/2} \mid \boldsymbol{\omega} \mid^{(n-m-1)/2} \Gamma(nm/2-\alpha)}{h(n,m) \mid C \mid^{n/2} (-\alpha-1)^{\alpha}\Gamma(-\alpha)} \times \\ & \times \Big(\frac{n}{2}Tr(C^{-1}\boldsymbol{\omega}) - (\alpha+1)\Big)^{\alpha-nm/2}, \end{aligned} \tag{6}$$

which is a $\mathscr{G}(\alpha, C, n)$ density. □

Corollary 1. *Let $\mathbf{\Omega} \sim \mathscr{G}(\alpha, C, n)$ and δ be a scaling factor, then $\delta\mathbf{\Omega} \sim \mathscr{G}(\alpha, \delta C, n)$.*

Proof. Let $\mathbf{Y} \sim W(C,n)$, where C is the expected value of $\mathbf{Y}$, and A is a square matrix of dimension m, then $A\mathbf{Y}A' \sim W(ACA', n)$. If we apply the scale transformation δ to the Wishart density $\mathbf{Y}$, and we define $A = diag(a_1, \ldots, a_m)$ with $a_i = \sqrt{\delta}$, $i = 1, \ldots, m$, then $\delta\mathbf{Y} = A\mathbf{Y}A' \sim W(\delta C, n)$, and, by Theorem 1, the density of $\mathbf{Z} = \delta\mathbf{\Omega}$ is

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) = & \frac{(n/2)^{nm/2} \mid \mathbf{z} \mid^{(n-m-1)/2} \Gamma(nm/2-\alpha)}{h(n,m) \mid \delta C \mid^{n/2} (-\alpha-1)^{\alpha}\Gamma(-\alpha)} \times \\ & \times \Big(\frac{n}{2}Tr\big((\delta C)^{-1}\mathbf{z}\big) - (\alpha+1)\Big)^{\alpha-nm/2}, \end{aligned} \tag{7}$$

which is a $\mathscr{G}(\alpha, \delta C, n)$ density. □

2.1 Dynamics of δ_t

The aim of this item is to extract a common factor amongst the securities, so the first step is to define how to model the elements of the covariance matrix in a way that allows to sort mutual and specific components. If we denote with $v_{ij,t}$ the generic ij element of the realised covariance matrix, we can express it with a multiplicative

model $v_{ij,t} = \lambda_t \sigma_{ij,t}$, where λ_t is a common factor for the elements in the matrix while $\sigma_{ij,t}$ is an element-specific factor assuming either positive or negative values. Then, the estimate of λ_t is

$$\tilde{\lambda}_t = \left(\prod_{ij \in \mathbf{S}} | v_{ij,t} | \right)^{\frac{2}{m(m+1)}}, \tag{8}$$

where $\mathbf{S}$ is the set of ij combinations of the elements in the lower triangular part of the realised covariance matrix, including the elements on the main diagonal. The scalar δ_t is modelled as the conditional mean of the Multiplicative Error Model (MEM) by [2], using $\zeta_t = T\tilde{\lambda}_t / \sum_{i=1}^{T} \tilde{\lambda}_i$ as source of information to ensure $\mathbb{E}[\delta_t] = 1$, that is

$$\delta_t = (1 - \theta_1 - \theta_2) + \theta_1 \delta_{t-1} + \theta_2 \zeta_{t-1}. \tag{9}$$

According to [7], the parameters of Eq. 9 can be estimated with a GARCH software by taking $\sqrt{\zeta}$ as dependent variable and setting the mean to zero.

2.2 Dynamic of C_t

The conditional mean of $\mathbf{Y}$ is endowed with an autoregressive moving average process resembling the CAW model by [3]. This model allows the predicted covariance matrix to depend on lagged covariance matrices as well as on their lagged predictions. A CAW of order (p,q) expresses C_t as

$$C_t = \mathbf{C}\mathbf{C}' + \sum_{i=1}^{p} \mathbf{A}_i C_{t-i} \mathbf{A}_i' + \sum_{j=1}^{q} \mathbf{B}_j \mathbf{Z}_{t-j} \mathbf{B}_j', \tag{10}$$

where $\mathbf{C}$ is a lower-triangular matrix, and $\mathbf{A}_i, \mathbf{B}_i$ are parameter matrices. The structure guarantees the symmetry and positive definiteness of the conditional mean C_t without imposing parametric restrictions on $(\mathbf{C}, \mathbf{A}_i, \mathbf{B}_i)$. In order to limit the number of parameters being estimated, $\mathbf{A}_i$ and $\mathbf{B}_i$ can be defined as diagonal matrices, and the intercept can be expressed as a function of the model long-run covariance obtained as follows. Let vec$(\cdot)$ denote the operator that stacks all columns of a matrix into a vector, and let vech$(\cdot)$ and ivech$(\cdot)$ denote respectively the operator that stacks the lower triangular portion (including the diagonal) of a matrix into a vector and the one reversing this operation. Then, the vector representation of the CAW (p,q) is

$$c_t = c + \sum_{i=1}^{p} \mathscr{A}_i c_{t-i} + \sum_{j=1}^{q} \mathscr{B}_j z_{t-j}, \tag{11}$$

where $c_t = vech(C_t)$, $z_t = vech(\mathbf{Z}_t)$, $c = vech(\mathbf{C}\mathbf{C}')$, and $(\mathscr{A}_i, \mathscr{B}_j)$ are $k \times k$ matrices, with $k = m(m+1)/2$, obtained as

$$\mathscr{A}_i = L_m (\mathbf{A}_i \otimes \mathbf{A}_i) D_m, \quad \mathscr{B}_j = L_m (\mathbf{B}_j \otimes \mathbf{B}_j) D_m, \tag{12}$$

where L_m and D_m denote the elimination and duplication matrices such that $vec(X) = D_m vech(X)$ and $vech(X) = L_m vec(X)$. The unconditional mean is given by

$$\mathbb{E}[z_t] = \left(I_k - \sum_{i=1}^{max(p,q)} (\mathscr{A}_i + \mathscr{B}_i) \right)^{-1} c, \tag{13}$$

and the covariance targeting variant is therefore

$$C_t = \tilde{c} + \sum_{i=1}^{p} \mathbf{A}_i C_{t-i} \mathbf{A}_i' + \sum_{j=1}^{q} \mathbf{B}_j \mathbf{Z}_{t-j} \mathbf{B}_j', \tag{14}$$

where $\tilde{c} = ivech(c)$. We adopt an analogous approach for modelling the conditional mean of $\mathbf{Y}$, using diagonal parameter matrices with $p = 1$, $q = 1$ and covariance targeting. Additionally, $\mathbf{Z}_{t-1}$ is replaced by $\mathbf{Z}_{t-1}/\zeta_{t-1}$, since realised covariances are filtered by the terms used for the common factor estimation. The parameters in Eq. 14 are estimated by maximum likelihood.

3 Empirical Analysis

We use high-frequency data at 1-minute level (390 daily observations) of the 26 assets of the Dow Jones Industrial Average (DJIA) which have always been included in the index in the period of analysis. The realised covariance matrices are computed as $RCOV_t = \sum_{i=1}^{I} r_{t,i} r_{t,i}'$, where $r_{t,i}$ is the vector of percentage log-returns for the i-th 1-minute interval of the t-th trading day. The sample period is 2 January 2003 to 24 June 2020 with a total of 4400 trading days.

We want to show that δ_t can be seen as a proxy for market variance, so we fit the model to the full sample of daily realised covariances and compare the results with the DJIA variance, computed with an EGARCH(1,1), and realised variance. The correlation of the common factor with the DJIA variance is above 0.9, meaning that δ_t well represents market movements. This close relation is also shown in Fig. 1 where δ_t is plotted with the DJIA realised variance (rescaled for comparison, since δ_t is constructed to have unit-mean).

Another way of displaying the crucial role of δ_t is to focus on its relevance within the conditional mean. Fluctuations in the covariance matrix are mostly captured by the common factor, which promptly adapts to sudden market changes, while residual movements are explained by the elements in C_t. This is clearly observable in periods marked by extreme instability such as the 2008 financial crises and the recent pandemic. When considering the median value of the off-diagonal elements of $RCOV_t$ and C_t, it can be noticed that C_t has a marginal effect on the overall conditional mean $\delta_t C_t$, having a correlation of just 0.02 with the median of the off-diagonal elements of realised covariances as well as of conditional means.

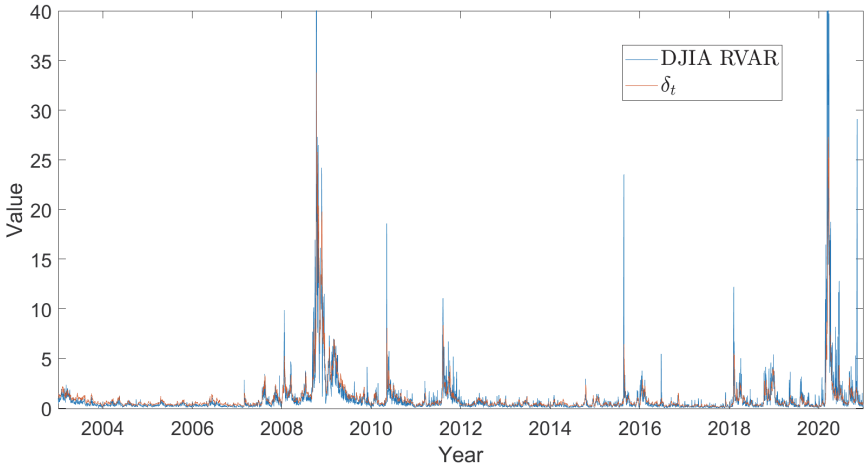


Fig. 1 Common factor and DJIA realised variance

4 Conclusions

In this work we propose a conditional autoregressive model for realised covariances using a novel distribution that identifies a common factor among the assets with the remarkable property of accurately tracking market variability. The model can be exploited to isolate the effect of the common factor and operate on the remaining component only for asset allocation applications, for instance in market risk tracking and hedging.

References

1. Bollerslev, T., Engle, R.F., Wooldridge, J.M.: A capital asset pricing model with time-varying covariances. J. Polit. Econ. **96**, 116–131 (1988)
2. Engle, R.: New frontiers for ARCH models. J. Appl. Econ. **17**, 425–446 (2002)
3. Golosnoy, V., Gribisch, B., Liesenfeld, R.: The conditional autoregressive Wishart model for multivariate stock market volatility. J. Econ. **167**, 211–223 (2012)
4. Gouriéroux, C., Jasiak, J., Sufana, R.: The Wishart autoregressive process of multivariate stochastic volatility. J. Econ. **150**, 167–181 (2009)
5. Gradshteyn, I.S., Ryzhik, I.M.: Table of integrals, series, and products. Elsevier (2007)
6. Harvey, A., Ruiz, E., Shephard, N.: Multivariate stochastic variance models. Rev. Econ. Stud. **61**, 247–264 (1994)
7. Lee, S.-W., Hansen, B.E.: Asymptotic theory for the GARCH (1, 1) quasi-maximum likelihood estimator. Econ. Theory **10**, 29–52 (1994)

Semi-parametric generalized linear mixed effects models for binary response for the analysis of heart failure hospitalizations

Modelli lineari generalizzati ad effetti misti semi-parametrici per risposta binaria per l'analisi di ospedalizzazioni per scompenso cardiaco

Alessandra Ragni, Chiara Masci, Francesca Ieva, Anna Maria Paganoni

Abstract In this work, we propose a semi-parametric mixed effects model for binary response with an Expectation Maximization (EM) algorithm for the parameters estimation and we apply it to an administrative database of Lombardy Region (Italy) related to patients hospitalised for hearth failure between 2000 and 2012 within different hospitals. The semi-parametric assumption provides the random effects of the model to be distributed according to a discrete distribution with an (a priori) unknown number of support points, inducing an automatic clustering of the higher level units (in our application, hospitals). This modelling induces hospitals within the same cluster to share the same random effects.

Abstract *In questo lavoro, proponiamo un modello ad effetti misti semi-parametrico con risposta binaria, con un algorithmo con tecnica EM per la stima dei parametri e lo applichiamo ad un dataset amministrativo della Regione Lombardia (Italia), relativo ad ospedalizzati per scompenso cardiaco tra il 2000 e il 2012 in diversi ospedali. La semi-parametricità risiede nell'assunzione di una distribuzione discreta degli effetti random del modello, con un numero (a priori) sconosciuto di punti di supporto, inducendo così un clustering delle osservazioni relative al livello più alto della gerarchia (nella nostra applicazione, gli ospedali). Con tale modello, gli ospedali nello stesso cluster condividono lo stesso effetto random.*

Alessandra Ragni
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, Milan, Italy - e-mail: alessandra.ragni@polimi.it

Chiara Masci
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, Milan, Italy - e-mail: chiara.masci@polimi.it

Francesca Ieva
CHDS – Center for Health Data Science, Human Technopole, Milan, Italy
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, Milan, Italy - e-mail: francesca.ieva@polimi.it

Anna Maria Paganoni
MOX - Modelling and Scientific Computing lab, Dipartimento di matematica, Politecnico di Milano, Milan, Italy - e-mail: anna.paganoni@polimi.it

Key words: Semi-parametric generalized linear mixed effects model, EM algorithm, Hearth Failure, administrative databases, hospitalization rates

1 Introduction

In this work, we intend to propose an innovative model called *Semi-Parametric Generalized Linear Mixed-effect Model* (SPGLMM), aiming at identifying sub-populations inside a hierarchical data structure, where groups within the same sub-population share the same random effect.

More in details, a clustering of the groups at the highest level of hierarchy is automatically performed by the algorithm by assuming the random effects of the Generalized Linear Mixed effects Model (GLMM) to be distributed as a discrete distribution with an a-priori unknown number of support points, creating an association between the group and the sub-population to which it belongs to; this is the novelty of the SPGLMM and in such assumption lies the non-parametricity of the algorithm. Semi-parametric mixed-effects models have been already proposed in the literature, but only for continuous and multinomial responses [1, 2]. SPGLMM generalizes this approach for binary and other responses in the exponential family (such as Poisson,...). A suitable Expectation Maximization (EM) algorithm for the parameters estimation is provided and, in order to test the proposed method, different simulations are performed.

In addition, an application to a real data case is presented. We apply SPGLMM to data retrieved from the administrative databases of Lombardy Region (Italy), related to patients hospitalized for Heart Failure (HF) between 2000 and 2012 [3, 4]. HF is a major and growing public health issue, characterized by high costs, steep morbidity and mortality rates [5]. Despite the advances in the understanding the pathophysiology of chronic HF and the improvement of therapy, HF mortality and morbidity rates remain high [6]. The analysis focuses on discovering sub-populations of hospitals which differently affect patients hospitalizations, with the aim of getting insights on the effects the different hospitals treatments have on the HF patients.

2 Semi-parametric generalized linear mixed effect model

Given $i = 1,...,N$ groups, GLMMs are generally defined such that, conditioned on the random effects $\mathbf{b_i}$ in the i^{th} group, the dependent variable $\mathbf{y_i}$ (n_i-dimensional vector) in the i^{th} group is distributed according to the exponential family with its expectation, being related to the linear predictor $\mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i}$ via a link function $g(\cdot)$:

$$g(E[\mathbf{y_i}|\mathbf{b_i}]) = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} \quad \text{for} \quad i = 1,...,N$$

where $\mathbf{X_i}$ and $\mathbf{Z_i}$ are, respectively, the $[n_i \times (P+1)]$ and $[n_i \times (Q+1)]$ matrices of fixed and random covariates in the i^{th} group; β is the $(P+1)$-dimensional vector of fixed coefficients and $\mathbf{b_i}$ the $(Q+1)$-dimensional vector of random coefficients in the i^{th} group.

In the parametric framework, the random coefficients $\mathbf{b}$ are assumed to be normally distributed: $\mathbf{b} \sim N(\mathbf{0}, \Sigma_{\mathbf{b}})$.

In our semi-parametric approach, we assume $\mathbf{b}$ to follow a discrete distribution, called P^*, with an a-priori unknown number of support points. More in details, P^* is a discrete distribution with M support points $(\mathbf{c_1}, ..., \mathbf{c_M})$ for $M \leq N$, where each $\mathbf{c_l} \in \mathbb{R}^{Q+1}$, $l = 1, ..., M$, corresponds to the l^{th} subpopulation. The number of discrete masses M is not chosen a-priori, but it is estimated together with the other parameters of the model. The interpretation is that each group i ($i = 1, ..., N$) is assigned with a certain probability w_l to a sub-population l with parameter values $\mathbf{c_l}$ allowing to identify a latent structure among the groups. The distribution of the random effect coefficient $\mathbf{b_i}$ is the following:

$$\mathbf{b_i} = \begin{cases} \mathbf{c_1} & p(\mathbf{b_i} = \mathbf{c_1}) = w_1 \\ ... & \\ \mathbf{c_M} & p(\mathbf{b_i} = \mathbf{c_M}) = w_M \end{cases} \qquad \sum_{l=1}^{M} w_l = 1, \qquad w_l \geq 0$$

3 The EM algorithm for SPGLMM

The algorithm is inspired by the one proposed in *Azzimonti et al. (2013)*[7] and consists in the computation of the *expected log likelihood* and its maximization with respect first to random effects and then to fixed effects, in an iterative framework. More in details, the maximization of the following quantity is performed:

$$(\beta^{up}, \mathbf{c_1}^{up}, ..., \mathbf{c_M}^{up}) = \underset{\beta, \mathbf{c_l}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{l=1}^{M} \mathbf{W_{il}} logL(\beta, \mathbf{c_l} | \mathbf{y_i})$$

where $L(\beta, \mathbf{c_l} | \mathbf{y_i}) = p(\mathbf{y_i} | \beta, \mathbf{c_l}) = \prod_{j=1}^{n_i} f(\mathbf{y_{ij}}, g(E[\mathbf{y_{ij}} | \mathbf{c_l}])$, being $f(\cdot)$ the density function of the response, and $\mathbf{W_{il}} = \frac{w_l p(\mathbf{y_i} | \beta, \mathbf{c_l})}{\sum_{k=1}^{M} w_k p(\mathbf{y} | \beta, \mathbf{c_k})} = \frac{p(\mathbf{b_i} = \mathbf{c_l}) p(\mathbf{y_i} | \beta, \mathbf{c_l})}{p(\mathbf{y_i} | \beta)} = p(\mathbf{b_i} = \mathbf{c_l} | \mathbf{y_i}, \beta)$.

It is possible to divide the maximization in two substeps:

1. Computation of $\mathbf{c_l}^{up}$, given β computed at previous iteration, as

$$\mathbf{c_l}^{up} = \underset{\mathbf{c_l}}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbf{W_{il}} logL(\beta, \mathbf{c_l} | \mathbf{y_i}) \quad \text{for} \quad l = 1, ..., M$$

2. Computation of β^{up}, given $\mathbf{c_l}$ computed at the previous step, as

$$\beta^{up} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{l=1}^{M} \mathbf{W_{il}} logL(\beta, \mathbf{c_l} | \mathbf{y_i})$$

The weights can be updated as $w_l^{up} = \frac{\sum_{i=1}^{N} \mathbf{W_{il}}}{N}$, which corresponds to the sample mean over the N groups of the l^{th} population; therefore, the estimation of the coefficients $\mathbf{b_i}$ of the random effects for each group is obtained as follows:

$$\tilde{\mathbf{b}}_\mathbf{i} = \mathbf{c}_{\tilde{\mathbf{l}}} \quad \text{where} \quad \tilde{l} = \underset{l}{\text{argmax}}\, \mathbf{W}_{\mathbf{il}} \text{ for } i = 1, ..., N$$

Finally, the support reduction of the discrete distribution stands on two criteria:

- two points $\mathbf{c_l}$ and $\mathbf{c_k}$ closer than a fixed threshold D collapse to a unique point $\mathbf{c}_{\mathbf{l,k}} = \frac{\mathbf{c_l}+\mathbf{c_k}}{2}$ with weight $w_{l,k} = w_l + w_k$;
- points with a weight w lower than a fixed threshold $\tilde{w}$ or not associated to any subpopulation are removed; when one or more mass points are deleted, the remaining weights are reparametrized in such a way that they sum up to 1.

4 SPGLMM for binary response and simulation study

In the simulation and in the case study, we focus on the case of a binary response variable. In such family of models, the link function $g(\cdot)$ is assumed to be the *logit* link. If we assume the random coefficients $\mathbf{b}$ to follow the discrete distribution P^* described in Section 2, the SPGLMM formulation for observation j in group i is the following:

$y_{ij}|\mathbf{c_l} \sim Ber(\pi_{ijl})$ for $i = 1, ..., N$, $j = 1, ..., n_i$ and $\pi_{ijl} = \mathbf{E}[y_{ij}|\mathbf{c_l}] = P(y_{ij} = 1|\mathbf{c_l})$

$$\eta_{ijl} = logit(\pi_{ijl}) = log\Big(\frac{\pi_{ijl}}{1-\pi_{ijl}}\Big) = \sum_{h=1}^{P} x_{ijh}\beta_h + \sum_{k=0}^{Q} z_{ijk}c_{lk}$$

$$p(\mathbf{y_i}|\beta, \mathbf{c_l}) = \prod_{j=1}^{n_i} \frac{exp\{y_{ij}\eta_{ijl}\}}{1+exp\{\eta_{ijl}\}} \quad \text{and} \quad L(\beta, \mathbf{c_1}, ..., \mathbf{c_M}|\mathbf{y}) = \sum_{l=1}^{M} w_l \prod_{i=1}^{N} p(\mathbf{y_i}|\beta, \mathbf{c_l}).$$

To validate the estimation algorithm proposed, we perform a simulation study: we consider $I = 10$ groups of data, where each group contains $n_{i=1,...,I} \sim U(70, 100)$ and we induce the presence of three subpopulations. We implement three different Data Generating Processes (DGP) each of them based on models including one fixed covariate $\mathbf{x_i} \sim N(\mathbf{0}, \mathbb{1}_{n_i})$, with the following differences: DGP A based on a **model A** having both random intercept and random covariate, DGP B based on a **model B** having only a random covariate, DGP C based on a **model C** having only a random intercept. The random covariate, when present, is distributed as $\mathbf{z_i} \sim N(\mathbf{0}, \mathbb{1}_{n_i})$. Table 1 reports the mathematical formulation of the three models. The thresholds D and $\tilde{w}$, described at the end of Section 3, are set to 1 and 0.05, respectively. Table 2 shows, for each of the three models, the coefficients (fixed and random) simulated and estimated by the algorithm, respectively in the upper and lower part of the table, in 100 runs of the algorithm.

model A	**model B**	**model C**
$\begin{cases} \eta_i = c_{01} + c_{11}\mathbf{z_i} + \beta_0\mathbf{x_i} & i=1,2 \\ \eta_i = c_{02} + c_{12}\mathbf{z_i} + \beta_0\mathbf{x_i} & i=3,...,7 \\ \eta_i = c_{03} + c_{13}\mathbf{z_i} + \beta_0\mathbf{x_i} & i=8,9,10 \end{cases}$	$\begin{cases} \eta_i = \beta_0 + c_{01}\mathbf{z_i} + \beta_1\mathbf{x_i} & i=1,2 \\ \eta_i = \beta_0 + c_{02}\mathbf{z_i} + \beta_1\mathbf{x_i} & i=3,...,7 \\ \eta_i = \beta_0 + c_{03}\mathbf{z_i} + \beta_1\mathbf{x_i} & i=8,9,10 \end{cases}$	$\begin{cases} \eta_i = c_{01} + \beta_0\mathbf{x_i} & i=1,2 \\ \eta_i = c_{02} + \beta_0\mathbf{x_i} & i=3,...,7 \\ \eta_i = c_{03} + \beta_0\mathbf{x_i} & i=8,9,10 \end{cases}$

Table 1 The three simulated models: A, B and C.

	model A			model B			model C	
	$\mathbf{c_0}$	$\mathbf{c_1}$	β_0	β_0	$\mathbf{c_0}$	β_1	$\mathbf{c_0}$	β_0
l=1	5	10			10		5	
l=2	2	5	-6	9	5	-6	2	-6
l=3	-10	1			1		-10	
l=1	4.96(1.06)	10.09(1.72)			10.73(1.61)		5.00(0.61)	
l=2	1.73(1.94)	5.03(1.03)	-6.15(0.73)	9.54(1.25)	5.29(0.95)	-6.21(0.56)	2.02(0.28)	-6.01(0.48)
l=3	-10.27(1.40)	0.04(0.53)			0.12(0.48)		-10.09(1.11)	

Table 2 Fixed and random effects coefficients: simulated (above) and estimated by the algorithm (below), in the three different models. Estimates are reported in terms of mean(std) computed on the 100 runs of the simulation in which the algorithm identifies 3 subpopulations.

5 An application to the heart failure hospitalizations in Lombardy Region

Data were supplied by Regione Lombardia (Italy) - Healthcare Division [3] within the research project described in *Mazzali et al.* [4]. Hospital discharge forms with HF-related diagnosis codes were the basis for identifying HF hospitalizations as clinical events, or episodes. In patients experiencing at least one HF event, we selected hospitalizations for causes related to cardiocirculatory pathologies (Clinical Classifications Software for ICD-9-CM codes [8] greater or equal than $'96'$ and less or equal than $'117'$). Moreover, we included in the analysis only patients who didn't leave the study before its end (except for death) and didn't change hospital at different hospitalizations. Our variable of interest is the binary variable that tells us whether a patient had at least a secondary hospitalization in the 365 days after the first one. We selected hospitals that received at least two patients (≥ 2) and in which at least a patient had a secondary hospitalization; this choice was performed because, for each hospital, the SPGLMM requires a logistic regression to be fitted on the fixed covariates of the patients cured in such hospital: both the classes need to be present.

Collected data include 60813 patients (51.54% females, mean$\pm$sd age 75.33 ± 12.06, median age 78, range $18-105$) hospitalized in 218 structures. Each hospital received on average$\pm$sd 278.95 ± 372.33 different patients. We built the model as follows: for each patient, the sex and the normalized number of comorbidities out of the twenty most significant (e.g. dementia, coagulopathy, hypertension, psychosis, arrhythmia,...), recorded at the first hospitalization, were set as fixed covariates; a random intercept for the hospital has been considered. Multiple runs, with different D, were performed.

With $D = 0.2$, the SPGLMM identified six groups with numerosities 27, 113, 1, 12, 57 and 4; in addition, 4 hospitals were not assigned to any of the groups. We report the obtained fixed coefficients and the correspondent p-value computed through likelihood ratio test: $\beta_{\text{Sex}} = -0.2166$ (p-value$= 2.97 \cdot 10^{-22}$), $\beta_{\text{NComorbidities}} = 0.0395$ (p-value$= 4.09 \cdot 10^{-5}$) and random intercept for the six groups:

$$\begin{cases} c_{0i} = -0.7081 & i = 1,...,27 \\ c_{0i} = -0.7605 & i = 28,...,140 \\ c_{0i} = -0.2897 & i = 141 \\ c_{0i} = -2.3994 & i = 142,...,153 \\ c_{0i} = -1.3446 & i = 154,...,210 \\ c_{0i} = -0.7495 & i = 211,...,214 \end{cases}$$

6 Discussion and further developments

The simulation results reported in Section 4 show that the proposed SPGLMM, for different random effects and fixed covariates, well performs in the individuation of the simulated groups. The study shows that the preliminary results obtained by the application of the SPGLMM algorithm to the heart failure hospitalizations, performing a clustering on the hospitals, provide informative insights on their characteristics; such result could in future support decision policies. Different numbers of subpopulations were detected for different thresholds D.

Developments of this work include the analysis of the selection of the best threshold D and other methods for the evaluation of the providers (i.e. sets of hospitals) performances, such as their profiling, outlier detection or the computation of the PVRE (Percentage of Variation due to Random Effects). Further goals also concern the application of SPGLMM to integer outcomes such as a Poisson or a Zero Inflated Poisson where the number of hospitalizations, for a certain patient in a certain amount of time after the first hospitalization, is the response.

References

1. Masci, Chiara, Anna Maria Paganoni, and Francesca Ieva. "Semiparametric mixed effects models for unsupervised classification of Italian schools." Journal of the Royal Statistical Society: Series A (Statistics in Society) 182.4 (2019): 1313-1342.
2. Masci, Chiara, et al. "Does class matter more than school? Evidence from a multilevel statistical analysis on Italian junior secondary school students." Socio-Economic Planning Sciences 54 (2016): 47-57.
3. Regione Lombardia. HFData project: Utilization of Regional Health Source Databases for Evaluating Epidemiology, short- and medium-term outcome and process indicators in patients hospitalized for heart failure. Progetto di Ricerca Finalizzata di Regione Lombardia - HFData-RF-2009-1483329 (2012)
4. Mazzali, Cristina, et al. "Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012." BMC health services research 16.1 (2016): 1-10.
5. Lloyd-Jones, D. et al. Executive summary: heart disease and stroke statistics–2010 update: a report from the American Heart Association. Circulation 121, 948–954 (2010).
6. Kalogeropoulos, A. et al. Progression to Stage D Heart Failure Among Outpatients With Stage C Heart Failure and Reduced Ejection Fraction. JACC Heart Fail 5, 528–537 (2017).
7. Azzimonti, Laura, Francesca Ieva, and Anna Maria Paganoni. "Nonlinear nonparametric mixed-effects models for unsupervised classification." Computational Statistics 28.4 (2013): 1549-1570.
8. Healthcare Cost and Utilization Project. "Clinical classifications software (CCS) for ICD-9-CM." (2018).

Issues in Data science

`etree`: Classification and Regression With Structured and Mixed-Type Data in `R`

`etree`: *Classificazione e Regressione Con Dati Strutturati e Misti in* `R`

Riccardo Giubilei, Tullia Padellini, Pierpaolo Brutti

Abstract As data continue to grow in complexity, so does the need for models and methods that directly account for their non-trivial structure because any simplification may induce loss of information. Energy Trees have been recently proposed following this principle as a unifying and statistically sound framework for classification and regression with structured and mixed-type covariates. In this work, we provide an illustration of `etree`, which is the R package where Energy Trees are implemented. We describe its origin, structure, main function, and other important features, such as methods for plotting and making predictions. The package currently covers functional data and graphs as structured covariates. However, thanks to its modular infrastructure, any other type of variable can be easily included.

Abstract *Con l'aumentare della complessità dei dati, cresce anche la necessità di avere modelli in grado di gestirne la struttura in quanto l'eventuale semplificazione può causare perdita di informazione. Seguendo questo principio, gli Energy Trees sono stati proposti recentemente come framework unificante e statisticamente rigoroso per fare classificazione e regressione con covariate di tipo strutturato e misto. In questo lavoro, presentiamo* `etree`*, ovvero il pacchetto R in cui gli Energy Trees sono implementati. Ne descriviamo le origini, la struttura, la funzione principale, e altre caratteristiche importanti, come ad esempio i metodi per costruire grafici e fare previsioni. Al momento, il pacchetto prevede l'utilizzo di dati funzionali e grafi come covariate strutturate. Tuttavia, grazie alla sua infrastruttura modulare, ogni altro tipo di variabile può essere facilmente integrato.*

Key words: Energy Trees, R package, Conditional Trees, Energy Statistics, Complex Data.

R. Giubilei
Luiss Guido Carli (Rome, Italy), e-mail: `rgiubilei@luiss.it`

T. Padellini
Bank of Italy (Rome, Italy), e-mail: `tullia.padellini@bancaditalia.it`

P. Brutti
Sapienza University (Rome, Italy), e-mail: `pierpaolo.brutti@uniroma1.it`

1 Introduction

Structured data objects such as functions, graphs, images, or time series are constantly growing in quantity and need of being analyzed. However, arbitrary simplifications, often resulting in loss of information, are usually required to apply standard statistical techniques. Energy Trees are a new statistical learning model that adequately accounts for the possibly non-trivial structure of the variables. More precisely, the model provides a unifying and statistically sound framework for classification and regression with structured and mixed-type data.

Energy Trees are a recursive binary partitioning model with piecewise constant fits. They are built starting from Conditional Trees [3] and replacing standard permutation tests with Energy tests of independence [7] from the Energy Statistics framework [6]. This choice allows for testing the association between variables defined in spaces that are not necessarily Euclidean, and not necessarily equal.

At each step of Energy Trees iterative procedure, an Energy test of independence is performed between the response variable and each of the J covariates. If the test of global independence (defined as the intersection of the J tests of partial independence) is not rejected at the specified significance level, the recursion is stopped; otherwise, the covariate most associated with the response in terms of p-value is selected for splitting. When the covariate is traditional, an Energy test of independence is performed for each possible split point, and the one yielding the strongest association with the response is chosen. When the selected covariate is structured, Energy Trees provide two alternative strategies: 1) transform the structured selected variable into Euclidean vectors through coefficient expansion, and then apply any splitting technique that works with numerical variables; 2) use distance-based clustering methods such as Partitioning Around Medoids (PAM) [5].

Energy Trees are implemented in R package `etree`, which is publicly available at `https://github.com/ricgbl/etree`. In this work, we briefly describe `etree`'s origin, structure, main function, and other important characteristics. Finally, we draw some concluding remarks and sketch directions for future work.

2 Package overview

R package `etree` is built upon `partykit` [4], which is the package where Conditional Trees are implemented. The new package attempts to follow `partykit` design principles, exploiting the rich and flexible infrastructure for objects of class `"constparty"`. However, Energy Trees are designed to work with structured and mixed-type variables, hence their implementation requires more flexibility and some further adjustments. First, the fundamental data structure is generalized from the `data.frame`s used in `partykit` to `list`s. A new class of objects, `"etree"`, allows defining characteristic methods. Energy tests of independence substitute permutation tests, and both splitting strategies are implemented. Additionally, specific handling is provided for each type of covariate considered in the package. The es-

sential exported functions are three: etree() (the main function), plot(), and predict(). The package also features about two dozens original internal functions and several modified versions of partykit internal functions.

The current version of etree covers functional data and graphs as structured covariates. However, thanks to a well-established and modular infrastructure, almost any other case can be easily implemented.

3 Main function

The main function in R package etree is the homonym etree(), which allows fitting an Energy Tree by simply specifying the response variable, the set of covariates, and possibly some other parameters. The function is specified in the same way regardless of the task type: the choice between classification and regression is automatically made depending on the nature of the response variable.

The function etree() is defined through the following call, where default values for arguments that admit one are included:

```
etree <- function(response,
                  covariates,
                  weights = NULL,
                  minbucket = 5,
                  alpha = 0.05,
                  R = 1000,
                  split_type = 'coeff',
                  coeff_split_type = 'test',
                  p_adjust_method = 'fdr')
```

The arguments of etree() can be shortly explained as follows:

- response: response variable, an object of class "factor" or "numeric" (for classification and regression, respectively);
- covariates: set of covariates, in the form of a list where each element is a different variable;
- weights: optional numeric vector of case weights;
- minbucket: minimum number of observations that each terminal node must contain;
- alpha: nominal level controlling the probability of type I error in the Energy tests of independence used for variable selection;
- R: number of replicates in every Energy test of independence;
- split_type: splitting method used when the selected variable is structured. Either "coeff" for coefficient expansion, or "cluster" for clustering;
- coeff_split_type: method to find the split point for the chosen component when the selected variable is structured and split_type = "coeff". Either "test", where Energy tests of independence are used, or "traditional", to employ Gini index for classification and RSS for regression;

- `p_adjust_method`: multiple-testing adjustment method for p-values, which can be set to any of the values provided by `stats::p.adjust()`. Default is `"fdr"` for False Discovery Rate.

The function `etree()` returns the full-grown fitted tree as an object of class `"etree"`. Methods for printing, plotting, and predicting with objects of this class are included in the package and explained in Section 5.

4 Distance and coefficient expansion techniques

When using Energy Trees, covariates can be in principle of any kind. A useful distinction is between traditional variables, including numeric and nominal, and structured ones, such as functions and graphs. Each type of covariate requires an appropriate distance, which is always necessary to compute the test statistics in the Energy tests of independence employed for variable selection. It is also needed by structured covariates when using the clustering approach to splitting. Additionally, structured covariates necessitate a proper technique for coefficient expansion when choosing that method for splitting.

Here, the types of covariates under consideration are four: numeric, nominal, functions, and graphs. Tables 1 and 2 include `etree` default choices of distance and coefficient expansion techniques, respectively, and the corresponding R functions used to implement them. Note that these choices can be easily changed to reflect personal preferences or application-specific needs, and that new types of variables can be quickly added.

Table 1 Choices of distance and corresponding functions for the four types of covariates available in the package.

	Technique	**R function**
Numeric	Euclidean	`stats::dist()`
Nominal	Gower	`cluster::daisy()`
Functional	L^2-norm	`fda.usc::metric.lp()`
Graphs	Edge Difference [2]	`NetworkDistance::edd()`

Table 2 Choices of coefficient expansion techniques and corresponding functions for the two types of structured covariates available in the package.

	Technique	**R function**
Functional	Cubic B-splines	`fda.usc::optim.basis()`
Graphs	Shell distribution [1]	`etree::graph_shell()`

5 Other functions

Fitted Energy Trees can be plotted by calling `plot()` on the `"etree"` object returned by `etree()`. Specifically, `etree` offers a `plot()` method for `"etree"` objects that is a modified version of the `partykit` one for the `"constparty"` class. Basic elements have been mostly kept intact from the original version. On the other hand, in `etree`, inner nodes are colored differently based on the splitting variable's type. Additionally, split points are on the edges, as in `partykit`, only for traditional variables, or for structured covariates when `split_type = "coeff"`; otherwise, the size of the corresponding kid node is displayed.

In `etree`, predictions are given by the `predict()` method for `"etree"` objects. When the `newdata` argument is not specified, `predict()` returns the fitted values for the response; otherwise, it computes predictions for the observations in the test set provided with `newdata`. The only step required to obtain both these types of results is as easy as calling `predict()` on the fitted Energy Tree returned by `etree()`, and possibly adding `newdata`. Then, the `etree` version of `predict()` automatically identifies the `split_type` used for fitting, and reacts accordingly. If `newdata` is provided and `split_type = "coeff"`, coefficient expansion is applied to the structured covariates contained in the test set, and, subsequently, splits are induced and predictions are made based on the rules established in the fitting procedure; when `split_type = "cluster"`, the distances between the medoids coming from fitting and the observations in the test set are calculated, and, subsequently, splits and predictions are obtained by assigning the new observations to the closest medoid at each split.

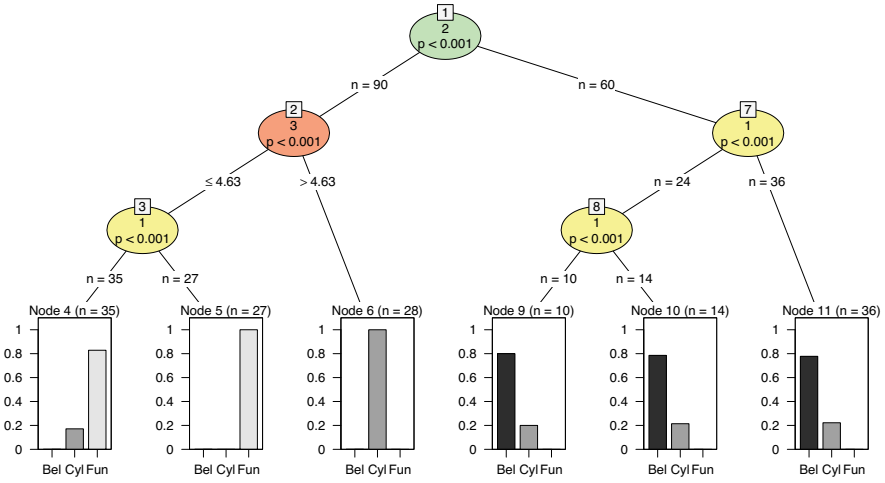


Fig. 1 Example plot of an Energy Tree fitted with `split_type = "cluster"` on a simulated classification dataset.

6 Concluding remarks and future work

The aim of this work is to illustrate R package `etree` for Energy Trees. Despite the complexity that Energy Trees potentially have to deal with, the interface and the functionalities for users are kept as simple as possible. The rich infrastructure of `partykit` has been embedded and generalized, enabling further possibilities. However, much room is left for improvement.

The graphical compartment should be enhanced, possibly using `ggplot2` and `ggparty` package to provide fancier and dynamic plots. A procedure for missing values handling through surrogate splits should be considered. The computational complexity could be improved in many ways: using faster versions of PAM, such as CLARA and FastPAM, for clustering; further optimizing the internal functions; compiling the package's underlying processes in C. Finally, `etree` is expected to implement any methodological development within the Energy Trees framework, which may include the introduction of new types of structured covariates and the case where also the response variable is structured.

References

[1] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.

[2] D. K. Hammond, Y. Gur, and C. R. Johnson. Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *2013 IEEE GlobalSip*, pages 419–422, 2013.

[3] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[4] T. Hothorn and A. Zeileis. `partykit`: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16:3905–3909, 2015.

[5] L. Kaufmann and P. Rousseeuw. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 1987.

[6] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[7] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Deep Learning framework for ungrouping coarsely aggregated vital rates.

Un modello di deep learning per disaggregare tassi raggruppati in classi.

Andrea Nigri

Abstract The aim of this study is to develop a versatile methodology for ungrouping vital rates leveraging deep learning algorithms based on neural networks.
Abstract Lo scopo di questo studio è quello di sviluppare una metodologia versatile per disaggregare dati demografici usando algoritmi di deep learning basati sulle reti neurali.

Key words: Ungrouping, Vital Rates, Deep Neural Network.

1 Introduction

The monitoring of changes and inequality among populations is a prime aim to assess population dynamics and social and public policies. Thus reliable predictions of age-specific vital rates are crucial in demographic studies. Data deficiency usually refers to incomplete or misreported information such as age exaggeration and age heaping, in both death and population data. Despite single age-specific data being desirable, ages are commonly grouped in bins of 5 years in most cases. This is the case of demographic data, which are followed by a broad or open-ended age class for older ages. Several methods have been proposed for the disaggregation problem about historical data and data from developing countries that lack functional systems of vital registration ([9]; [13]; [8]; [12]). Using countries from the Human Fertility Database (HFD), [8] derived age-specific fertility rates from abridged data comparing 10 different methods. The authors concluded that the modified Beers method ([3]) provided the best fit. Similarly, [12] using schedules observed in the HFD and in the US Census International Database (IDB), proposed a calibrated spline (CS) as a more accurate and flexible alternative to Beers interpolation method that requires more computation. Among mortality modeling, the main approaches aimed at ungrouping

Andrea Nigri
Bocconi Univeristy.e-mail: andrea.nigri@unibocconi.it

histograms or abridged life tables were based on parametric assumptions for the underlying distribution ([6]; [7]; [5]). For fitting a nonparametric density to binned data histosplines ([2]), kernel density estimators ([1]), and local likelihood estimation. One of the most prominent frameworks has been proposed by [11] whom developed a versatile method for ungrouping histograms based on the composite link model, with a penalty added to ensure the smoothness of the target distribution. Estimates are obtained by maximizing a penalized likelihood.
Nevertheless, estimations often require harmonizing, comparability, and coherence among countries and gender. Working on a single-gender population, the aforementioned models do not guarantee the latter properties, jeopardizing the comparability across time and among countries. Thus relying on coarsely grouped data may hinder accurate data analysis. Aiming at providing a more comprehensive perspective on the indirect age-specific vital rates estimation from grouped data, this paper introduces a multi-population (countries and genders) approach for splitting the abridge demographic rates. The model leverages deep learning algorithms based on deep neural networks (DNN) to uncover age-specific vital rates.
The proposed model represents an advance in mortality modeling, offering the advantage of an indirect and complementary way to approximate death rates specific for age. It can be valuable in contexts where population-level mortality studies are hindered by financial or time constraints, for national registries that do not support the open data system.

2 Method

In this section, I will describe the DNN model used to derive the country-specific ungrouped death rates by age and time. Let $\mathbf{M^{(1)}} = \log(m_{a^{(1)},t})_{a^{(1)} \in A^{(1)}, t \in T}$ be a matrix with death rates for single ages, where rows denote age and columns calendar years, and $\mathbf{M^{(5)}} = \log(m_{a^{(5)},t})_{a^{(5)} \in A^{(5)}, t \in T}$ the matrix with grouped death rates in 5 years age classes. We set $A^{(1)} = \{0, 1, ..., 100\}$, $A^{(5)} = \{0, 1-4, 5-9, ..., 100-110\}$ and $T = \{t_1, t_2, ..., t_n\}$. Then, for an DNN architecture composed of three hidden layers $k = 3$, the theoretical relationship defining the matrix $\mathbf{M^{(1)}}$ given the vector of 5-years age grouped rates $\mathbf{M^{(5)}}$ is represented by:

$$\mathbf{M^{(1)}} = f^{(3)}(\mathbf{W^{(3)}}(f^{(2)}(\mathbf{W^{(2)}} f^{(1)}(\mathbf{W^{(1)}}\mathbf{M^{(5)}} + \mathbf{b^{(1)}}) + \mathbf{b^{(2)}}) + \mathbf{b^{(3)}}), \qquad (1)$$

where $f^{(k)}$ is the activation function, $W^{(k)}$ is the matrix of weights, and $b^{(k)}$ is the bias used to control the triggering value of the activation function. Thus, $f^{(1)}(\mathbf{W^{(1)}}\mathbf{M^{(5)}} + \mathbf{b^{(1)}}) = \mathbf{H^{(1)}}$ is the first hidden layer that accepts the vector $\mathbf{M^{(5)}}$ as input.

The DNN model is based on a training algorithm that involves an unconstrained optimisation problem aiming to minimise the prediction error. The idea is to adjust the weights of the network connections to minimise a measure of the difference between the actual and the desired output ($\mathbf{M^{(1)}}$ and $\hat{\mathbf{M}}^{(\mathbf{1})}$), respectively, known as the loss function $\mathcal{L}$. Here, I present a simple example of how to extend the DNN

model for the multiple populations case. Consider the case in which the full Human Mortality Database (HMD) is used to train a model and then to predict the single age-specific mortality profile. I still model the functional relationship between coarsely binned data and single-year age grouped rates, as numerical inputs and outputs. Thus, I extend the framework to the multi population model by adding other demographic features, such as country, year and sex, using what is known as embedding layer, that ([10]) allows a low-dimensional representation learning, mapping categorical variables into a vector space. In the present study I deal with the following variables:

$$c \in \mathcal{C} = \{\text{Italy,..,Russia}\}; g \in \mathcal{G} = \{\text{Male, Female }\}; a \in \mathcal{A} = \{0,..,110\}; t \in \mathcal{T} = \{1990, ..., 2015\}. \tag{2}$$

Embedding layers map these features into real-valued vectors, where for instance $z_C(c)$ is the new representation of countries. Therefore, $i = (c,g,a,t) \in I = \mathcal{C} \times \mathcal{G} \times \mathcal{A} \times \mathcal{T}$ might be considered the categorical features space. Once embedding vectors have been defined for each categorical variable, all variables, categorical and not, are concatenated into a single feature vector: $x_{\mathbf{M}^{(5)},i} = (\mathbf{M}^{(5)}, z_I(i))$, which is used as input to the sub-neural network (previously described) in order to predict the ungrouped death rates in year t, at age a for country c gender g related to levels of observed coarsely binned data.

2.1 Implementation

Let $\left\{\log(m_{a^{(5)},t})\right\}_{t=t_0}^{t_s}$, for $t_0 < t_s$, be the country-specific observed grouped death rates. Then, each series is split into a train-validation set and a test set, where the first one is used for fitting the model's parameters, while the second one to test the model's prediction and calculate the error. Specifically, the time frames 1970-2000 is used as a train-validation set, according to the 80%-20% splitting rules. The best DNN setting during the training phase is used to obtain predictions in the test phase which takes place on the time frames 2001-2015.

Hence, let t_τ, with $t_0 < t_\tau < t_s$, be the calendar year corresponding to the last realization in the training-validation set. The values of $\log(m_{a^{(5)},t})$ over the period (t_0, t_τ), $\left\{\log(m_{a^{(5)},t})\right\}_{t=t_0}^{t_\tau}$, represent the input for train-validation, while the corresponding output is $\left\{\log(\hat{m}_{a^{(1)},t})\right\}_{t=t_0}^{t_\tau}$. The values of grouped death rates over a subsequent period, $\left\{\log(m_{a^{(5)},t})\right\}_{t=t_\tau+1}^{t_s}$, represent the input for test, while the corresponding output is $\left\{\log(\hat{m}_{a^{(1)},t})\right\}_{t=t_\tau+1}^{t_s}$. Thereby, denoting ψ_{nn} as a composition of functions defined on the basis of the NN architecture, the model can be described by:

$$\left\{\log(\hat{m}_{a^{(1)},t})\right\}_{t=t_\tau+1}^{t_s} = \psi_{nn}\left(\left\{\log(m_{a^{(5)},t})\right\}_{t=t_\tau+1}^{t_s} \middle| \hat{W}\right) \tag{3}$$

where $\left\{\log(\hat{m}_{a^{(1)},t})\right\}_{t=t_\tau+1}^{t_s}$ is the matrix of death rates in single year age group, in the test set obtained by ψ_{nn}, that involves the NN weights $\hat{W}$ estimated during the network training.

3 Results

I consider historical mortality data collected by the [4] for all available countries and both genders. Aiming to assess the multi population model robustness and consistency toward the historical data, I carry out an out-of-sample test by using 30 years for training-validation (1970-2000) and the remaining years (2001-2015) are used for model forecasting/test. I now validate the DNN model forecasting performance using illustrative applications. These examples are dedicated to investigating whether the approaches can capture (a) regular and irregular trends over time (b) dynamics of age-specific mortality improvements. The analysis includes numerical and graphical representations of the goodness of fit. To assess the models' accuracy, I calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on the out of sample period, which in the present analysis corresponds to 2001-2015 time window. Despite the DNN estimations derived from the DNN training on the whole HMD, here, for brevity's sake, I provide results for only six countries in table 1. Overall, the DNN provides remarkable accuracy.

Table 1: Out-of-sample test: MAE and RMSE for DNN, by country and gender.

Country	**Male**		**Female**	
	MAE	RMSE	MAE	RMSE
Australia	0.083	0.120	0.106	0.168
Denmark	0.126	0.223	0.150	0.247
Italy	0.067	0.103	0.0836	0.120
Japan	0.072	0.095	0.106	0.127
Russia	0.074	0.093	0.083	0.104
USA	0.06	0.075	0.068	0.081

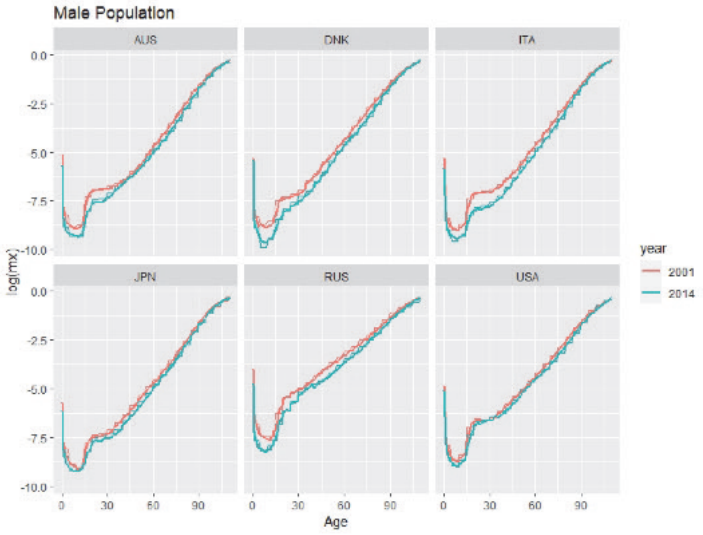


Fig. 1: Estimated ungrouped mortality rates, by countries for 2001, 2014 based on training period 1970-2000. The stepped-line refers to the observed rates collected in 5-year age groups. Solid lines are the DNN reconstructed ungrouped death rates

4 Discussion

This manuscript contributes to the current literature on the demographic methods for ungrouping vital rates. I propose a Deep Neural Network framework to ungroup death rates from rates collected in 5-year age groups, using a multi population approach. This method represents an advance among mortality modeling, indeed the model could be used to estimate vital rates for single ages in regions or subpopulations where present information is lacking but past data are available or from surrounding countries or populations.

References

1. Blower G, Kelsall JE. Nonlinear kernel density estimation for binned data: convergence in entropy. Bernoulli. 2002;8(4): 423–449.
2. Boneva LI, Kendall DG, Stefanov I. Spline transformations: three new diagnostic aids for the statistical data-analyst. J R Stat Soc Series B. 1971;33(1):1–71.
3. de Beer C (2011). A new relational method for smoothing and projecting age-specific fertility rates: TOPALS. Demographic Research 24(18): 409-454.
4. Human Mortality Database (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 01/02/2022. urlhttps://www.humanmortality.org.

5. Hsieh JJ. Construction of expanded continuous life tables a generalization of abridged and complete life tables. Math Biosci. 1991;103(2):287–302.
6. Kostaki A. The Heligman-Pollard formula as a tool for expanding an abridged life table. J Off Stat. 1991;7(3): 311–323.
7. Kostaki A, Panousis V. Expanding an abridged life table. Demogr Res. 2001;5(1):1–22.
8. Liu Y, Gerland P, Spoorenberg T, Kantorova V, Andreev K (2011). Graduation methods to derive age-specific fertility rates from abridged data: a comparison of 10 methods using HFD data. Presentation at the First HFD Symposium, MPIDR, Rostock, Germany.
9. McNeil, Donald R, Trussell TJ, Turner JC (1977). Spline interpolation of demographic data. Demography 14(2): 245–252.
10. Richman, R. (2021). Mind the Gap - Safely Incorporating Deep Learning Models into the Actuarial Toolkit. Available at SSRN: https://ssrn.com/abstract=3857693
11. Rizzi, S., Gampe, J., and Eilers P. H. C. (2015). Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. American Journal of Epidemiology. Vol. 12, No. 2
12. Schmertmann C (2012). Calibrated spline estimation of detailed fertility schedules from abridged data. MPIDR Working Paper WP-2012-022. Rostock, Germany.
13. Smith L, Hyndman R, Wood S. (2004). Spline interpolation for demographic variables: the monotonicity problem. Journal of Population Research 21 (1), pp. 95–97.

Inside the metaverse: analysis of the state of the art and development of a new usage approach based on quality and ethics

Dentro il metaverso: analisi dello stato dell'arte e sviluppo di un nuovo approccio di utilizzo basato su qualità ed etica

Vito Santarcangelo[1], Emilio Massa[1], Saverio Gianluca Crisafulli[1], Antonio Ruoto[1], Angelo Lamacchia[1], Alessandro D'Alcantara[1], Alessandro Verderame[1], Massimiliano Giacalone[2]

Abstract The paper analyses, starting from the first coining of the term, the state of the art related to the technologies and possible approaches of the metaverse, with particular focus on the potentialities expressed by blockchain-based metaverses and their criticalities. The paper then discusses in detail the concept of tokens, in order to explain the dynamics of regulation of ownership and transactions within the metaverse, up to the exposition of the revolutionary concept of ethical metaverse, object of industrial patent, based on the concepts of quality and social responsibility in a Benefit Impact (BIMPACT) Assessment perspective.

Abstract *Il paper analizza, partendo dalla prima coniazione del termine, lo stato dell'arte relativo alle tecnologie e ai possibili approcci del metaverso, con focus particolare alle potenzialità espresse dai metaversi basati su blockchain e alle relative criticità. Il lavoro tratta dunque nel dettaglio il concetto di token, al fine di spiegare le dinamiche di regolamentazione della proprietà e delle transazioni all'interno del metaverso, fino ad arrivare all'esposizione del rivoluzionario concetto di metaverso etico, oggetto di privativa industriale, basato sui concetti di qualità e di responsabilità sociale in ottica Benefit Impact (BIMPACT) Assessment.*

Key words: metaverse, blockchain, smart contract, NFT, SFT, token, digital twin, ar cloud, virtual reality, augmented reality, IPFS, ethics, quality, metavalue, CSR, BIMPACT

---

[1] V. Santarcangelo, E. Massa, S.G. Crisafulli, A. Ruoto, A, Lamacchia, A. D'Alcantara, A. Verderame, iInformatica Srl, Trapani (Italy); vito@iinformatica.it;

[2] M. Giacalone, Università degli Studi Federico II, Napoli (Italy); massimiliano.giacalone@unina.it

1 Introduction

One of the current trends in the technologic world is linked to the concept of metaverse, an innovative digital interactive experience. The term metaverse was coined by Neal Stephenson in Snow Crash (1992) [1], the sci-fi cyberpunk book described as a kind if virtual reality shared through internet, where you are represented in 3D by your own avatar and where everyone can create everything they wish; as a prototypical form of metaverse we can consider Second Life, a MUVE (multi-user virtual environment) born in 2003 where it is possible, logging in through your own digital avatar, to explore a virtual world, to interact with other users, to join activities and to create or exchange goods and services [2].

Even if the technology available during the development of Second Life allowed a quite good interaction with other users thank to the maturity of internet-related technology and to a good digital reproduction of real places thanks to graphic engines, several mechanics of real life were missing, making the title tend more towards a video game and less towards a virtual reality or metaverse. The first difference with the current metaverse was the lack of digitization of the natural right of ownership, that includes the artificial creation of scarcity, of the uniqueness, of the value, of a currency and other features that bring with them some technical challenge, and has been overcome thanks to blockchain technology; the second one, on the other hand, concerned the perception and immersion in the digital environment and the fruition of such contents, which has been bridged thanks to the evolution of hardware devices for the fruition of digital contents such as VR viewers, different sensor devices and new software technologies such as AR or AI [3]. The evolution of the concept of metaverse is therefore linked to all the mentioned technologies that allow to virtualize the features of reality and make all its virtual aspects more immersive and real.

The aim of this paper is to carry out a state-of-the-art analysis on metaverse technologies and approaches, highlighting their potentials and limitations, clarifying the concept of token (regulation of ownership and transactions in the metaverse) and introducing a revolutionary concept of ethical metaverse based on the concepts of quality and social responsibility (evolution of the BIMPACT metric defined by B Lab [19]).

2 Technologies of the metaverse

The concept of metaverse in recent years is destined to evolve into the coexistence of real and virtual contents, overcoming the concept of gamification of virtual environment (simple implementation in computer graphics of a non-real environment). In order to enable this vision, a digital model of the physical/real world is needed, where immersive 3D technologies, such as VR and AR, can add those digital interactions that can create an infinite set of real-world based experiences [4]. The real world-based metaverse involves the possibility of creating

simulations and interactions with the reality around us, going beyond what is possible in Second Life. To make this evolution of metaverse possible, three new technologies are fundamental: Digital Twin, AR Cloud and Virtual Reality. Digital Twin is a 3D technology able to create a digital model of an entity that exists in the real world, in order to interface it through information layers, sensors and IoT systems. A Digital Twin allows to acquire real-time information from the physical model to interact with it through digital systems [5]. The AR Cloud, also known as spatial computing, is an augmented reality technology implemented in the cloud, capable of adding information layers to the real world, creating hybrid environments in which one can simultaneously use traditional tools, augmented reality and virtual reality, with the particularity that the digitally generated contents are persistent, since the data are tracked with respect to the physical environment and stored in the cloud, and are always accessible by every user in the metaverse [6]. An example is the digital platform Spatial, a well-known AR collaboration tool. Virtual Reality is the one that can guarantee the highest level of immersion and presence in the digital experience, replacing the view of the real world with a digitally produces scene that can represent a duplicate of the real world or places from a totally imaginary universe [7]; it can be used through special devices, such as visor, and in combination with the Digital Twin and the AR Cloud it provides the flow from the real world to the virtual world.

By exploiting the various potentials of these technologies, it is possible to implement the real/virtual worlds of the metaverse.

3 Metaverse and blockchain

Another technology that characterises the metaverse is the blockchain [18]. If a gaming metaverse (Minecraft, Roblox) unites multiplayer dynamics with the possibility for the users to create contents [8], a blockchain metaverse (Decentraland, The Sandbox) allows its users to own properties just like in the real world; it is possible using the blockchain technology and in particular using the smart contracts, through which it is possible to create digital currencies (token) and collection of assets (NFTs and SFTs) describing their behavior and characteristics, making them customizable, guaranteed by the decentralized blockchain network and independent of the single platform hosting a metaverse, thus visualizing a reality with more similar features to the one we know but richer and without limits [9].

The tokens of the metaverse are the basis for the functioning of virtual worlds and are mainly divided into three categories. The Governance Tokens, similar to shares, is basically a digital asset that allows token owners to vote and to take decisions that will influence the future of the protocol to which the token corresponds. System Tokens, which are needed to perform actions within the system, such as selling and buying in internal marketplaces and developing new functions, and whose value is essentially linked to the dynamics of the metaverse. Finally, NFTs (non-fungible tokens), which are unique digital objects whose

transaction/ownership is recorded on the blockchain. NFTs have played a key role in enabling the market for digital resources, that would have been reproducible without the possibility of recognizing their actual ownership otherwise. An NFT can for instance be a work of art, as well as a real estate property [10].

Most of the metaverses use one or more standards for their tokens, that allows to describe the functions and the events that the smart contract can implement [11]. The most used standards are ERC-20 for the fungible tokens, ERC-721 for the non-fungible tokens (NFT) and finally ERC-1155 for semi-fungible tokens (SFT), i.e. having the functionality of ERC-20 and ERC-721 to obtain the property of semi-fungible token. The blockchain is therefore used to record transactions and allows token holders to exchange them without any restriction. However, it does not allow the asset corresponding to the registered NFT/SFT to be saved within the transaction, limiting itself to a unique information on the same. In this regard, is introduced the IPFS (Interplanetary File System) protocol, which allows to decentralize data storage where instead of pointing to a URI, one points to a content hash [12]. This information is unique, immutable, and always accessible. All transactions are regulated by smart contracts: from buying and selling activities, to commercial agreements, to private agreements, everything goes through the writing of a smart contract whose nature is absolutely unchangeable and conditioned in time to the rules established in the contract itself, ensuring fairness and security between the parties involved [13].

A blockchain metaverse bases its success on participation, it has an autonomous and independent economic system whose main resource is land, a real estate property that allows it to give place and shape to its virtual activities, characterized by a real estate market that fully reflects the logic of operation typical of real estate, with areas/properties that are worth more than others. The properties of a metaverse are NFTs that can be traded on platforms such as OpenSea. All meta-verses are characterized by the presence of avatars, constructions, vehicles and objects of any nature, without limits in terms of creativity; these resources, defined as assets, can be created and sold in the form of NFTs and constitute an essential engine for the economy of the metaverse [14]. Each metaverse is therefore characterized by its own creative style, and from this technological infrastructural complexity/variety we derive the different interpretations/definitions that characterize the metaverse. The winning metaverse will therefore be the one equipped with a highly dynamic, technological and quality infrastructure, that is flexible to the various demands of the market and of its users. The concept of quality of the metaverse is particularly important, as it is closely related to its user experience and compliance, i.e. to the technologies used to guarantee a high level of UX (user experience), which can be calculated by considering aspects of usability and sentiment analysis feedback, together with strict compliance with mandatory regulations (lawfulness of the actions that can be carried out through the metaverse), personal data, commercial/industrial secrecy and intellectual property managed through it. These parameters represent, in fact, the main criticalities of the metaverse that open up important analysis scenarios to be addressed.

4 Prospects for an ethics-based metaverse: an experimental quality model

As mentioned in the previous paragraph, the development of the metaverse will most likely depend on the level of trust of the users [15], trust that cannot be consolidated and based on mere gamification and economic logic [16].

For this reason, in this paragraph we introduce a model of experimental development of an ethical metaverse that aims to enhance the territorial experience in order to raise awareness of the geography, products and customs of a territory, using the metaverse as an international showcase by putting in place value actions based on ethics and not on economic values, on what we could therefore define as "metavalues" related to ethics and know-how, which can then allow those who make ethics their goal, in reality and in the metaverse, to be able to interact more and distinguish themselves by virtue [17].

A valoristic dimension that rewards those rich in virtues, and allows business or interactions to develop only if in real life or in the metaverse itself, models of social responsibility or scientific progress and dissemination are followed. Engaging in the real dimension by achieving scores in a BIMPACT key, proving service activities, and contributing to the progress of the community in a technical-scientific perspective thanks to the development of new technology therefore represents a dimension of meritocratic enhancement of the ethical metaverse through the production of meta-values. A mathematical formulation of the concept of metavalue can be expressed by the sum of the contributions of two weighted (Θ,τ) integrals calculated on two-time intervals of reference t1 and t2, which integrate the temporal trends over time of actions of social responsibility (csr) and know-how (kh) carried out in reality (ar) or in the metaverse (av). The trends are therefore user behavioural functions that represent the tokens of user actions in the metaverse and map their actual certified actions in the real world.

$$mvl = \theta \int_{t1}^{t2} (ar_{csr}(t) + ar_{kh}(t))dt + \tau \int_{t1}^{t2} \left(av_{csr}(t) + av_{kh}(t)\right) d(t)$$

The core of this model is based on the fundamental concept of metavalue. Metavalue is the coinage of this ethical metaverse and is a function of the level of ethics expressed in the real world and/or in the metaverse, as well as the know-how that can be expressed through work tables, training provided and publications certified in a smart contract perspective in the metaverse or by intellectual property and scientific publications that can be obtained in an OSINT perspective or from institutional databases (e.g. UIBM, SIAE). Technology transfer thus becomes proof of know-how and commitment to the progress of science and technology. Then, metavalue is obtained considering in a fiscal year (t1,t2) the benefit investments (from fiscal balance) and BIMPACT score adding the contribution know-how (number of patents granted and papers' citations in that year). The value obtained is combined by score calculated by metaverse. This virtuous metaverse becomes a source of quality and an example of best practice, with a view to the social responsibility of the individual and of the company.

5 Conclusion

This work aims to provide a synthetic representation of the metaverse complexity, its potential, some of its criticalities and a possible model from a quality and ethical perspective. This shows how this tool can represent a new trend for developers in the IT world, on a par with operating systems to be adapted to the required purposes. It is therefore possible to imagine specific productive metaverses based on Industry 4.0, investment metaverses to create a network of investors, companies and consultants from all over the world, and also possible applications for the improvement of culture, research into healthcare (through research teams and hospitals working together in real-time synergy in the metaverse) and for geopolitical purposes and institutional management of relations thanks to easier sharing of virtual experience.

References

1. Stephenson, Neal. Snow crash: A novel. Spectra, 2003.
2. Descy, Don E. "Second Life." TechTrends 52.1 (2008): 5.
3. SIVASANKAR, GA. "Study Of Blockchain Technology, AI and Digital Networking in Metaverse." (2022).
4. Mystakidis, Stylianos. "Metaverse." Encyclopedia 2.1 (2022): 486-497.
5. Qi, Qinglin, et al. "Enabling technologies and tools for digital twin." Journal of Manufacturing Systems 58 (2021): 3-21.
6. Biggio, Federico. "Protocols of immersive web: WebXR APIs and the AR Cloud." Proceedings of the 3rd International Conference on Web Studies. 2020.
7. Burdea, Grigore C., and Philippe Coiffet. Virtual reality technology. John Wiley & Sons, 2003.
8. Han, Jeongmin, Jeongyun Heo, and Eunsoon You. "Analysis of Metaverse Platform as a New Play Culture: Focusing on Roblox and ZEPETO." Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021). CEUR Workshop Proceedings, Da Nang, Vietnam (Oct 2021). 2021.
9. Ynag, Qiinglin, et al. "Fusing Blockchain and AI with Metaverse: A Survey." arXiv preprint arXiv:2201.03201 (2022).
10. Ante, Lennart. "Non-fungible token (NFT) markets on the Ethereum blockchain: Temporal development, cointegration and interrelations." Available at SSRN 3904683 (2021).
11. Norvill, Robert, et al. "Standardising smart contracts: Automatically inferring ERC standards." 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE, 2019.
12. Vimal, S., and S. K. Srivatsa. "A new cluster p2p file sharing system based on ipfs and blockchain technology." Journal of Ambient Intelligence and Humanized Computing (2019): 1-7.
13. Buterin, Vitalik. "Ethereum white paper." GitHub repository 1 (2013): 22-23.
14. Jeon, Hyun-joo, et al. "Blockchain and AI Meet in the Metaverse." Advances in the Convergence of Blockchain and Artificial Intelligence (2022): 73.
15. Morozov, Evgeny. “The Net Delusion: The Dark Side of Internet Freedom”. PublicAffairs (2011).
16. Barberis, Mauro. Ecologia della rete: Come usare internet e vivere felici. Mimesis, 2021.
17. Santarcangelo,V. et. al (2022), Sistema metaverso etico semantico adattativo, UIBM.
18. Giacalone, M. et al. (2021), Big data for corporate social responsibility: blockchain use in Gioia del Colle DOP, Quality & Quantity volume 55, pages1945–1971
19. Wiman,A. et al. (2017), The B Impact-Assessment - a GLOBAL VALUE tool showcase