SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università
degli Studi
della Campania
Luigi Vanvitelli

SIS 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

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN Draft

Contents

Investigating effects of air pollution on health: a challenge for statisticians 77

Explainable Artificial Intelligence methods 98

Advances in Classification and Data Analysis 109

Family demography 133

New Frontiers in the theory of composite indicators 164

4 Contributed Sessions 848

Applications in Machine Learning 849

Bayesian modelling and inference 1 874

Methods of causal inference 905

Bayesian modelling and inference 2 1024

Advances in Categorical and Preference data 1056

Covid-19 Assessment and Evaluation 1 1087

Methods and applications in economics and finance 1362

Latent Class models 1393

Methods for health studies 1424

Applications in Economics 1967

Image analysis and visual methods 1999

Statistical assessment and empirical estimation 2027

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

2

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 μg/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.

Francesca Dominici

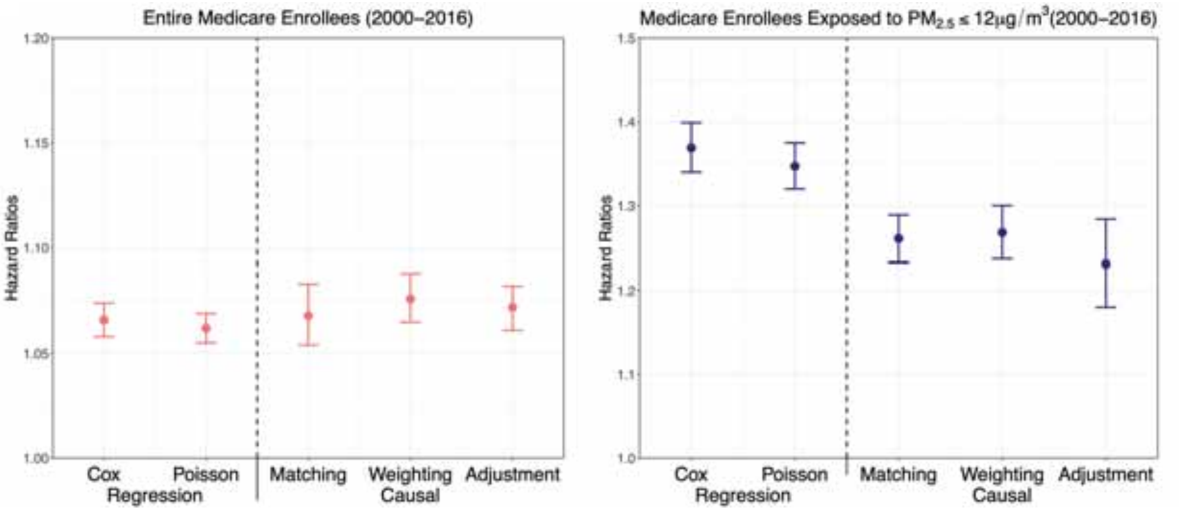


Figure 2: HR and 95% CIs.
The estimated HRs were obtained under five different statistical approaches (two traditional approaches and three causal inference approaches). HRs were adjusted by 10 potential confounders, four meteorological variables, geographic region, and year.

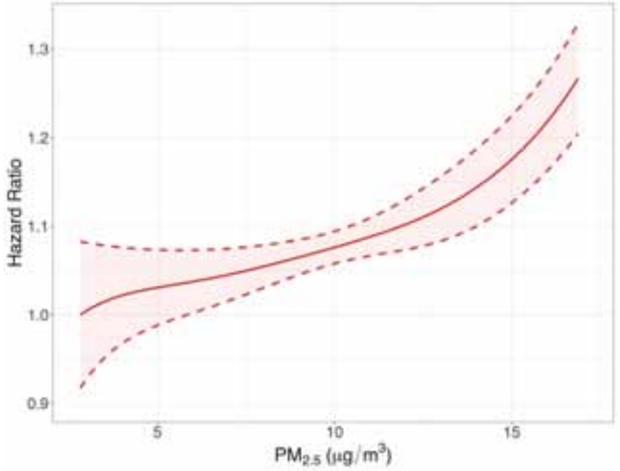


Figure 2. Estimated causal exposure-response curve relating $PM_{2.5}$ to all-cause mortality among Medicare enrollees (2000–2016) with associated 95% confidence bands obtained via bootstrap, only adjusting for one pollutant NO_2 as a potential confounder. We define the baseline rate as the estimated 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 μg/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

11

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

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 ormai ampiamente stabilito che i processi cerebrali sono caratterizzati da eterogeneit a pi 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'eterogeneit nelle neuroscienze, in cui si presume che i dati delle serie temporali siano organizzati in unit 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 unit neuronali. In primo luogo, discuteremo modelli per l'analisi multi-soggetto che identificheranno sottogruppi di*

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

popolazione caratterizzati da modelli di attivit 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 $Y_{iv} = (Y_{iv1}, \dots, Y_{ivT})^T$ be the $T \times 1$ vector of the BOLD response data at the v-th region in the ith subject, with $i = 1, \dots, N, v = 1, \dots V$, 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], $\beta_{iv} = (\beta_{iv1}, \dots, \beta_{ivp})^T$ is a $p \times 1$ vector of regression coefficients and $\varepsilon_{iv} = (\varepsilon_{iv1}, \dots, \varepsilon_{ivT})^T$ 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 hierarhical 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, DAngelo 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 ($A_t > 0$). DAngelo 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.

DAngelo 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].

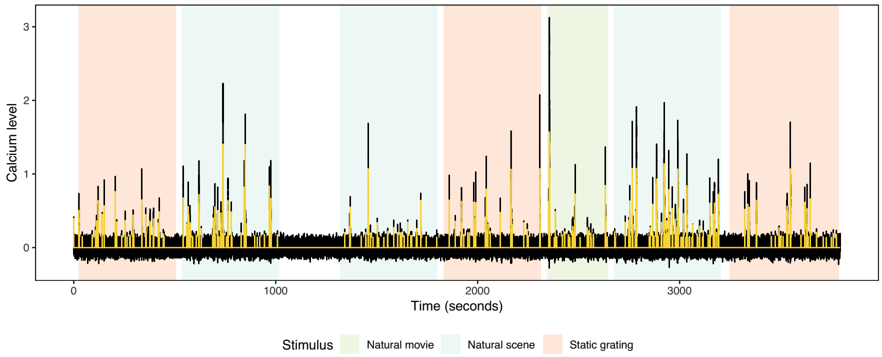


Fig. 1 Observed fluorescence trace y_t from the Allen Brain Observatory data (black line), and visual stimulus (shaded areas). The yellow line represents the estimated spike train in [15].

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

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

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 accounting 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

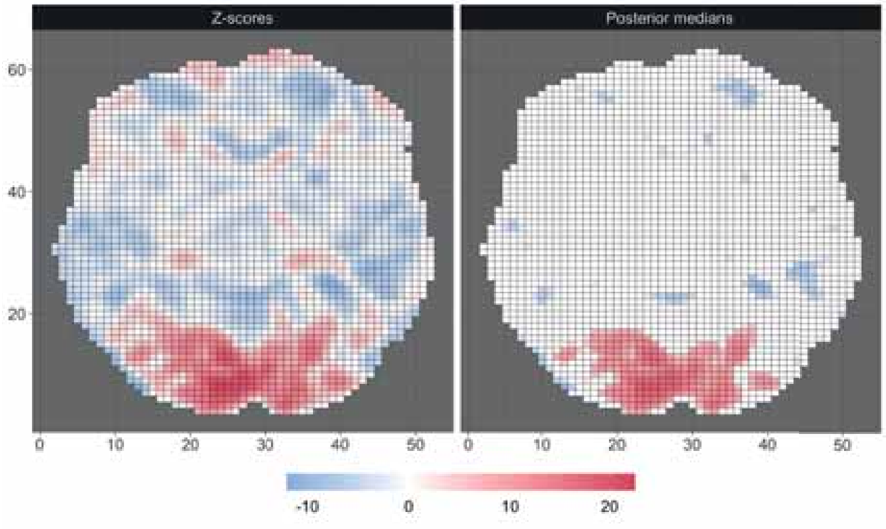


Fig. 2 Heatmaps comparing the pixel-specific z-scores (left panel) versus their regularized estimates obtained with the HSmix model (right panel - posterior medians).

(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

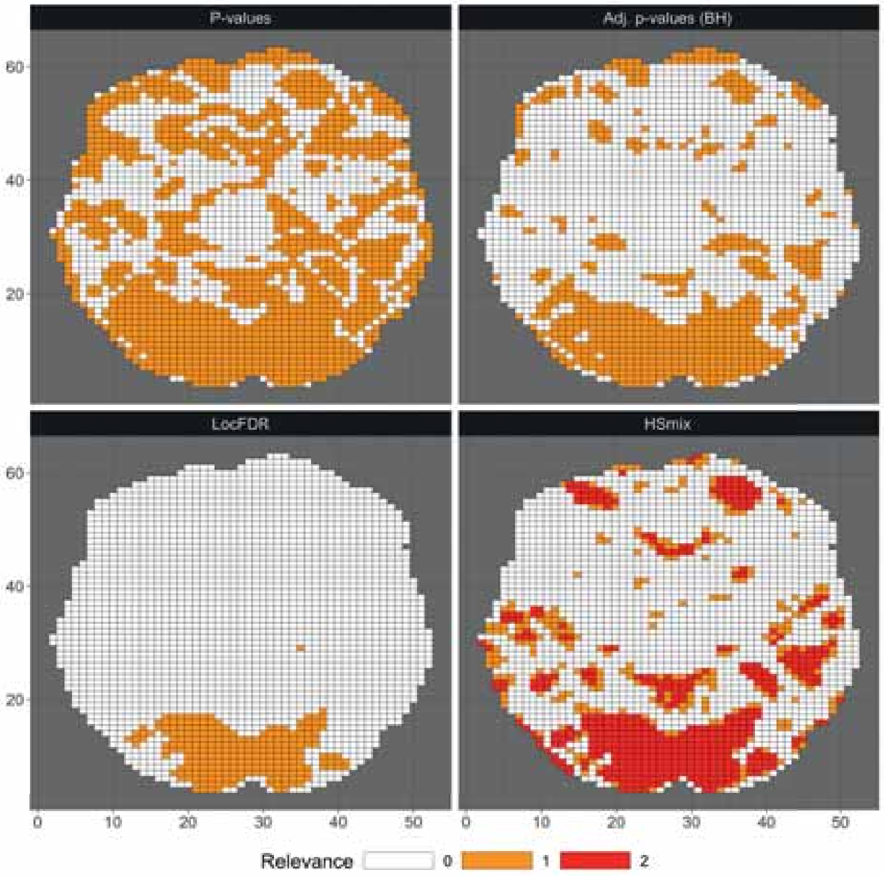


Fig. 3 Comparison of the discoveries obtained with four different screening procedures. The Relevance level 0 indicates no activation.

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, Tlay 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 DAngelo, 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 Adal, 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 factor 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 $\Sigma = \mathrm{diag}(\sigma_1^2,\dots,\sigma_p^2)$ $(\sigma_j^2 > 0)$ a diagonal covariance matrix. Let $\mathscr{N}_p$ denote the p–dimensional Gaussian distribution. As in [4], we assume the following model

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

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

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

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

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

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

$$\{\mu_1,\ldots,\mu_m\} \mid \Lambda \sim \text{DPP}(\rho,\Lambda,K_0) \tag{5}$$
$$w_1,\ldots,w_m \mid m \sim \text{Dirichlet}(\alpha,\ldots,\alpha) \tag{6}$$
$$\Delta_1,\ldots,\Delta_m \mid m \overset{\text{iid}}{\sim} \text{inv-Wishart}(\nu_0,\Psi_0) \tag{7}$$
$$\sigma_1^2,\ldots,\sigma_p^2 \overset{\text{iid}}{\sim} \text{inv-Gamma}(a_\sigma,b_\sigma), \tag{8}$$

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 $i = 1,\ldots,50$.

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].

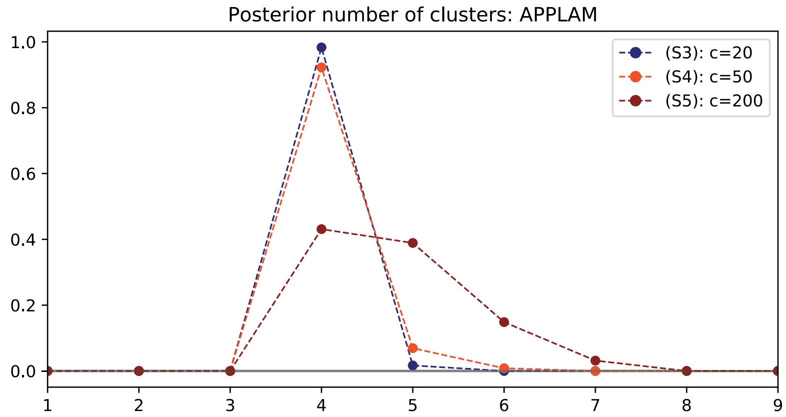


Fig. 1: Posterior distributions of the number of clusters

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 $V = \{1, \dots, q\}$ and set of edges $E \subseteq V \times V$. Let also $(X_1, \dots, X_q)$ 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,

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

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

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

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 $u = 1, \dots, q$. 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, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_q)^\top$ 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 $j = 1, \dots, q$, 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 $(X_1, \dots, X_q)$ can be equivalently written as

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

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 $(\boldsymbol{\mu},\boldsymbol{\Omega}) \sim \mathscr{N}\mathscr{W}(a_\mu,\boldsymbol{m},a_\Omega,\boldsymbol{U})$, 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

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

with I-Ga denoting an inverse-gamma distribution and $a_j^{\mathscr{D}} = a_\Omega + |\mathrm{pa}_{\mathscr{D}}(j)| - q + 1$. Most importantly, the resulting prior is conjugate to the sampling density in (4). Hence, for given n i.i.d. observations $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ from the Gaussian DAG model (1), the marginal likelihood $p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\,|\,\mathscr{D})$ 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 $\boldsymbol{x}_i = (x_{i,1}, \dots, x_{i,q})^\top$, $i = 1, \dots, n$, 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 $K \le n$ be the number of *unique* values among $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ and $\xi_1, \dots, \xi_n$, $\xi_i \in \{1, \dots, K\}$, 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,

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

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 $\boldsymbol{\theta}^{(k)} = \big(\boldsymbol{\eta}^{(k)}, \boldsymbol{D}^{(k)}, \boldsymbol{L}^{(k)}, \mathscr{D}^{(k)}\big)$, $k = 1, \dots, K$, 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 $\xi_1, \dots, \xi_n$ 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 $j = 1,\dots,q$, $l \in \text{pa}_{\mathscr{D}}(j)$, $k \in \{1,2\}$ 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.

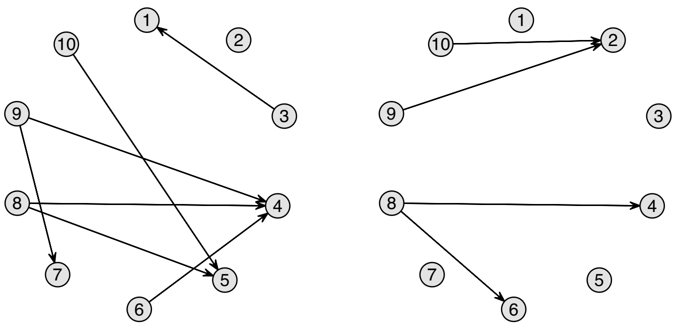


Fig. 1: Simulation study. True randomly generated DAG structures, $\mathscr{D}_1$, $\mathscr{D}_2$.

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 $t = 1,\dots,S$, our algorithm returns an n-dimensional vector of cluster indicators $\boldsymbol{\xi}^{(t)} = \left(\xi_1^{(t)},\dots,\xi_n^{(t)}\right)$, with $\xi_i^{(t)} \in \left\{1,\dots,K^{(t)}\right\}$ where $K^{(t)}$ is the number of distinct clusters, together with a collection of $K^{(t)}$ distinct cluster-specific parameters $\left\{\boldsymbol{\theta}_1^{(t)},\dots,\boldsymbol{\theta}_{K^{(t)}}^{(t)}\right\}$.

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 $i = 1,\dots,n$ 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

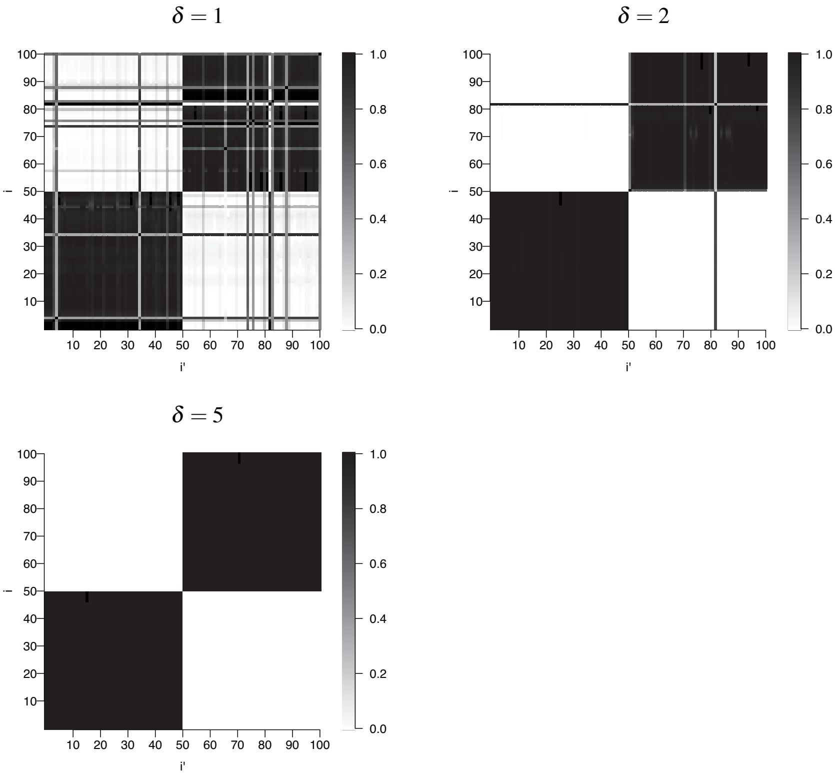


Fig. 2: Simulation study. Posterior similarity matrix with subjects arranged by true cluster membership (1:50 from cluster 1, 51:100 from cluster 2) for values of $\delta \in \{1,2,5\}$.

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

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.

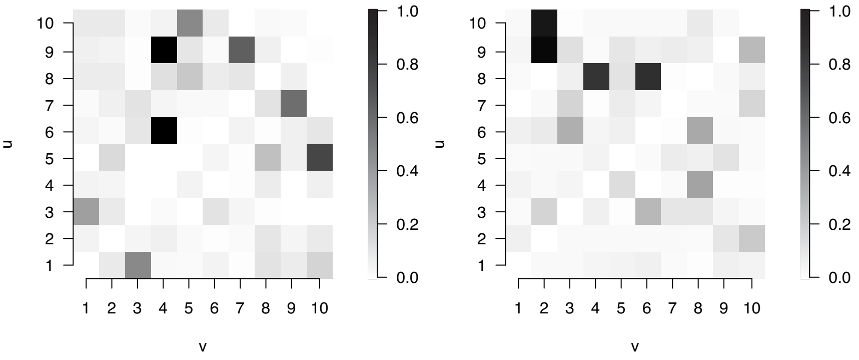


Fig. 3: Simulation study. Posterior probability of edge inclusion for each directed edge (u,v), for subjects $i \in \{10,60\}$ whose true cluster memberships are $\xi_{10} = 1$ and $\xi_{60} = 2$.

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).

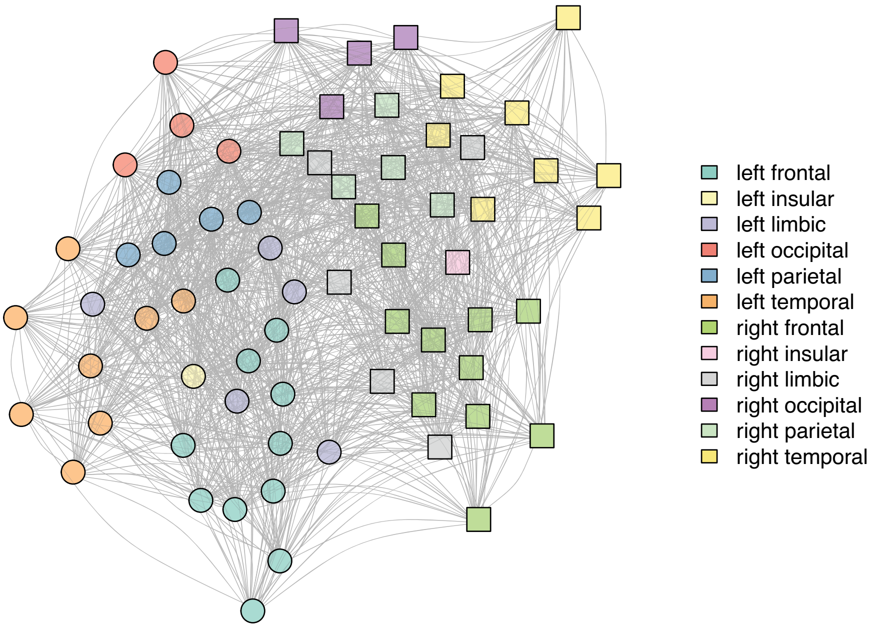


Fig. 1 Graphical representation of the considered network of brain regions. Node positions are obtained via force-directed placement (Fruchterman and Reingold, 1991), node shapes denote hemispheres (round=left, square=right), and colors correspond to the 12 hemisphere-lobe combinations.

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

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

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

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

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

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

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.

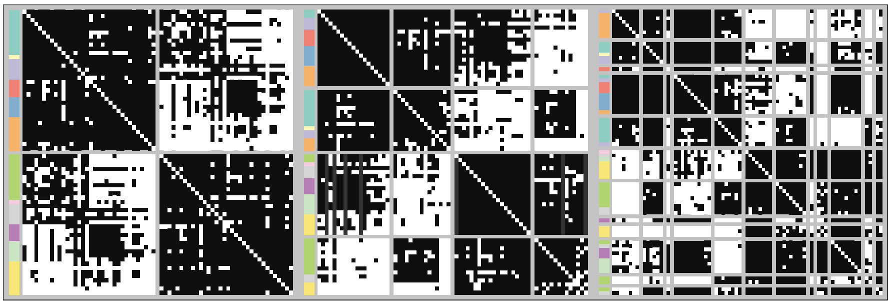
The availability of the urn schemes (2) and (4) allows to derive a collapsed Gibbs sampler. This actually holds for the whole class of Gibbs-type priors, and hence for any ESBM. See Legramanti et al. (2022b) for details on the Gibbs sampler, and https://github.com/danieledurante/ESBM for an R implementation.

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 $\mathbf{H}_j = diag(h_j^{(1)}, \dots, h_j^{(d)})$ 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

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

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

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

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

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

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.

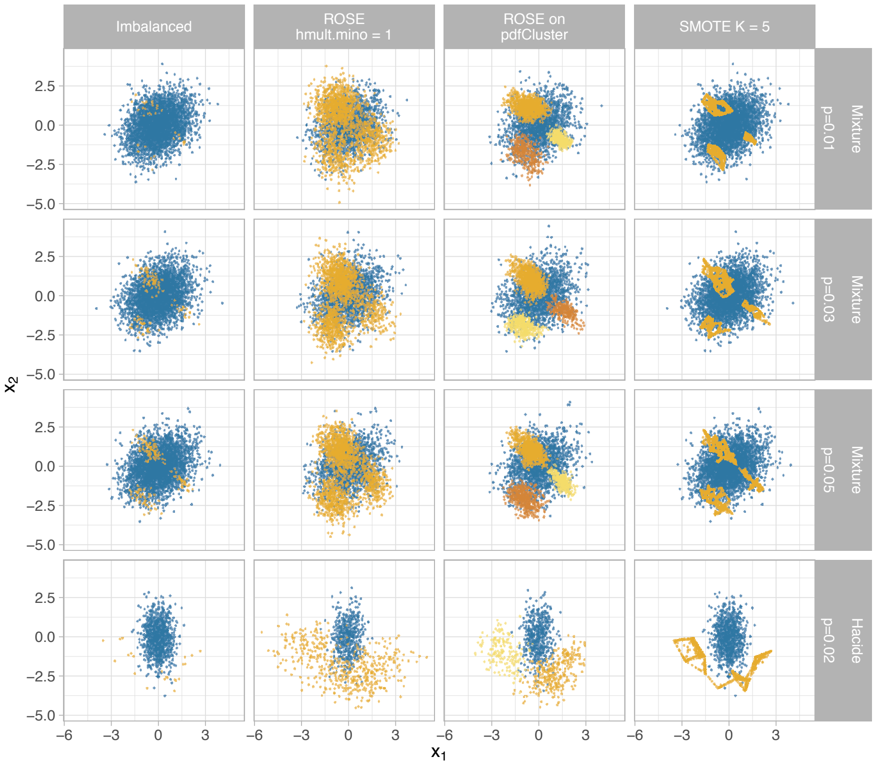


Fig. 1 Training sets with different proportion of the minority class (rows: $p = (0.01, 0.03, 0.05)$ for mixture data and $p = 0.02$ for hacide data). Each training set contains 3500 data points for the three mixture data examples and 1000 data points for the hacide data. The first column shows the imbalanced 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 approaches 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 $i = 1,\dots,N_p$ and $j = 1,\dots,M_p$. 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, $P(Y_{ij} = 1)$:

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

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 Garcìa, 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 $\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}^T = \boldsymbol{R}$ 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 ($\text{PM}_{2.5}$). Due to this, they conducted a study to determine if an increase in $\text{PM}_{2.5}$ resulted in a higher COVID-19 mortality rate. They found that an increase of

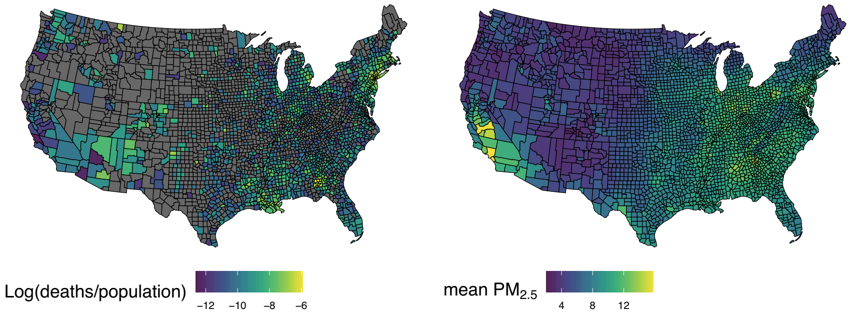


Fig. 1 $\mathbf{PM_{2.5}}$ exposure and COVID-19 mortality by US county: (left) the log COVID-19 mortality rate (i.e., log(deaths/population)) through May 12, 2020 (counties with no deaths are shaded gray) and (right) Average $\mathrm{PM}_{2.5}$ ($\mu g/m^3$) over 2000-2016.

1 $\mu g/m^3$ in ambient fine particulate matter ($\mathrm{PM}_{2.5}$) 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 $\mathrm{PM}_{2.5}$ was calculated by averaging results from an established exposure prediction model for years 2000-2016. This resulted in mortality counts and $\mathrm{PM}_{2.5}$ measures for $n = 3109$ counties, see Fig. 1). The long-term average $\mathrm{PM}_{2.5}$ 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 $\mathrm{PM}_{2.5}$ rate is a smoother spatial process than mortality, likely because the $\mathrm{PM}_{2.5}$ exposure estimates are generated from predictive models. In addition to $\mathrm{PM}_{2.5}$ 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 $\mathrm{PM}_{2.5}$ 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 $\mathrm{E}(Y_i \mid X_i, Z_i, \boldsymbol{C}_i) = \lambda_i = r_i(1-p_i)/p_i$. 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

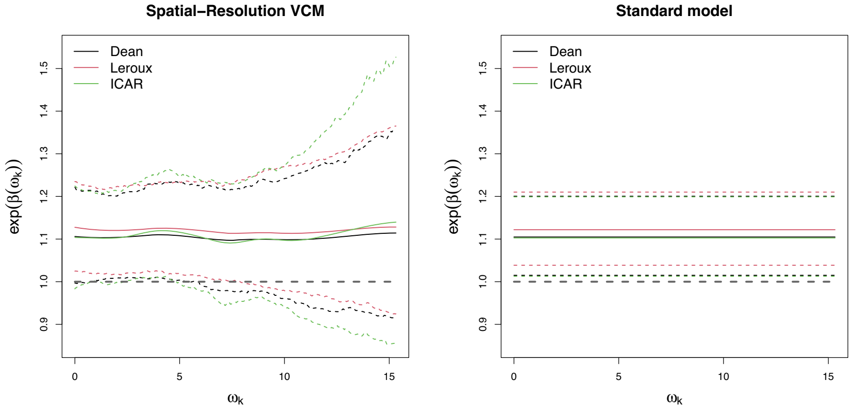


Fig. 2 The left plot contains estimated coefficient values as a function of spatial resolution based on the spatial-resolution varying coefficient model for the Dean, Leroux, and ICAR areal data models. The right plot is the standard spatial generalized linear mixed model approach.

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 $\text{PM}_{2.5}$, 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 ($p = 1, \dots, P$) and t^{th} time point ($t = 1, \dots, T$), 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 **s** are the mixture weights, representing the probability of an observation coming from source $1, \dots, k, \dots, K \longrightarrow \infty$. 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.

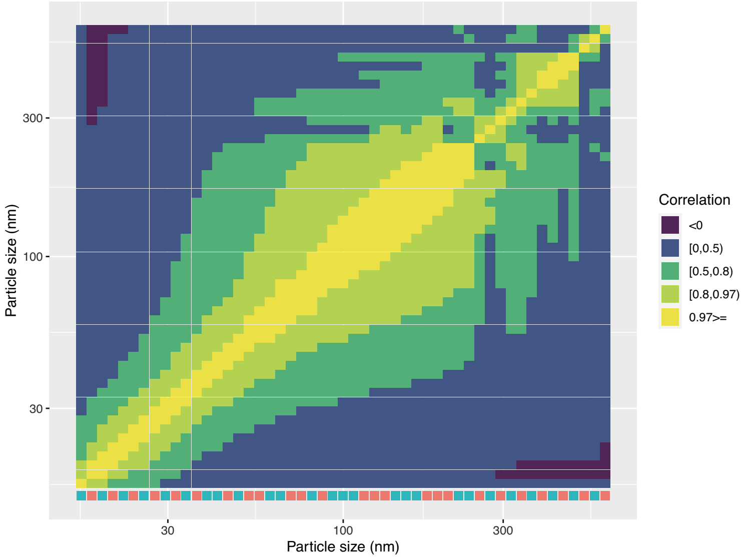


Fig. 1 Correlation matrix between data in original size bins. Resulting aggregated bins are indicated at the bottom by colored segments (preliminary results).

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.

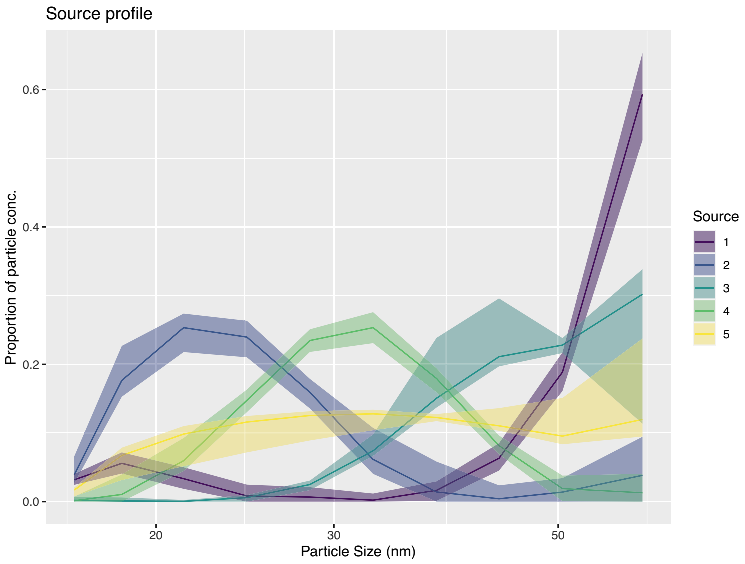


Fig. 2 Particle size distribution for the five sources identified by the model for summer period (preliminary results).

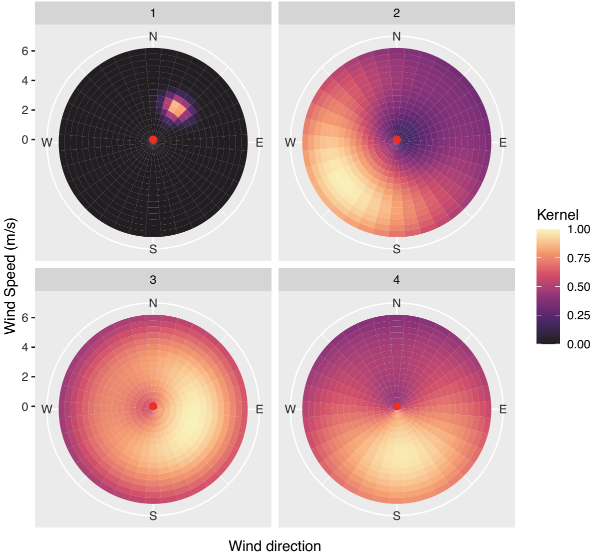


Fig. 3 Wind kernels for $K - 1$ sources (preliminary results).

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 $\rho(|\mathbf{s} - \mathbf{s}\prime|; \phi)$. Thus, $\mathbf{X}$ and $\mathbf{g}$ are jointly normal with the following:

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

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].

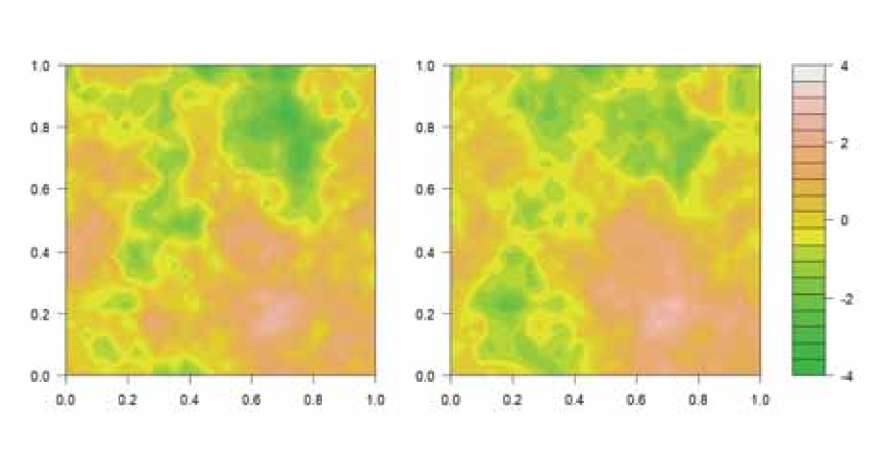


Fig. 1 Exposure (left) and spatial error process $g(\mathbf{s}_i)$ (right) are highly correlated ($\delta = 0.7$). Variables are scaled to have zero mean and unit variance.

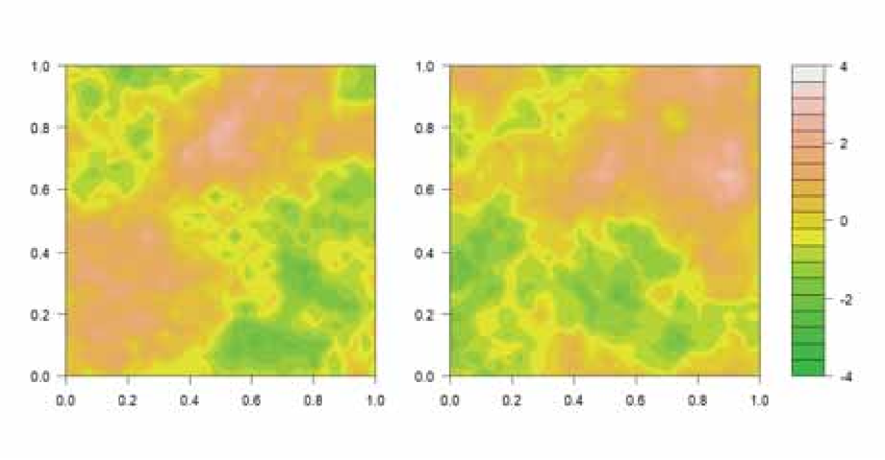


Fig. 2 Exposure (left) and spatial error process $g(\mathbf{s}_i)$ (right) are uncorrelated ($\delta = 0$). Variables are scaled to have zero mean and unit variance.

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].

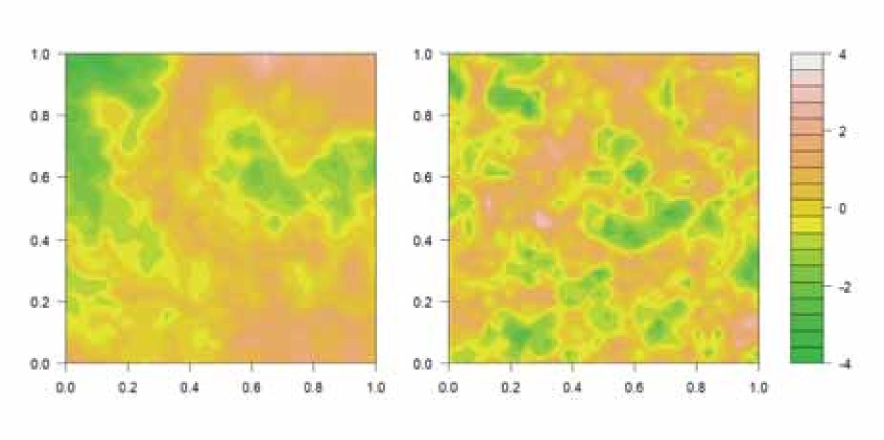


Fig. 3 The exposure (left) is generated using a large spatial range ($\phi_x = 0.5$), while the spatial error process $g(\mathbf{s}_i)$ (right) using a smaller one ($\phi_g = 0.05$). Variables are scaled to have zero mean and unit variance.

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-

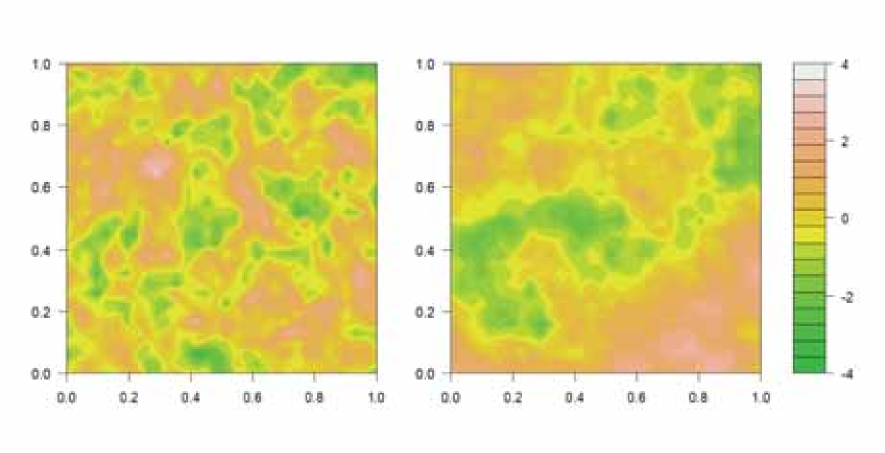


Fig. 4 The exposure (left) is generated using a small spatial range ($\phi_x = 0.05$), while the spatial error process $g(\mathbf{s}_i)$ (right) using a larger one ($\phi_g = 0.5$). Variables are scaled to have zero mean and unit variance.

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 **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 $\mathbf{X}_{t-q} = \mathrm{diag}(\mathbf{x}_{t-q}) \in \mathbb{R}^{K\times K}$ be the matricized version of $\mathbf{x}_{t-q}$, then it follows:

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

where $\tilde{\mathbf{x}}_{t-q} = \mathrm{vec}(\mathbf{X}_{t-q})$. 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

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

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

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 $\pi(\gamma|\Sigma_1, \Sigma_2) \sim \mathscr{G}a\Big(\underline{a} + \frac{1}{2}\big[\underline{\nu}_1 G + \underline{\nu}_2 N\big], \underline{b} + \frac{1}{2}\big[\mathrm{tr}(\underline{S}_1 \Sigma_1^{-1}) + \mathrm{tr}(\underline{S}_2 \Sigma_2^{-1})\big]\Big)$

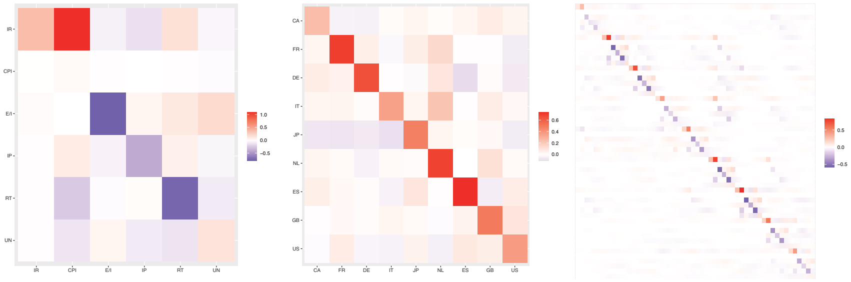
where $\bar{\Omega}_i = \big[\underline{\Omega}_i^{-1} + (\mathscr{X}_i \mathscr{X}_i)' \otimes \Sigma_i^{-1}\big]^{-1}$, $\bar{\beta}_i = \bar{\Omega}_i\big[\underline{\Omega}_i^{-1}\underline{\beta}_i + (\mathscr{X}_i \otimes \Sigma_i^{-1})\mathrm{vec}(\mathscr{Y}_i)\big]$ and $\hat{S}_i = (\mathscr{Y}_i - \hat{\mathscr{B}}_i \mathscr{X}_i)(\mathscr{Y}_i - \hat{\mathscr{B}}_i \mathscr{X}_i)'$ 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.

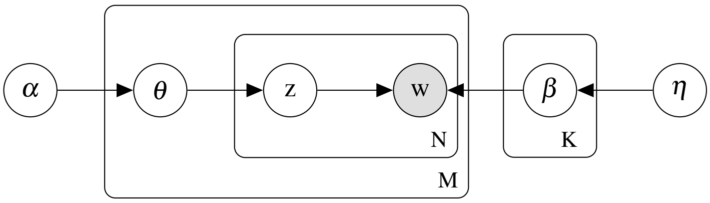


Fig. 1 Graphical model for LDA.

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

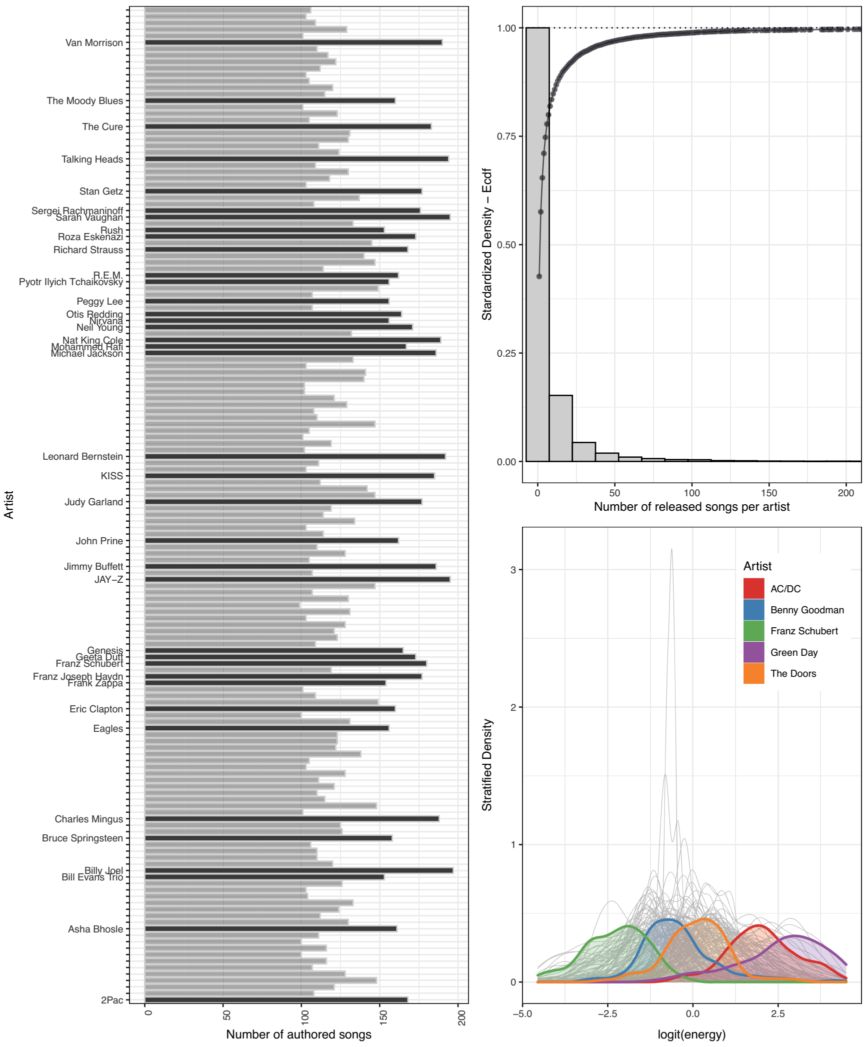


Fig. 1 Descriptive plots of the `Spotify` dataset. Left panel: barplots displaying the number of authored songs for each selected artist. Top right: histogram and empirical c.d.f. for the number of songs per artist (entire dataset). Bottom-right: density plots of the energy distributions stratified by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

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

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

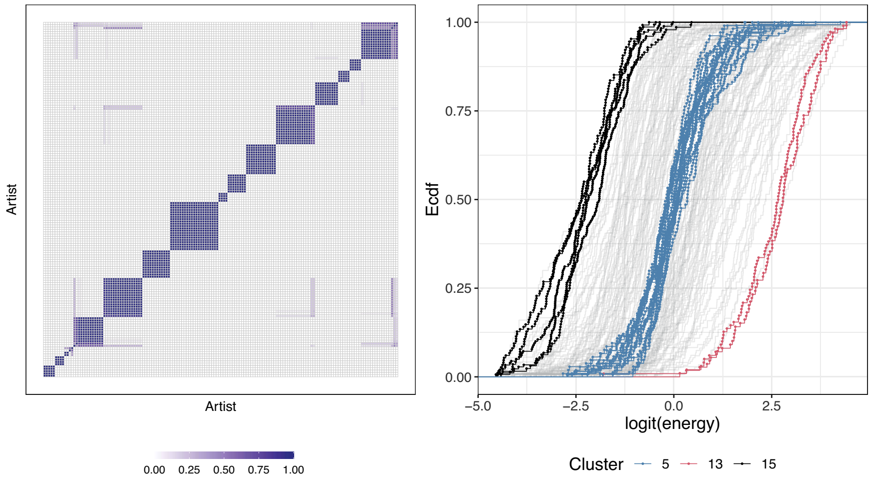


Fig. 2 Left panel: posterior co-clustering matrix between artists. Right panel: empirical cdfs for every logit-energy distribution considered. Three distributional clusters are highlighted.

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.

$$\begin{aligned}&f\left(\mathscr{X}_{it}=\mathbf{X}_{it}|\mathscr{X}_{i1}=\mathbf{X}_{i1},\ldots,\mathscr{X}_{it-1}=\mathbf{X}_{it-1},S_{i1}=s_{i1}\ldots,S_{it}=s_{it}\right)\\&=f\left(\mathscr{X}_{it}=\mathbf{X}_{it}|S_{it}=s_{it}\right),\end{aligned}$$

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

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

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 $\pi_{ik}=\Pr\left(S_{i1}=k\right)$, $k=1,\ldots,K$, and the transition probabilities

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

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 $14\times7=98$ 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 algorithm 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 simulated 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

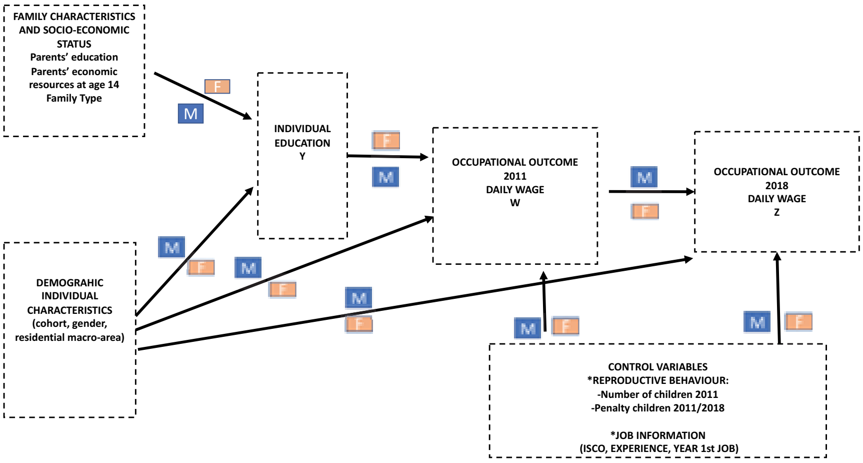
We use the AD-SILC[1] database, which is constructed by matching longitudinal information from administrative archives held by the National Institute of Social Security (INPS) with survey data collected by the National Institute of Statistics (ISTAT). In our database, we have information on the socio-economic 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.

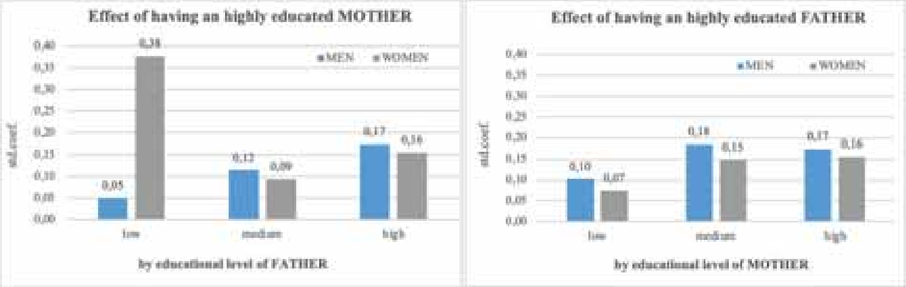
We assess the presence of gender bias in the effect of socio-economic status on both indicators of occupational outcomes by estimating two separate models for men and women. The comparison between the direct and indirect effects of family socio-economic conditions for the two genders highlights gender 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).

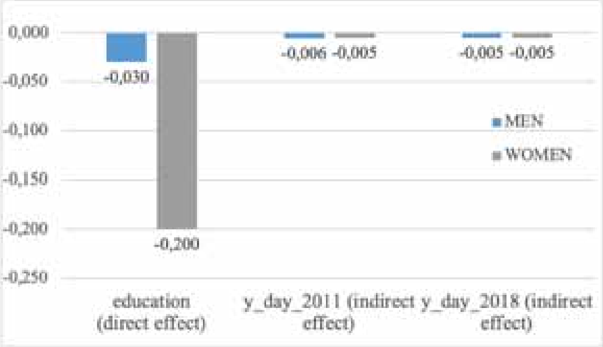
When other factors are equal, family disruption seems to have a small but significant direct effect on men's educational levels (see Figure 3).

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*)*



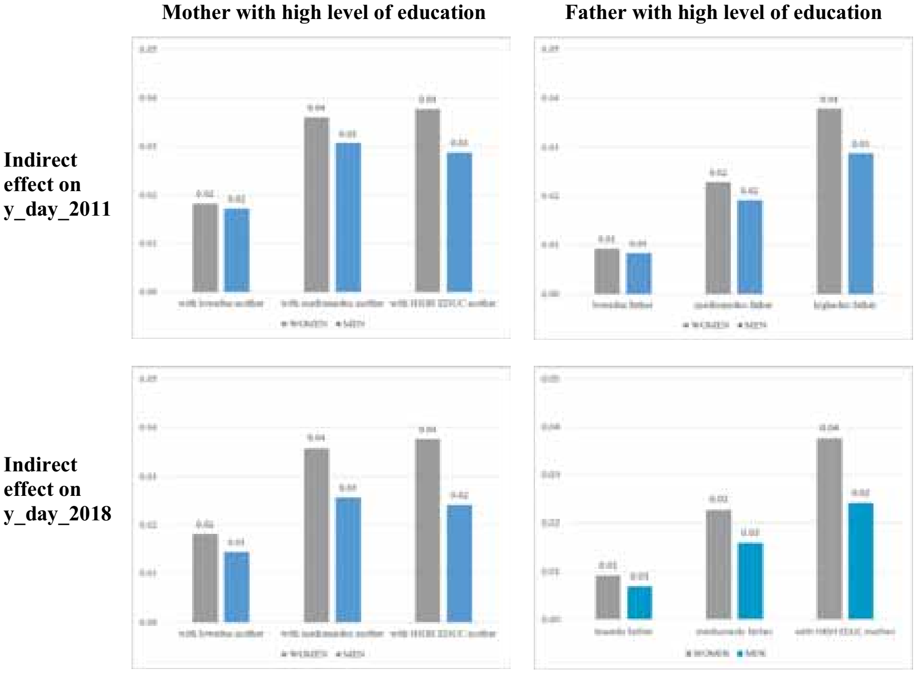
* The men coefficient is significant, with *p* value < 0.05. The women coefficients are not significant.
Source: Authors' elaborations on AD-SILC data

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*



Mother with high level of education **Father with high level of education**

Indirect effect on y_day_2011

Indirect effect on y_day_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

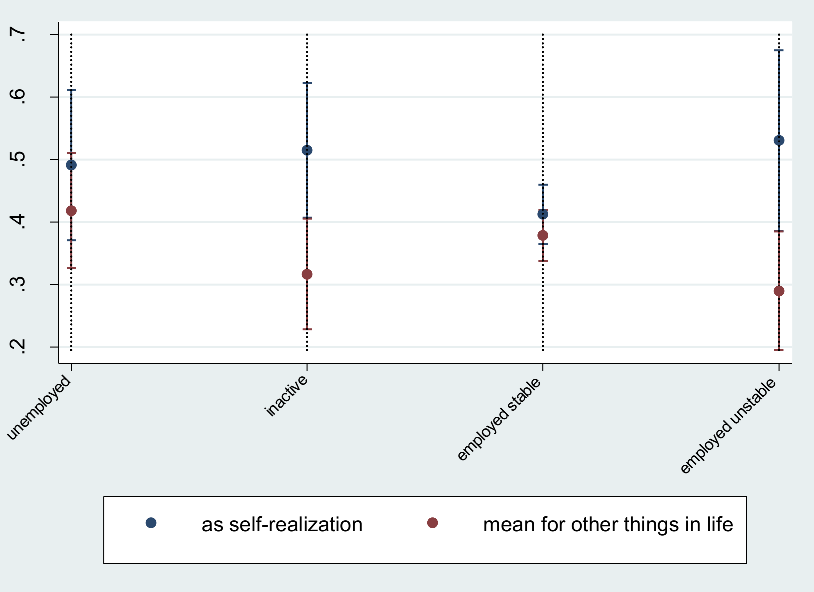
Preliminary results from random effects logit models show that the probability to reduce - over the survey period - the motivation towards having children (and in particular for those declaring they would feel fully realized in their life without children) is associated with holding a tertiary degree 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

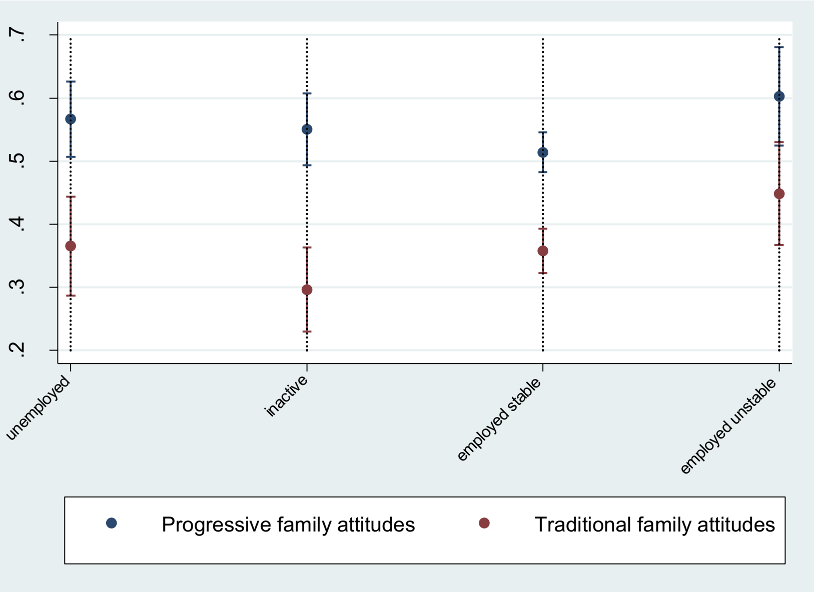
However, when we consider the mediation role of the cultural dimension things slightly change. In particular, for those considering their work as an important dimension of self-realization, becoming inactive or holding an unstable job contract increases the probability of reducing the relevance of having 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.

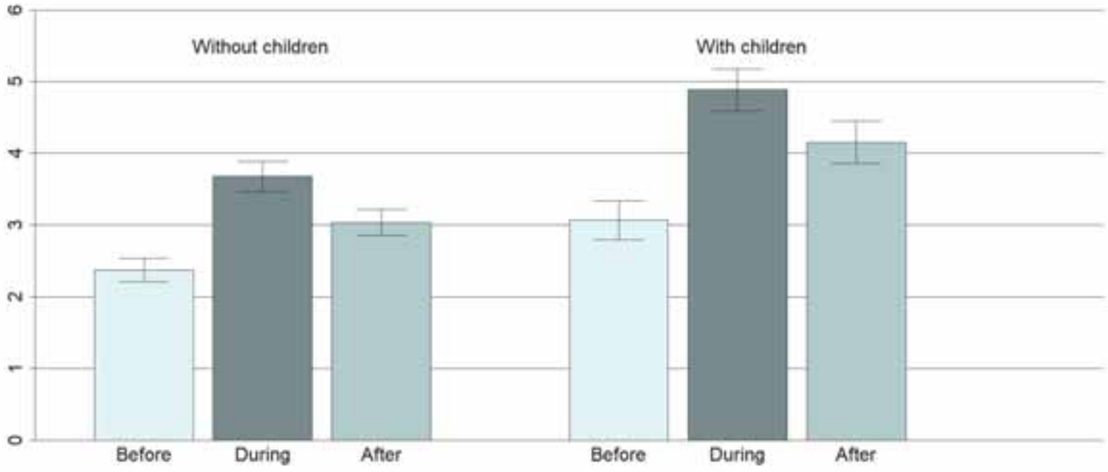


Figure 1: Women's *average daily hours of unpaid domestic work.*

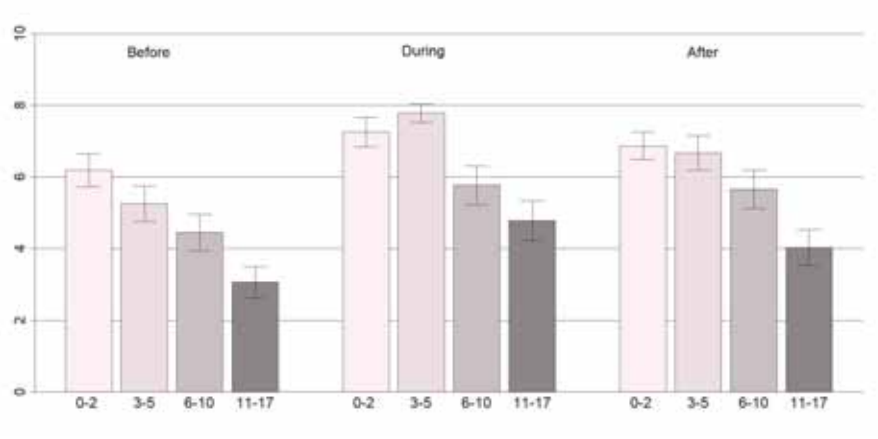


Figure 2: *Mothers' average daily hours of childcare according to the age of the youngest child.*

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).

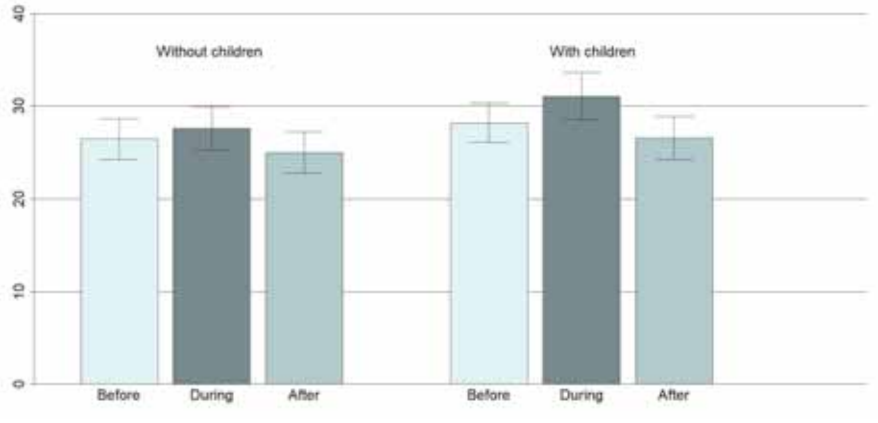


Figure 3: *Partner's average share of unpaid care and domestic work.*

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.

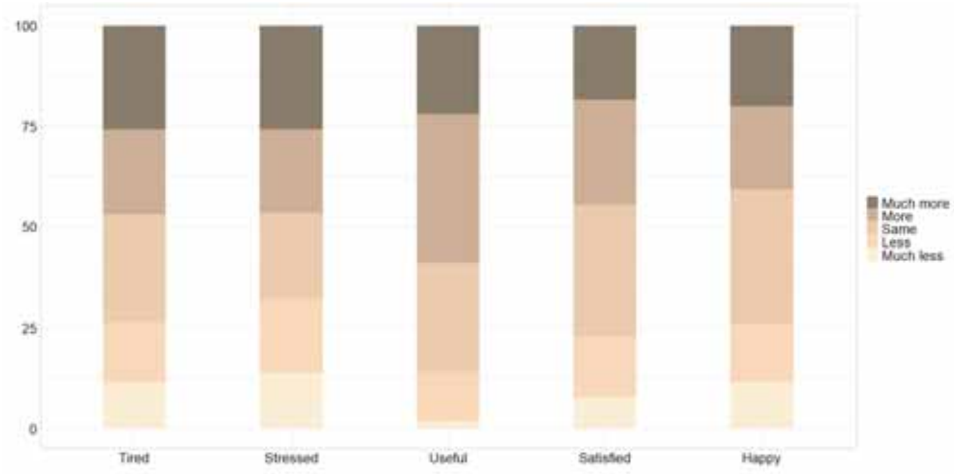


Figure 4: *Did you feel more or less ... than usual while spending time on childcare during the lockdown?*

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.

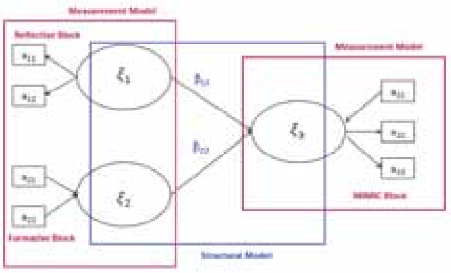


Fig. 1 PLS-PM: structural model and measurement model

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 primarily 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:

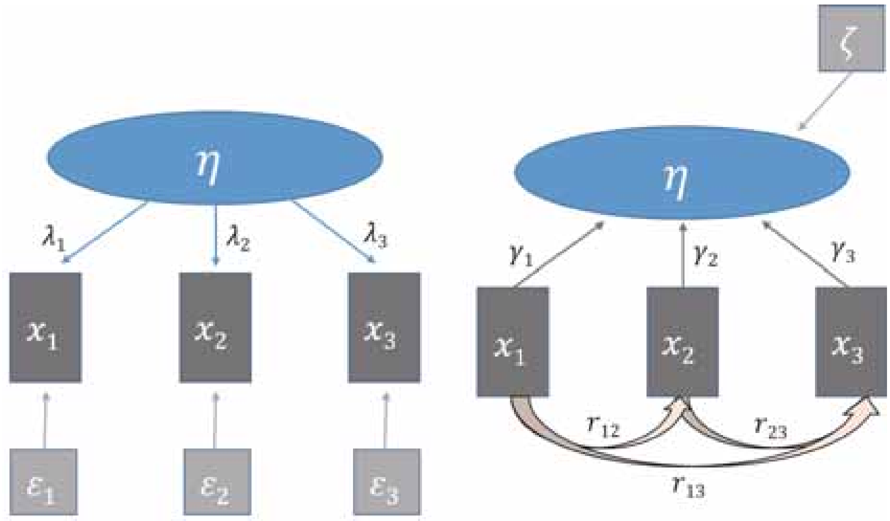
$$\mathbf{V} \equiv \left\{v_{it} : i = 1,\ldots,N; t = 1,\ldots,T\right\} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1T} \\ v_{21} & v_{22} & \cdots & v_{2T} \\ \vdots & \ddots & \ddots & \vdots \\ v_{N1} & v_{N2} & \cdots & v_{NT} \end{pmatrix} \tag{2}$$

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

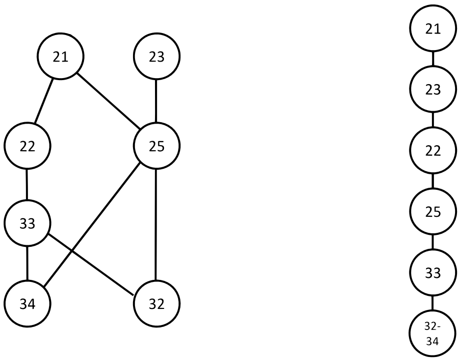


Fig. 1 Left panel: Hasse diagram of the selected Belgium regions; Right panel: final ranking

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 empirical 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:

$$\widehat{t}_{rf} = \sum_{k\in U} \widehat{m}_{rf}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S} \frac{\left(1-\psi_k^{(b,S)}\right)\left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})\right)}{\pi_k}. \qquad (13)$$

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 $P_{k,1}, \dots, P_{k,n_k}$. Following Cordy (1993), the designs should be such that the n_k-tuple $[P_{k,1}, \dots, P_{k,n_k}]$ 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 $i \neq h = 1, \dots, n_k$. 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 $\widehat{T}_k/T_k$ 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

$$\widehat{T} = \sum_{j \in S(p)} \frac{y_j}{\pi_j} = \lambda(B) \sum_{j \in S(p)} \frac{y_j}{\lambda(B_j)} \tag{9}$$

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

$$\widehat{T} = \lambda(B) g(p) \tag{11}$$

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 $\widehat{T}_1, \ldots, \widehat{T}_K$ are available for the totals $T_1, \ldots, T_K$ of K attributes, then for many functions of totals f, the plug-in estimator $f(\widehat{T}_1, \ldots, \widehat{T}_K)$ 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,

$$|B(\,\bar{y})| \leq \frac{S_\rho S_y}{\bar{\rho}} = CV_\rho\, S_y, \tag{1}$$

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

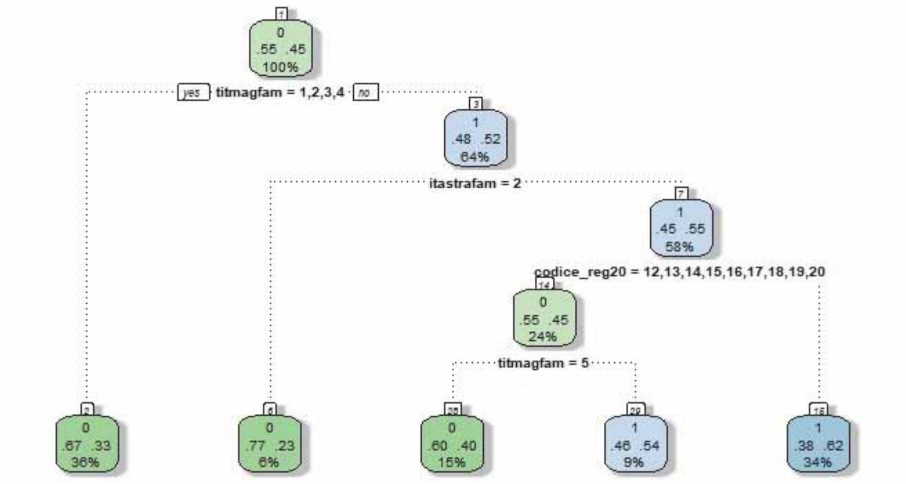
The approach proposed by van Berkel et al. (2020) is, here, implemented on the list component of the 2018 PCPH data enriched by linking additional variables from administrative sources.

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_ρ 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

$$\min_{p_g}\left(\sqrt{\left[\sum_{g=1}^{G} p_g^2\, F_g - \left[\sum_{g=1}^{G} p_g F_g\right]^2\right]} \Bigg/ \sum_{g=1}^{G} p_g F_g\right) = CV_\rho$$

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

$$n_{cawi} \sum_{g=1}^{G} F_g \left(1 - p_{cawi,g}\right) f_{capi,g} p_{capi,g} = C$$

$$n_{cawi} \sum_{g=1}^{G} \left[F_g \, p_{cawi,g} + F_g \left(1 - p_{cawi,g}\right) f_{capi,g} p_{capi,g}\right] \geq R$$

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, $cv(\hat{\theta}) = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}/\bar{\hat{\theta}}$, 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.

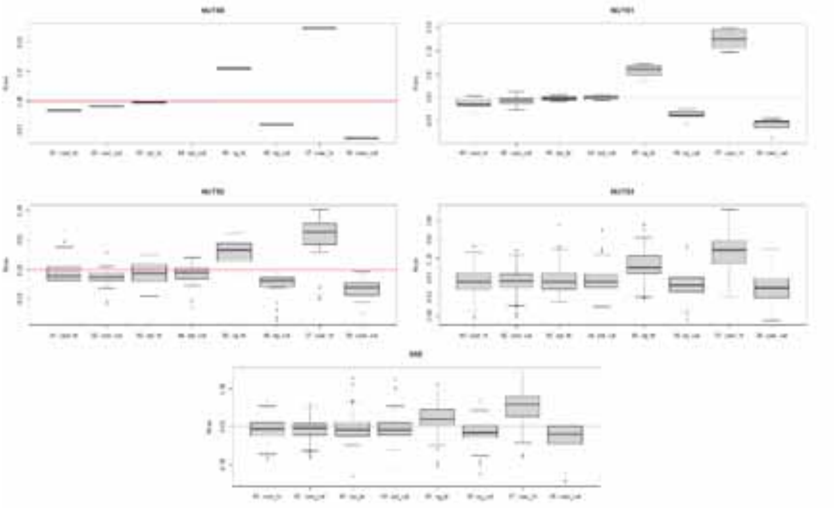
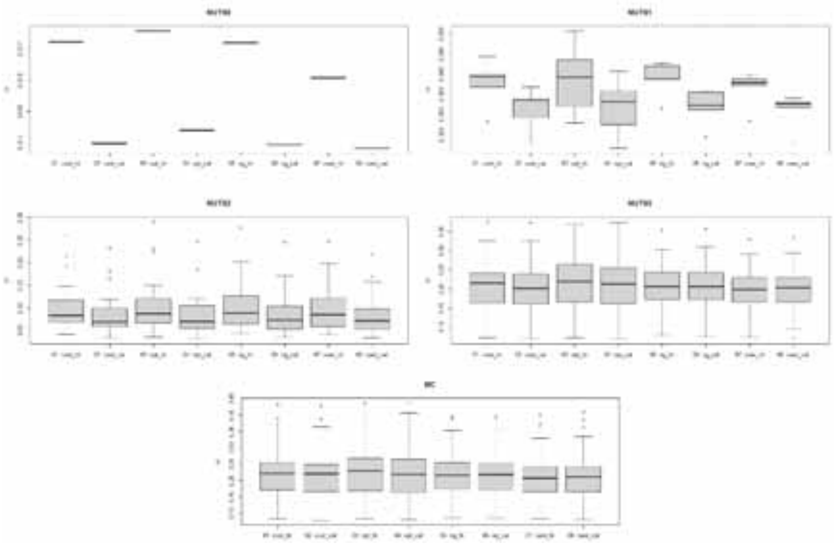


Figure 3: Normalized normalized root-mean-squared error (***NRMSE***) for the estimate of the unemployment rate at NUTS0, NUTS1, NUTS2, NUTS3 and MC level, under the different scenarios and with different estimators.



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

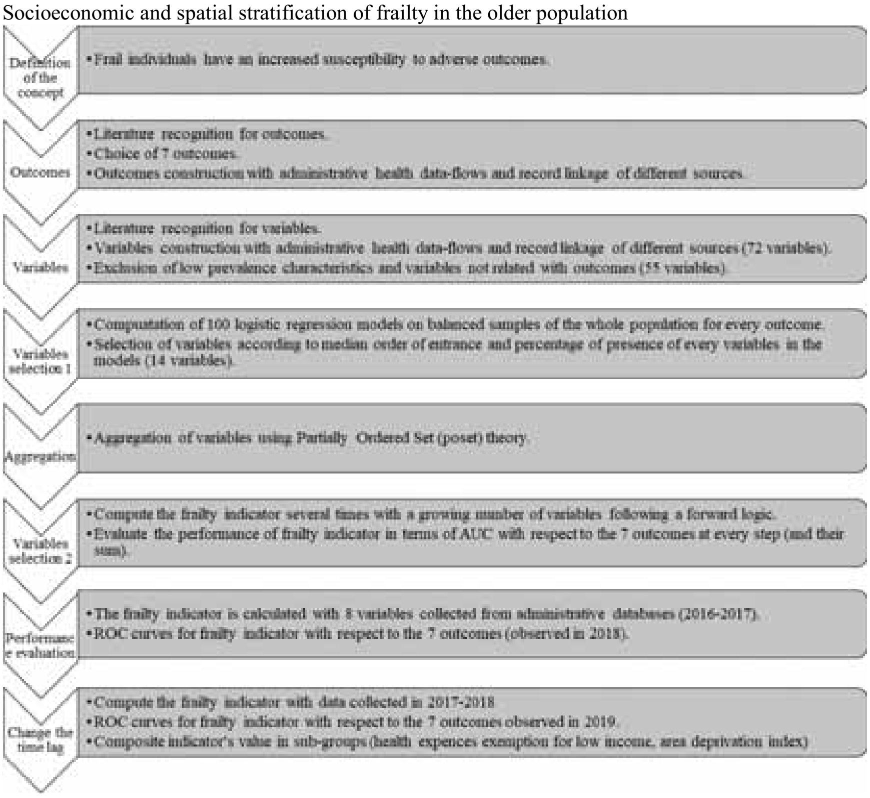
Methodological steps to build the frailty indicator start from a definition of frailty that finds wide support among scholars and defines frail individuals as individuals with increased susceptibility to adverse health outcomes [9] (Figure 1).

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.



- Frail individuals have an increased susceptibility to adverse outcomes.

- Literature recognition for outcomes.
- Choice of 7 outcomes.
- Outcomes construction with administrative health data-flows and record linkage of different sources.

- Literature recognition for variables.
- Variables construction with administrative health data-flows and record linkage of different sources (72 variables).
- Exclusion of low prevalence characteristics and variables not related with outcomes (55 variables).

- Computation of 100 logistic regression models on balanced samples of the whole population for every outcome.
- Selection of variables according to median order of entrance and percentage of presence of every variables in the models (14 variables).

- Aggregation of variables using Partially Ordered Set (poset) theory.

- Compute the frailty indicator several times with a growing number of variables following a forward logic.
- Evaluate the performance of frailty indicator in terms of AUC with respect to the 7 outcomes at every step (and their sum).

- The frailty indicator is calculated with 8 variables collected from administrative databases (2016-2017).
- ROC curves for frailty indicator with respect to the 7 outcomes (observed in 2018).

- Compute the frailty indicator with data collected in 2017-2018
- ROC curves for frailty indicator with respect to the 7 outcomes observed in 2019.
- Composite indicator's value in sub-groups (health expences exemption for low income, area deprivation index)

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.

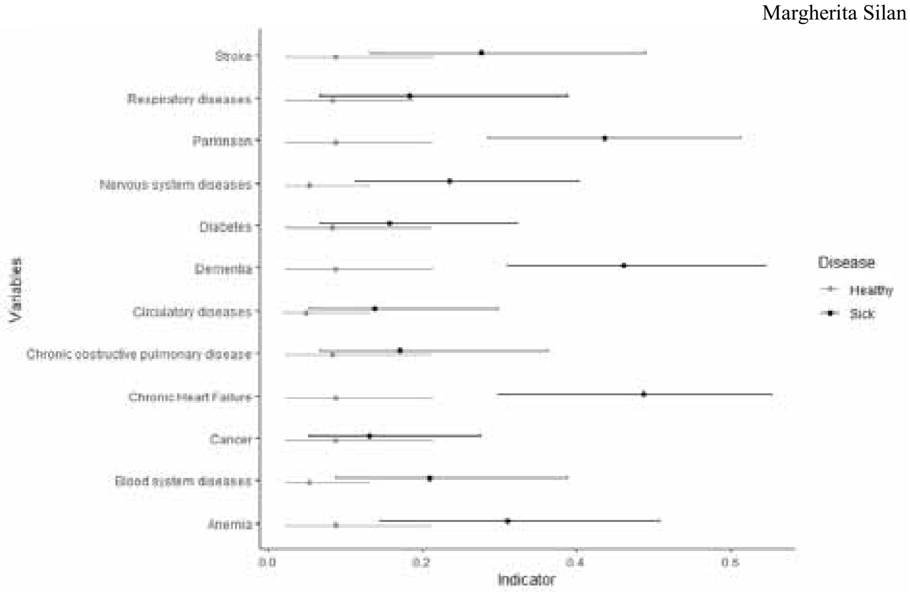


Figure 2: Frailty indicator by chronic conditions. Median and interquartile range.

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).

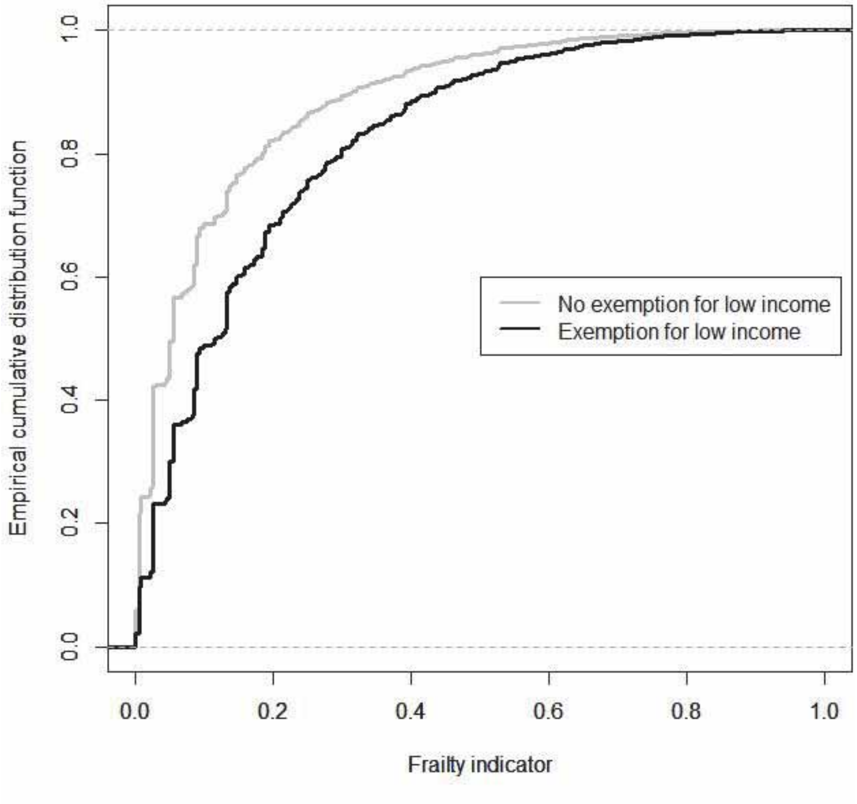


Figure 3: Frailty indicator by health expenses exemption for low income.

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.

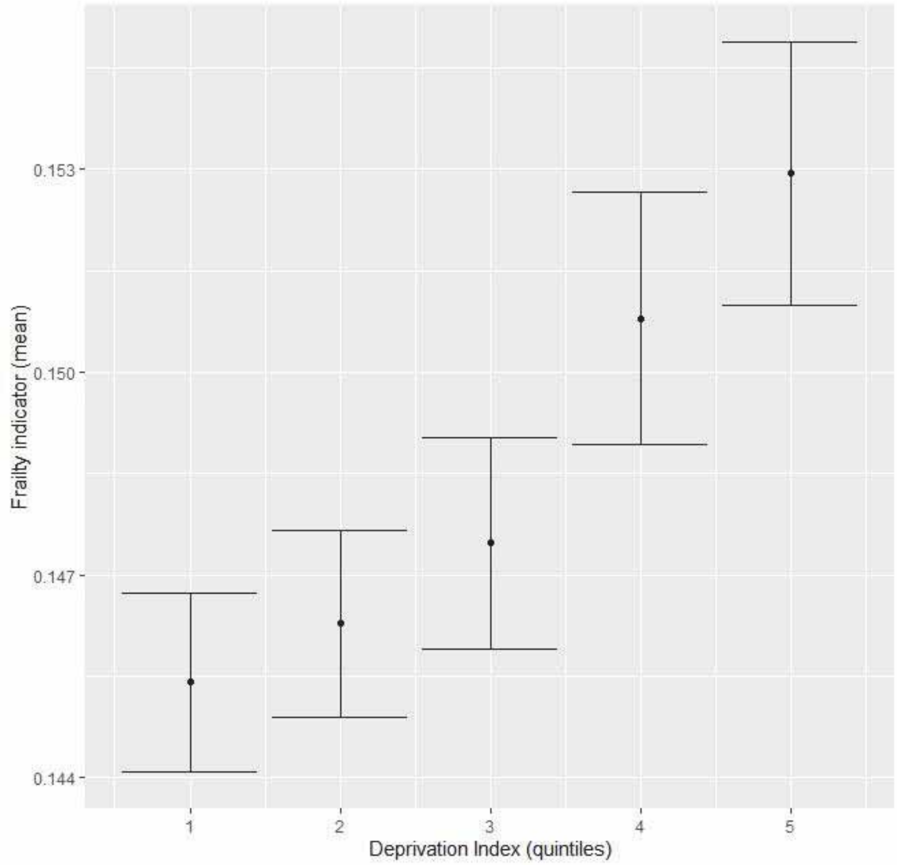


Figure 3: Frailty indicator (means and 95% confidence intervals) by Caranci deprivation index (based on 2011 Census).

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_pubblicazi oni_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.

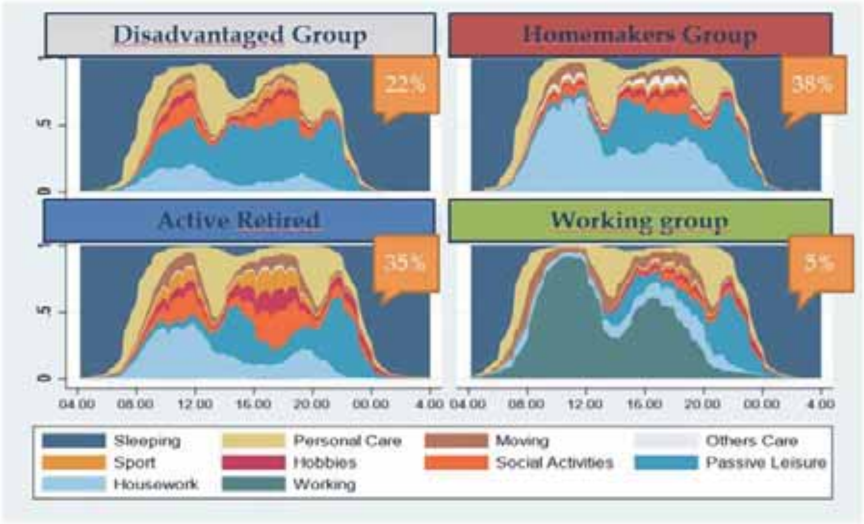


Figure 1: *Elderlies' time use profiles along the 24 hours. Chronograms.*
In the y axis the proportion of old people performing a certain activity in the time slot.
In the orange square the proportion of those belonging to each identified group.
Source: Authors' own elaboration on ISTAT Time use survey (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.

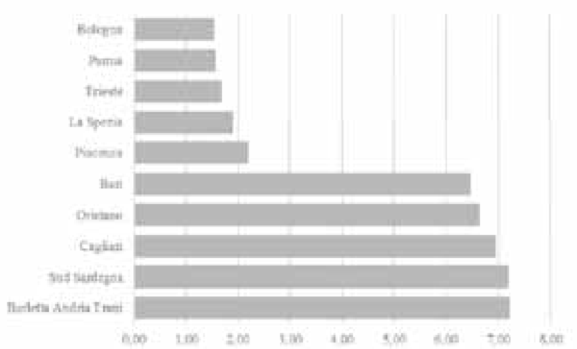


Figure 1: *Relative variations in the mean age of resident population between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

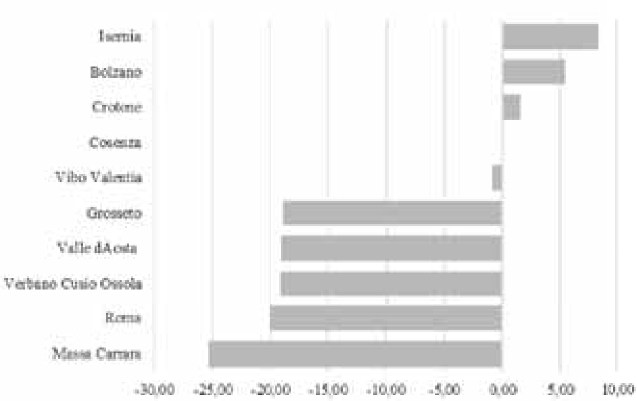


Figure 2: *Relative variations in the Total Fertility Rate of Italian women between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

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).

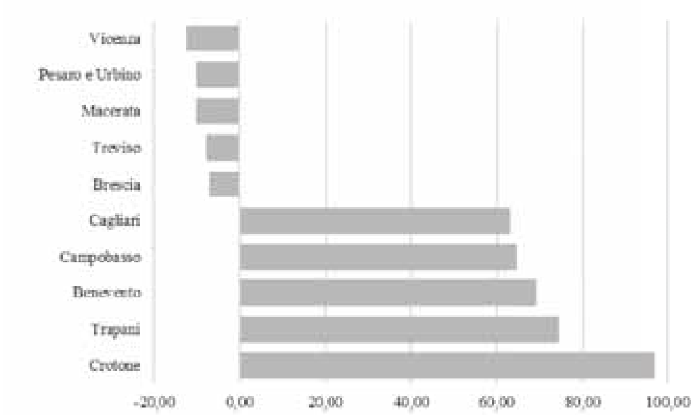


Figure 3: *Relative variations in the share of foreigners among total resident population between 2011 and 2019 by provinces, lowest 5 and highest 5 values.*

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.

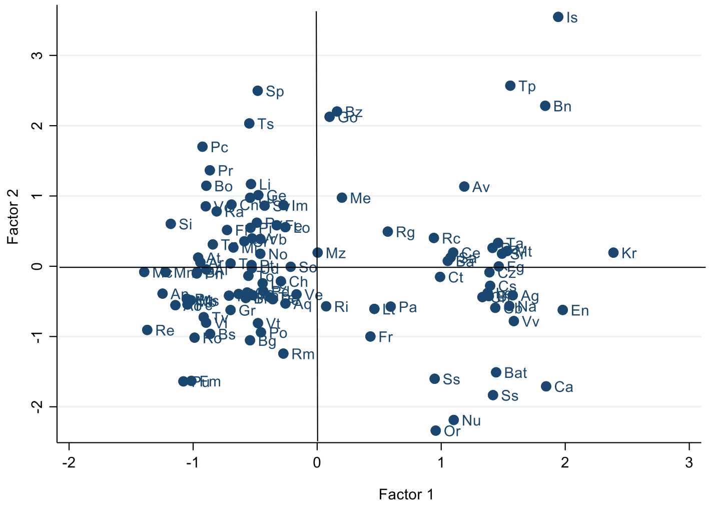


Figure 4: *Positions of provinces on the factor plane (first and second factor) resulting from principal components factor analysis.*

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 implementation 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 **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 **X** takes place in two steps. The observed values for all figures in a given year t are selected from the multivariate time array **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]:

$$\widetilde{\mathbf{X}}_i \equiv \{\mathbf{x}_{it} : t = 1, \ldots, T\}. \tag{3}$$

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%.

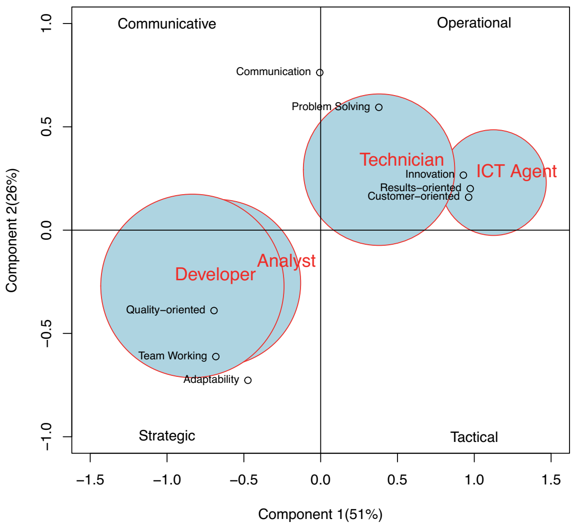
Table 4 Contribution of soft skills to 2 latent factors

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.

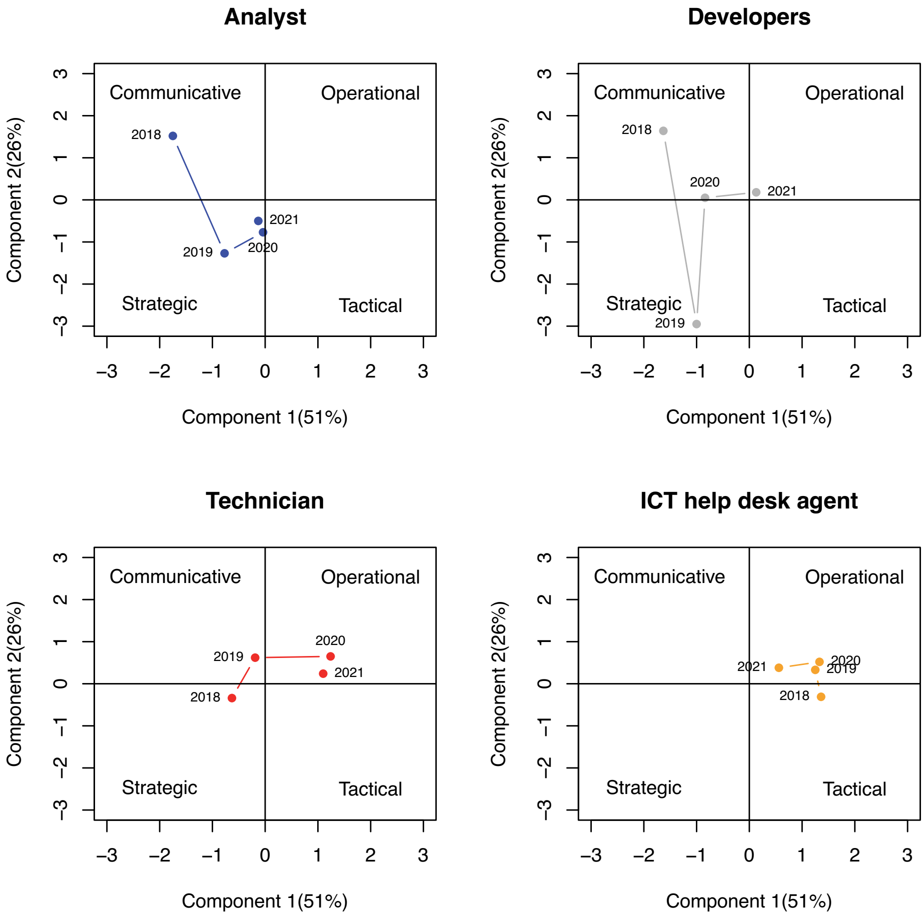
Fig. 1 Bubble graph of the WFA for job titles and soft skills



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

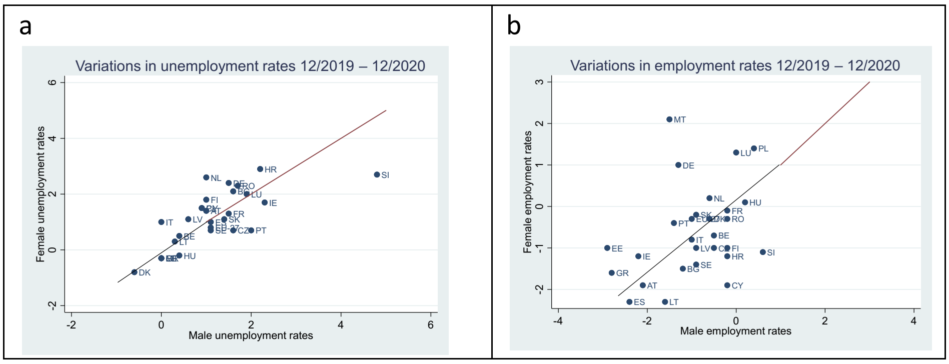
[1] The term defamilisiation has been recently introduced in the economic literature on the gender gap to indicate how the welfare state facilitates female autonomy and economic independence from the family [5].

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.



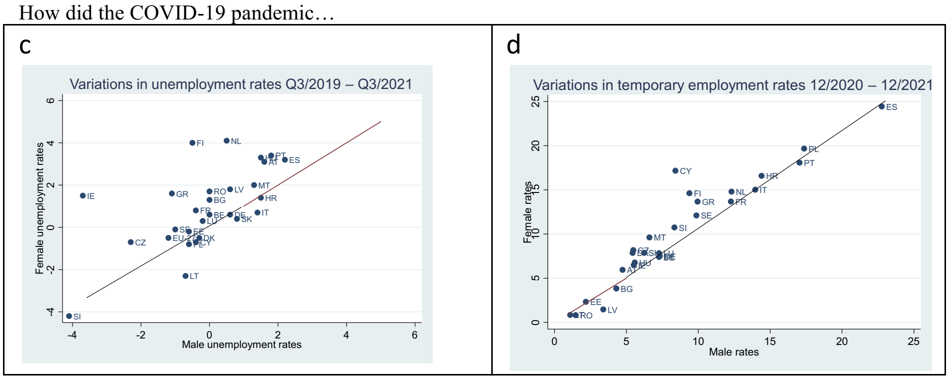


Figure 1: Variations in the employment and unemployment rates by gender and levels of temporary work in EU countries.

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]:

$$\overline{Y}^M - \overline{Y}^F = \beta^M\left(\overline{X}^M - \overline{X}^F\right) + \left(\beta^M - \beta^F\right)\overline{X}^F \qquad (1)$$

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:

$$\text{ATE}= E(Y_{1i}–Y_{0i}) \equiv E(\beta_i) \qquad (2)$$

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)

$$\text{ATET} =E(Y_{1i}–Y_{0i}|t=1) \equiv E(\beta_i|t=1) \qquad (3)$$

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	-	-

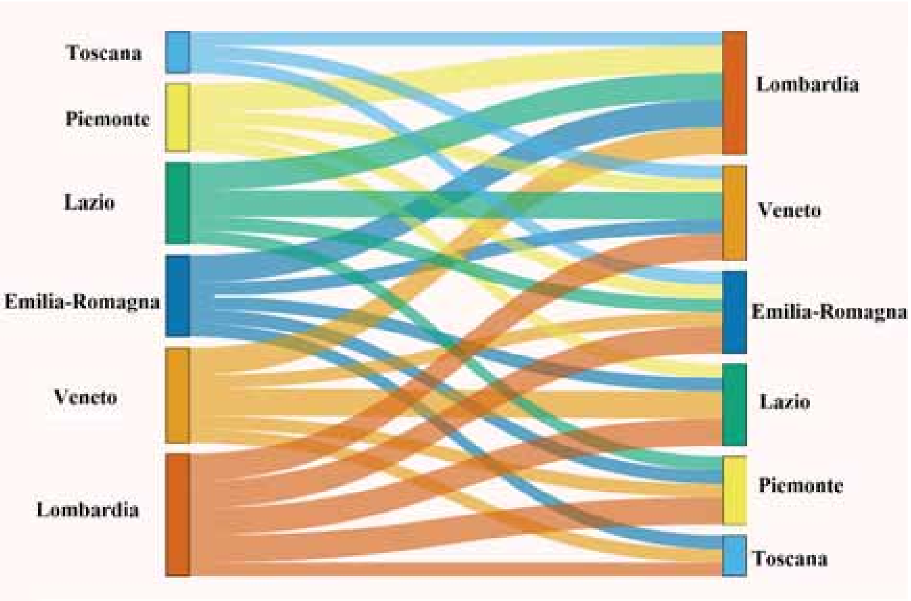


Figure 1: Region similarity ≥ 0.5

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).

$$P_{jk}^{L} = \frac{\sum_{i \in N_{jk}} p_{ik}\, q_{ij}}{\sum_{i \in N_{jk}} p_{ij}\, q_{ij}} \qquad\qquad P_{jk}^{P} = \frac{\sum_{i \in N_{jk}} p_{ik}\, q_{ik}}{\sum_{i \in N_{jk}} p_{ij}\, q_{ik}}$$

$$P_{jk}^{F} = \sqrt{P_{jk}^{L} \times P_{jk}^{P}} \quad (1)$$

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

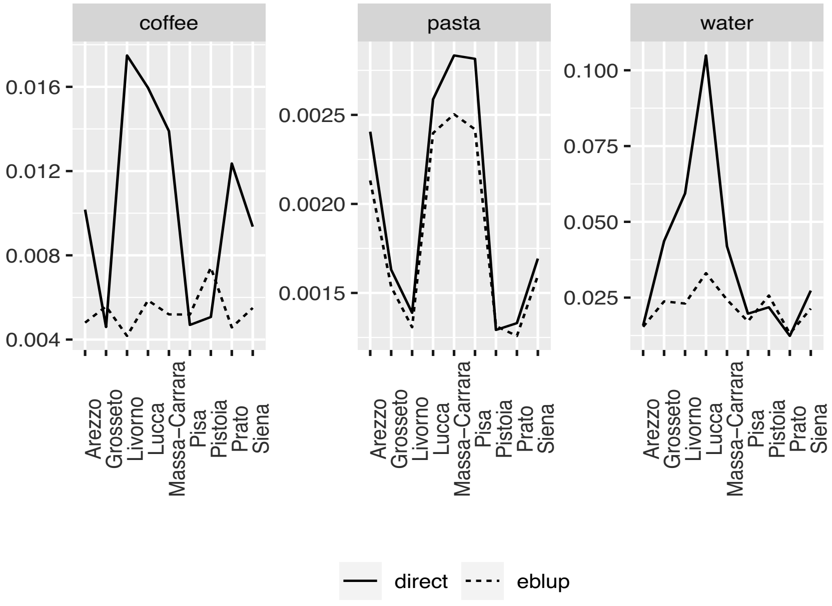


Figure 1: Gain in efficiency (Florence =1 as reference)

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:

$$\begin{cases} y_d | \lambda_{1d}, \lambda_{2d}, \phi_d, p \overset{ind}{\sim} FB(\lambda_{1d}, \lambda_{2d}, \phi_d, p) \quad \forall d = 1, \ldots, m \\ \text{logit}(\lambda_{2d}) = x_d^T \beta + v_d \quad v_d \sim N(0, \sigma_v^2) \end{cases} \tag{2}$$

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 practical 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 the 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 sullla metodologia integrazione dii 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, $i = 1,\dots,m$. 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 $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\dots,m$.

By using the non-probability sample we can estimate $\bar{Y}_i$ by:

$$\bar{Y}_{B_i} = N_{B_i}^{-1} \sum_{j \in U_i} y_{ij},$$

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 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, 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 no-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 liner 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 a working population model holds for sample B. We assume 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 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 $\mathbf{x}_{ij}$. 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 liner random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized liner random intercept model where $h(\cdot)$ is the log link function and the individual $y_{ij}i$ 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:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i)}(y_{ij} - \hat{\mu}_{ij}) + \sum_{j \in A_i} \hat{\mu}_{ij}\right\}, \tag{4}$$

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 tU 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.

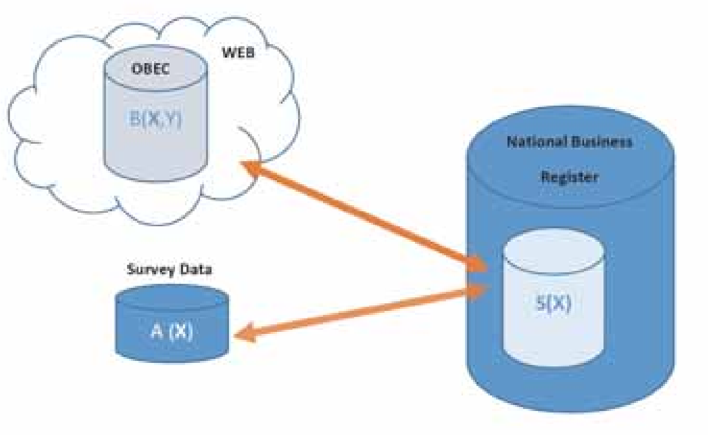


Fig. 1 Application Setting

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

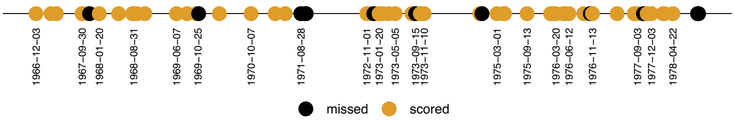


Fig. 1 Penalty history over time of the player Gerd Müller for the time period from 1964 until 1979 (successful penalties in yellow, failures in black).

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.

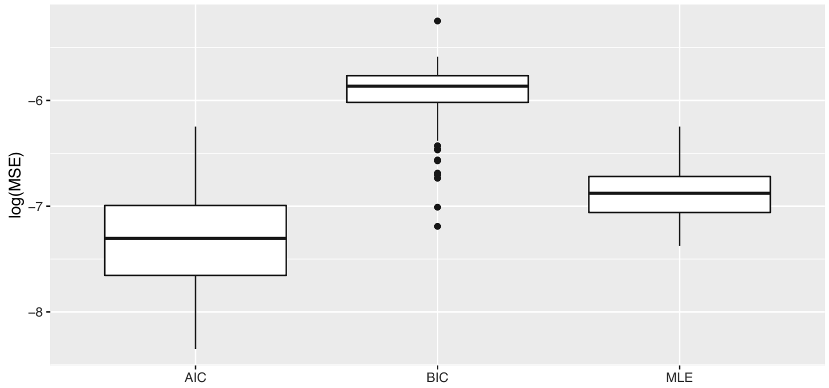


Fig. 2 Boxplots of the MSE obtained in 100 simulation runs. "AIC" and "BIC" denote the LASSO-HMM fitting schemes with λ chosen by AIC/BIC. "MLE" denotes unpenalised HMM.

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

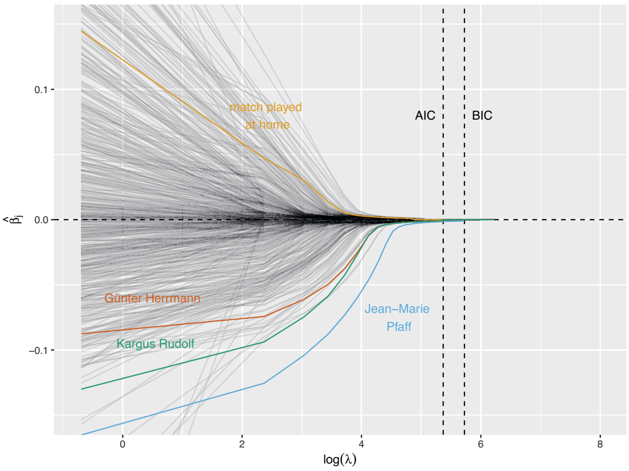


Fig. 3 Coefficient paths of all covariates considered in the LASSO-HMM models. Dashed vertical lines indicate the penalty parameters λ as selected by AIC and BIC, respectively. For BIC, no covariates are selected, whereas for the AIC the player-specific effect of Jean-Marie Pfaff is selected. The player-specific 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

$$\mathbf{y}^H = \mathbf{X}\beta^H + \mathbf{Z}^{(Hl)}\mu + \mathbf{Z}^{(Hp)}\gamma + \varepsilon^H, \qquad (1)$$

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 $\mu \sim N(0, \sigma^2_\mu)$ and $\gamma \sim Exp(\lambda_\gamma)$ in the likelihood specification for the effects of lineups and players, respectively. The idiosyncratic error term is $\varepsilon^H \sim N(0, \sigma^2_{\varepsilon^H})$. 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

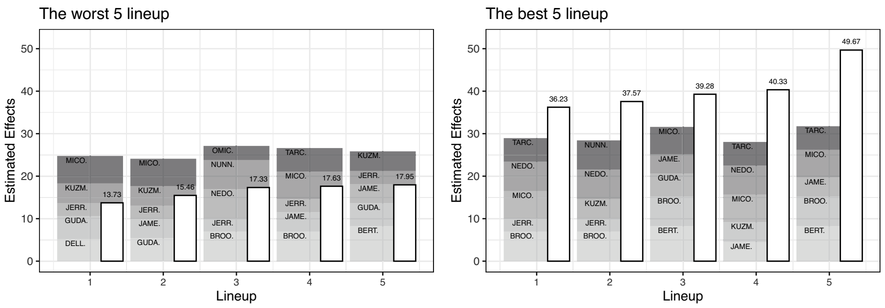


Fig. 1 The composition of total lineup effects (identified with the white bar in the figures) for the worst (left-panel) and the best (right-panel) five lineups in Milan team, including player effects and negative (or positive) synergies due to lineup composition.

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].

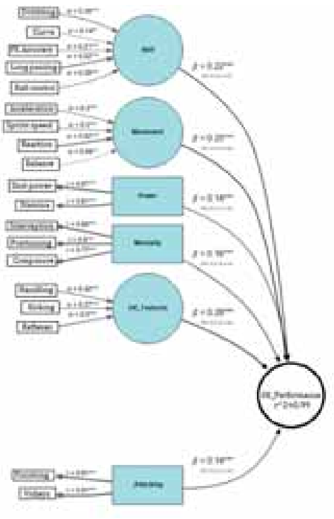


Fig. 1 PLS-SEM GK performance model after 5000 bootstrap resampling.

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-----	1111111111111111111111

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*).

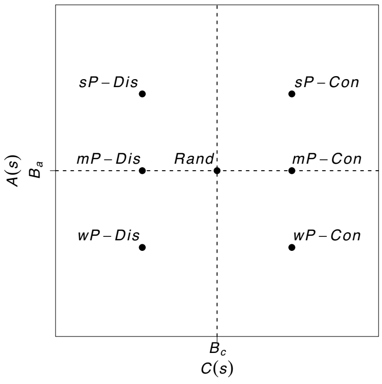
4. An industry that is geographically dispersed and positively spatially autocorrelated may be categorized as a *strongly polarized dispersed* industry (*sP-Dis*) as it is depicted by employment dispersion in neighboring regions.
5. An industry that is geographically dispersed but not spatially autocorrelated 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.

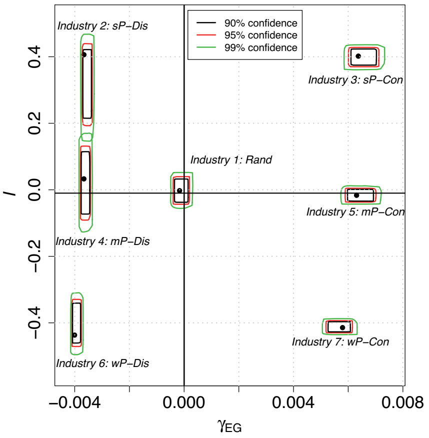
We specified $n = 49$ regions, as cells of a 7-by-7 lattice, and 7 industries, each with a total employment of 5000 and a total number of firms of 200 units. For each industry, the values of the vector $\mathbf{z}$ of the 200 firms' employment shares in the industry's total employment have been generated 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, $i = 1,2,\dots,n$, 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

$$wGM_i = \prod_{j=1}^{m} z_{ij}^{w_{ij}}, \quad i = 1,2,\dots,n, \tag{1}$$

where $w_{ij} \in [0,1]$ is the weight attributed to the j-th indicator for the i-th unit, $i = 1,2,\dots,n$ and such that $\sum_{j=1}^{m} w_{ij} = 1$.
For $i = 1,2,\dots,n$ 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}, $j = 1,2,\dots,m$.

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.

$$wS^2_{0,i} = F(wGM_i) = \sum_{j=1}^{m} w_{ij}(\ln z_{ij} - \ln wGM_i)^2, \quad i = 1,2,\ldots,n. \tag{3}$$

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 $wS^2_{0,i_1} > wS^2_{0,i_2}$, 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:

$$pwGM^{\pm}_i = wGM_i \exp\left\{\pm wS^2_{0,i}\right\}, \quad i = 1,2,\ldots,n. \tag{4}$$

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

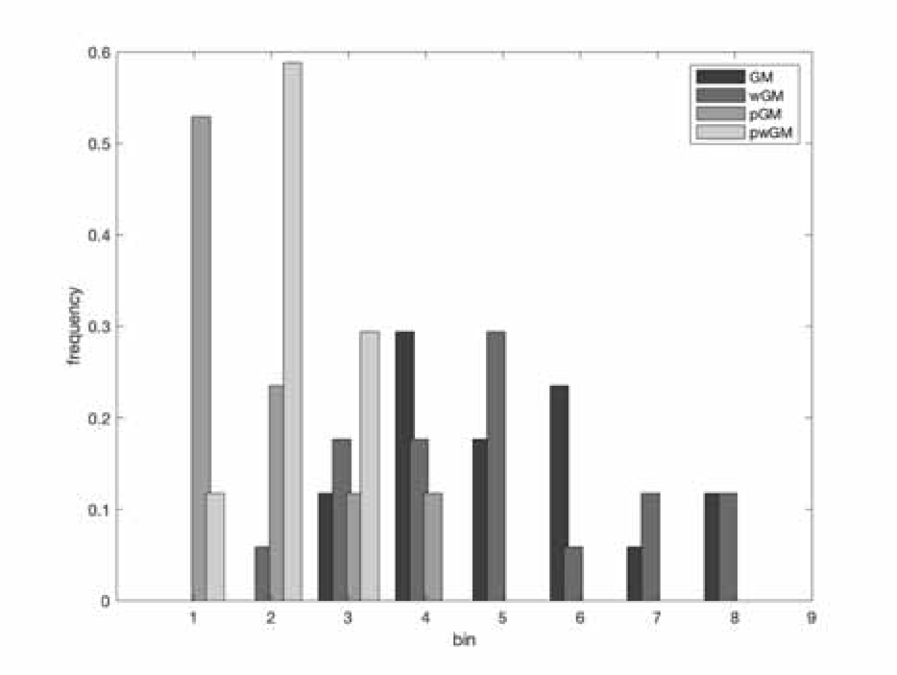


Fig. 1 **Distribution** comparisons

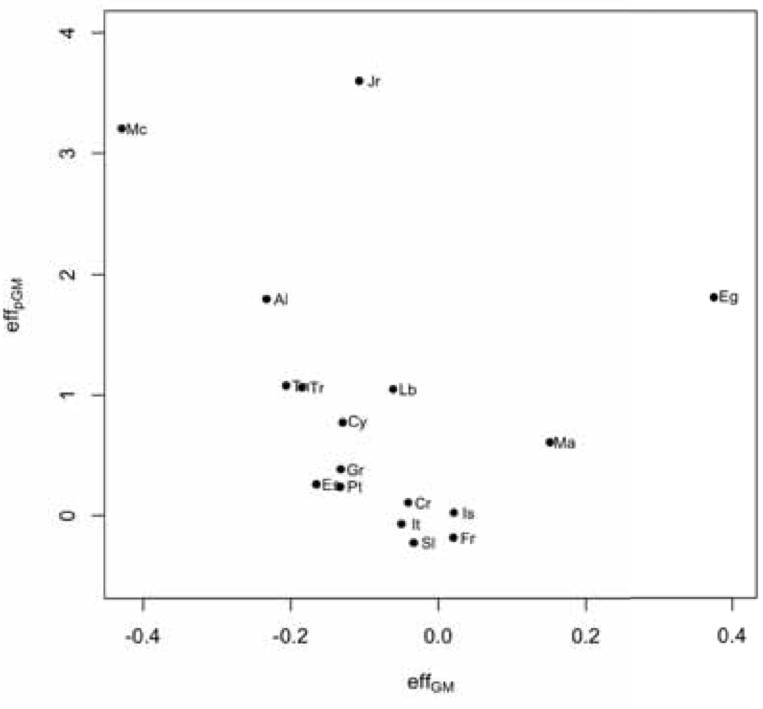


Fig. 2 Comparison of the role of weights

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 indicators. 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)):

$$w_{ij} = \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2 \text{ if } d_{ij} < h_i, 0 \; otherwise \qquad (2)$$

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)):

$$CV = \sum_{i=1}^{n} \left(\bar{y}_i - \hat{\bar{y}}_{\neq i}(h_i)\right)^2 \qquad (3)$$

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 $\mathrm{h_i}$ 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).

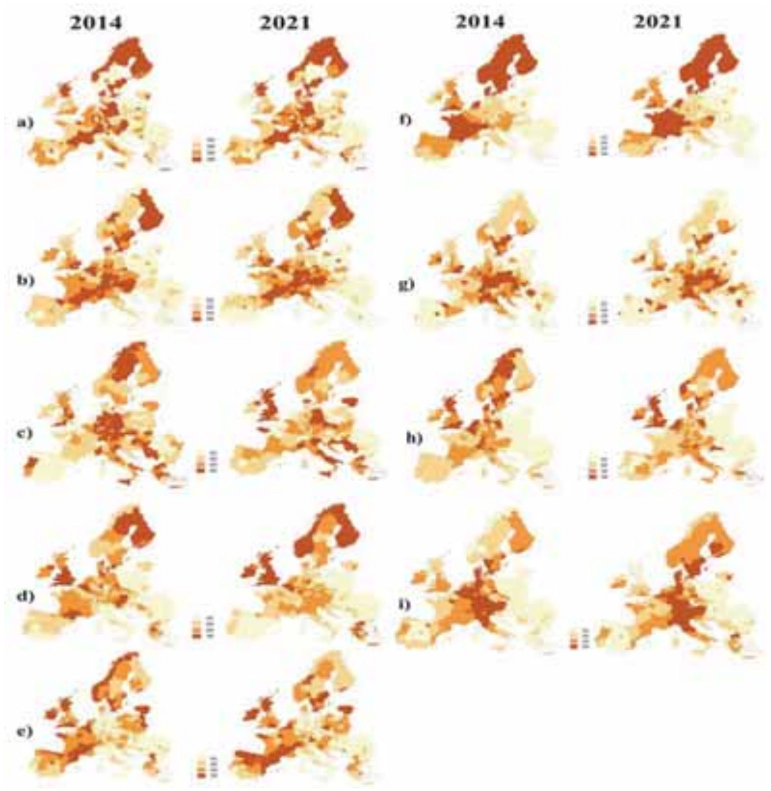
Moreover, The European Commission proposes the Regional Innovation Index (RII). The RII is a composite indicator calculated as the unweighted average of the scores of RIS variables. It combines the output side of innovation (e.g., the number of patent applications per billion GDP) and input variables (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

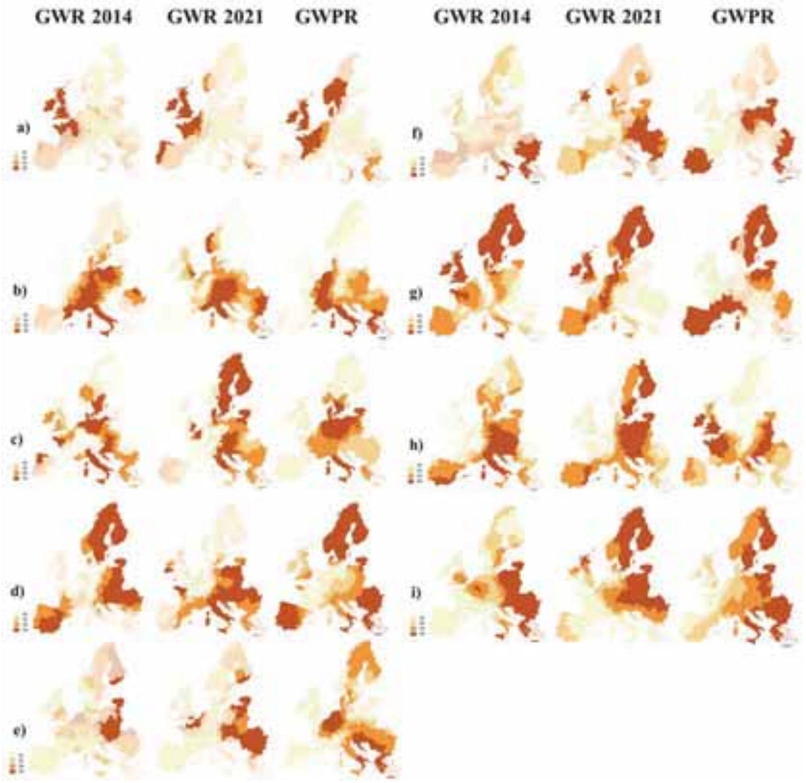
[1] Notably, for the three models (GWR 2014 and 2021, and GWPR) the optimal bandwidth procedure converges towards adaptative bi-square kernel but it highlights three different nearest neighbours: 85 (GWR 2014), 62 (GWR 2021), and 93 (GWPR). This is not surprising since CV procedure is based on the value 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:

$$X_s = F_s Q Q^\top D_s A_s^\top = (F_s Q)(A_s D_s Q)^\top = F_s D_s Q Q^\top A_s^\top = (F_s D_s Q)(A_s Q)^\top \ . \quad (1)$$

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:

$$S_W = S_T - S_B = X^\top X - \bar{X}^\top (G^\top G)\bar{X} = X^\top X - X^\top HX \ , \tag{3}$$

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:

$$S_W = I_s - F^\top HF = F^\top (I_n - H)F \ . \tag{5}$$

Then, the Fisher's LDA of F requires the solution of:

$$F^\top HFQ = Q\Lambda \ , \tag{6}$$

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.

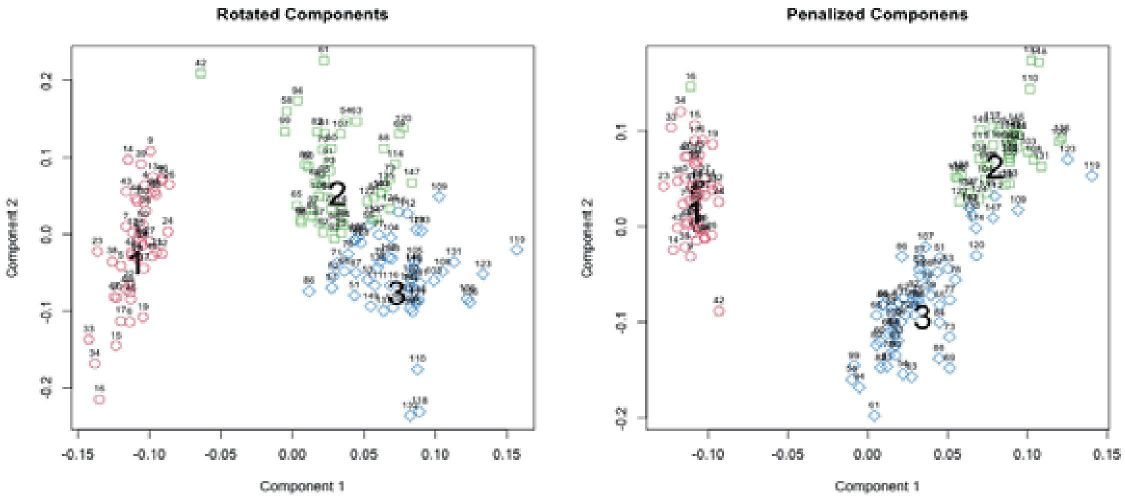


Fig. 1 Iris data: plots of the flowers on rotated (left) and penalized PCs.

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
```

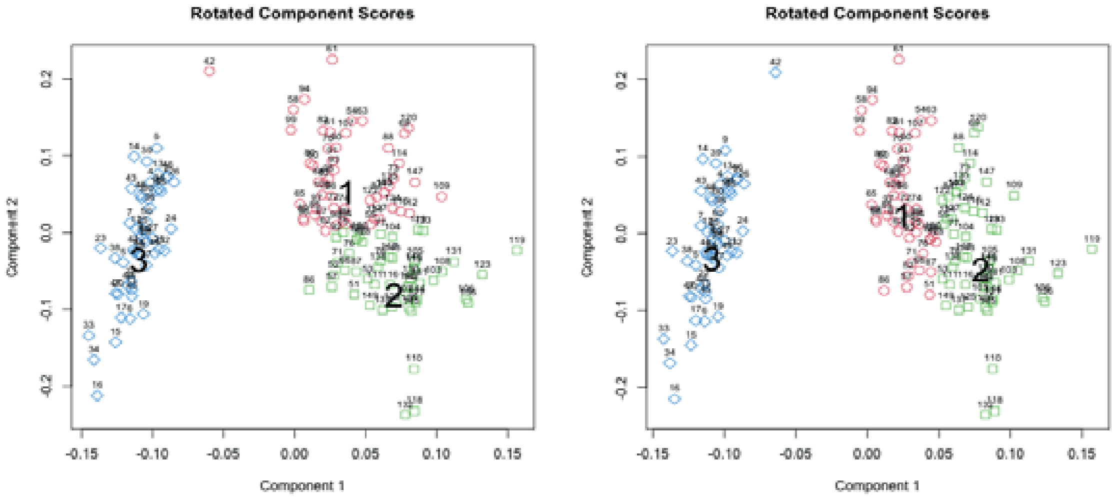


Fig. 2 Iris data: clustering rotated PCs, with Euclidean (left) and Manhattan allocation.

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 $V = \{1,\dots,p\}$ 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 $\boldsymbol{y} = (y_1,\dots,y_p)'$, 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 $\mathbf{y}$, $y_{-(i,j)}$, 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

$$y_j = \beta_{j,0} + \sum_{i \neq j} \beta_{j,i} y_i + \varepsilon_j, \quad j = 1,\dots,p.$$

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 $x_{n,j} = c \quad if\ \gamma_{j,c-1} \leq y_{n,j} \leq$

$\gamma_{j,c}$, $c = 1,\dots,C$; $\gamma_{j,0} = -\infty, \gamma_{j,C} = \infty$. A uniform prior distribution is chosen for the threshold parameters, truncated to the region $\{\gamma_{j,c} \in \mathscr{R}, \gamma_{j,c-1} \leq \gamma_{j,c} \leq \gamma_{j,c+1}\}$, $c = 1,\dots,C-1$, $\forall j$, 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).

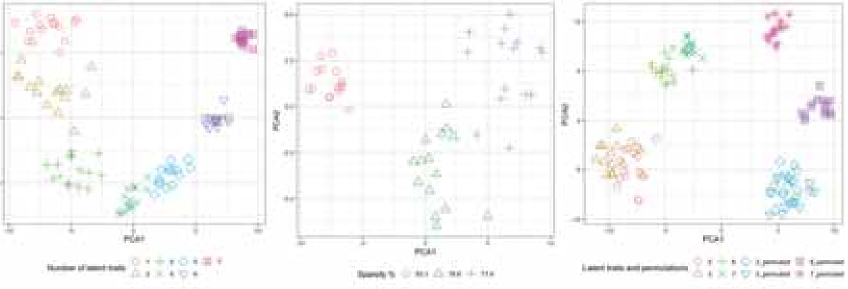


Fig. 1 Plot of the 2 first principal component scores for the simulated networks with varying numbers of latent traits, sparsity levels and item-latent traits structures.

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 $\boldsymbol{0}$ 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 $E(\boldsymbol{F}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \boldsymbol{I}_m$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi}$, 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:

$$\min_{\boldsymbol{\Lambda},\boldsymbol{\Psi}} Q(\boldsymbol{\Lambda}), \text{ subject to } \quad \ell(\boldsymbol{\Lambda},\boldsymbol{\Psi}) \le \ell^*, \tag{6}$$

where ℓ^* ($\ell^* \ge \hat{\ell}$) 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 $(\hat{\theta}_1, \dots, \hat{\theta}_n)$, 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 $i = 1, \dots, n$. Moreover, note that under typical mixture model assumptions for clustering, we have that $\hat{\theta}_1 \neq \dots \neq \hat{\theta}_n$. 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 $(\theta_1, \dots, \theta_{K_n})$. Therefore, the corresponding posterior distribution is

$$p(K_n, \boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{c=1}^{K_n} \prod_{i:c_i=c} k(y_i; \theta_c) \times \mathrm{EPPF}(n_1, \dots, n_{K_n}) \times P(d\boldsymbol{\theta}), \tag{3}$$

which can be equivalently represented as the cost function $-\log(p(K_n, \boldsymbol{c}, \boldsymbol{\theta} \mid \boldsymbol{y}))$, i.e.

$$C(K_n, \boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) = C_{\mathrm{lik}}(K_n, \boldsymbol{c}, \boldsymbol{\theta}; \boldsymbol{y}) + C_{\mathrm{part}}(K_n, \boldsymbol{c}; \alpha) + C_{\mathrm{base}}(K_n, \boldsymbol{\theta}),$$

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

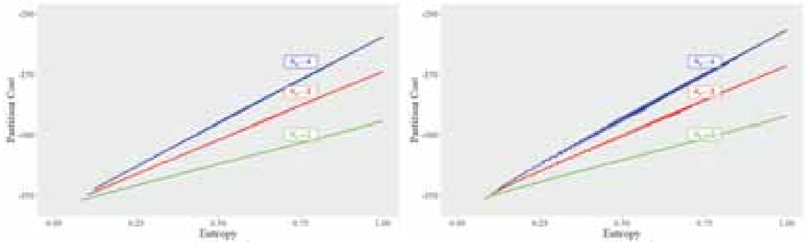
$$C_{\text{base}}(K_n, \boldsymbol{\theta}) = -\sum_{c=1}^{K_n} \log P_0(d\theta_c),$$

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 $(n_1, \dots, n_{K_n})$, i.e.

$$S(n_1, \dots, n_{K_n}) = -\sum_{c=1}^{K_n} \frac{n_c}{n} \log_{K_n} \frac{n_c}{n}.$$

Overall the EPPF acts favoring frequencies $(n_1, \dots, n_{K_n})$ 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 $(n_1, \dots, n_{K_n})$. 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:

$$\boldsymbol{c}^* = \underset{\hat{\boldsymbol{c}}}{\operatorname{argmin}} \, \mathbb{E}[L(\boldsymbol{c}, \hat{\boldsymbol{c}}) \mid y_{1:n}] = \underset{\hat{\boldsymbol{c}}}{\operatorname{argmin}} \sum_{\hat{\boldsymbol{c}}} L(\boldsymbol{c}, \hat{\boldsymbol{c}}) \mathbb{P}(\boldsymbol{c} \mid y_{1:n}),$$

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 $\{\boldsymbol{c}_m, m = 1,\dots,M\}$, $\lambda$
Output: point estimate $\boldsymbol{c}^*$
Compute $S(\boldsymbol{c}_m)$ for $m = 1,\dots,M$
Compute $w_m = \exp\{\lambda S(\boldsymbol{c}_m)\}$ for $m = 1,\dots,M$
$\bar{w}_m \leftarrow w_m / \sum_m w_m$ for $m = 1,\dots,M$
Generate $\{\tilde{\boldsymbol{c}}_m, m = 1,\dots,M\}$, sampling with replacement from $\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_M\}$ with prob. $\{\bar{w}_m, m = 1,\dots,M\}$
$\boldsymbol{c}^* \leftarrow \operatorname{argmin} \sum_{m=1}^{M} \sum_{\hat{\boldsymbol{c}}} L(\tilde{\boldsymbol{c}}_m, \hat{\boldsymbol{c}})$
```

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

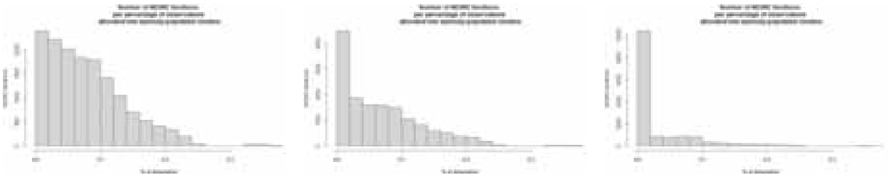
$$\bar{L}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \exp\{\lambda S(\boldsymbol{c})\} L(\boldsymbol{c}, \hat{\boldsymbol{c}}),$$

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.

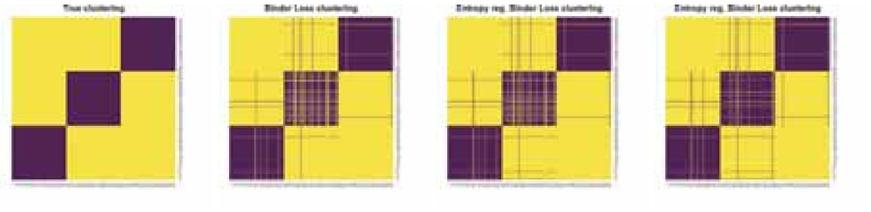
4 Simulation study

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.

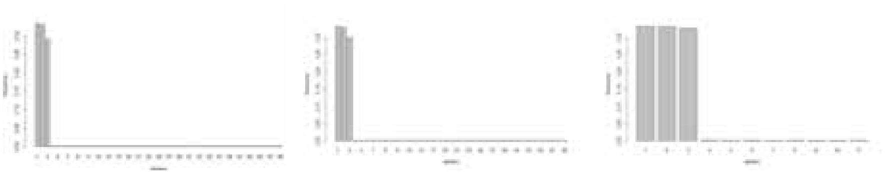
5 Results for the wine dataset

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 $\lambda = 20$.



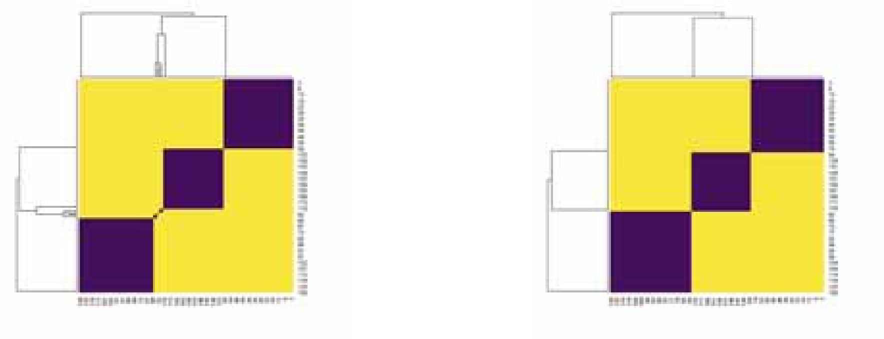
(a) Without entropy regularization.

(b) With entropy regularization for $\lambda = 10$.

(c) With entropy regularization for $\lambda = 20$.

Fig. 4: Estimated clusters' frequencies.

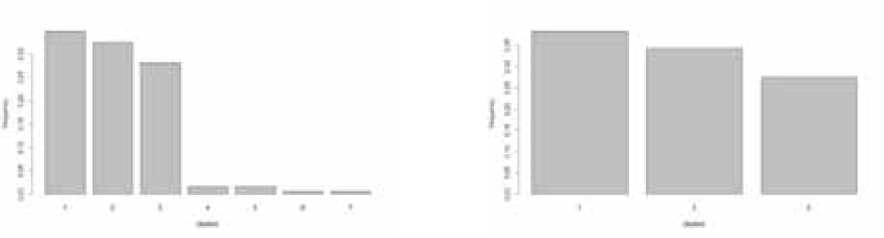
in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. Here we refer to the clustering identified by the three types of wines as "ground truth". We use the 13 constituents to estimate a Dirichlet process 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.

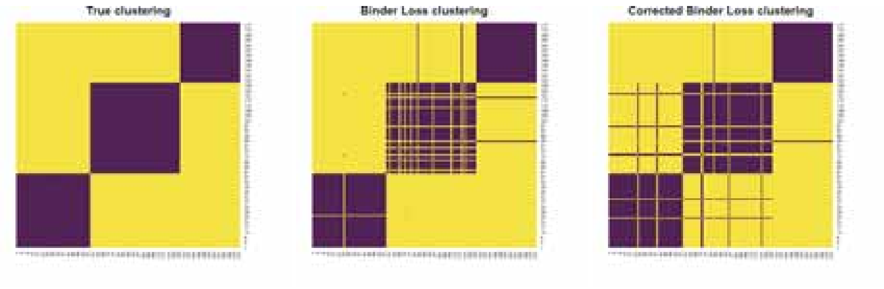
Fig. 5: Estimated partitions for the wine dataset. Darker squares denote couples of observations clustered together, observations are ordered based on co-clustering.



(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,

$$E = \big\{ x \in \mathscr{X} : \Pi(U(\theta_0)|X = x) \geq s \big\}.$$

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).

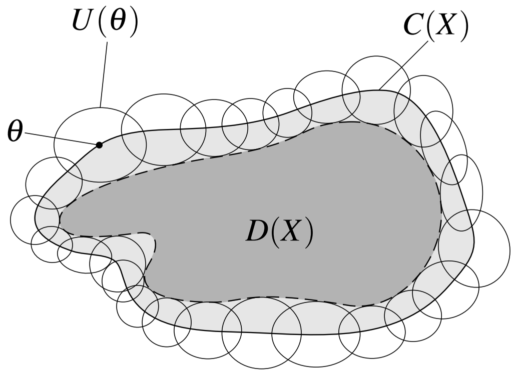


Fig. 1 The relation between a credible set $D(X)$ and its U-enlargement $C(X)$ in Venn diagrams: the extra points θ in $C(X)$ not included in the credible set $D(X)$ are characterized by non-empty intersection $U(\theta)\cap D(X) \neq \varnothing$. [From (Kleijn, 2021)]

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.)

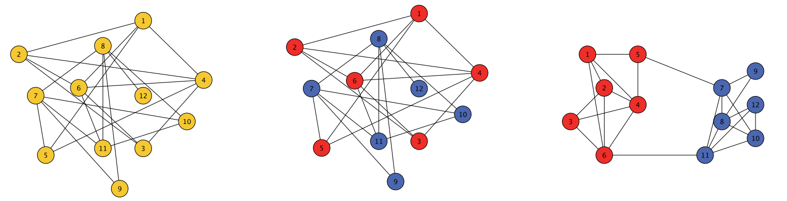
Coming back to the role of the posterior probability β, sharpness of the lower bound (3) is crucial for the construction of confidence sets from credible sets: if the bound is sharp, the required credible level is lowest, enlargement radii are lowest and the resulting confidence sets are smallest; versions 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),

$$\Bigl(\bigl(\sqrt{a_n}-\sqrt{b_n}\bigr)^2-2\Bigr)\log(n)+\log(\log(n)) \to \infty, \tag{8}$$

(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 $k_n = 0$):

$$E_{\theta_{0,n}}\Pi\bigl(\{\theta_{0,n}\}\bigm| X^n\bigr) \geq 1-\tfrac{1}{2}n\rho(p_n,q_n)^{n/2}\,e^{n\rho(p_n,q_n)^{n/2}}, \tag{9}$$

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 $0<\lambda_n<1/2$,

$$\begin{aligned} &E_{\theta_{0,n}}\Pi\bigl(k(\theta,\theta_{0,n})\leq\lambda_n n\bigm| X^n\bigr)\\ &\qquad\geq 1-\frac{1}{2}\Bigl(e\lambda_n^{-1}\rho(p_n,q_n)^{n/2}\Bigr)^{\lambda_n n}\Bigl(1-e\lambda_n^{-1}\rho(p_n,q_n)^{n/2}\Bigr)^{-1}, \end{aligned} \tag{10}$$

which means that the posterior achieves almost-exact recovery at Hamming rate $k_n=\lambda_n n$, whenever,

$$\lambda_n n\Bigl(\log(\lambda_n)+\frac{n}{4}\bigl(\sqrt{p_n}-\sqrt{q_n}\bigr)^2-1\Bigr) \to \infty.$$

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.

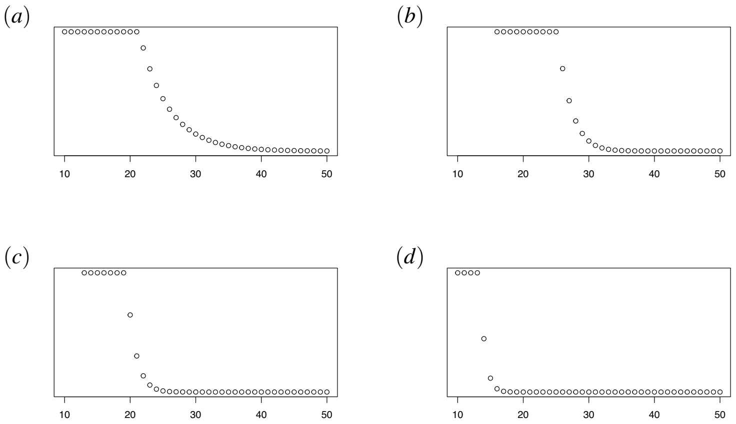


Fig. 3 Credible levels $1-\gamma$ required for a confidence set of confidence level 0.95 as a function of graph size $10 \leq n \leq 50$, with fixed edge probabilities $p = 0.9$ and $q = 0.1$ and various enlargement radii: in (a) we do not enlarge the credible set, in (b) – (d), we enlarge by Hamming 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 $\mathbb{N}_+ = \{1,2,\dots\}$ be the set of positive integers and, for a given $d \in \mathbb{N}_+ \setminus \{1\}$ and any $\lambda \in \mathbb{R}$, set $[d] = \{1,\dots,d\}$ and $\lambda_d := (\lambda,\dots,\lambda) \in \mathbb{R}^d$. In what follows, the computation of maxima as well as sums, products, etc. involving vectors in $\mathbb{R}^d$ operate component by component. Let $X_i, i = 1,2,\dots$ be i.i.d. d-dimensional r.v.s with distribution F. If there exist norming sequences $a_m \in (0,\infty)^d$ and $b_m \in \mathbb{R}^d$, $m = 1,2,\dots$ 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

$$C_\eta(u) = \exp\left(-d\int_{\mathscr{S}} \max\{(-\ln u_d)w_1, (-\ln u_1)w_2, \dots, (-\ln u_{d-1})w_d\} \mathrm{d}\eta(w)\right),$$

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

$$\mathbb{P}\left(\cap_{j=1}^d \{X_{i,j} > F_j^{\leftarrow}(e^{-1/t})\}\right) \sim dt^{-1} \int_{\mathscr{S}} \min(w_1, \dots, w_d) \mathrm{d}\eta(w),$$

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

$$C_\eta(u) = C_\eta(u_1^{1/k}, \dots, u_d^{1/k})^k,$$

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 $P : \mathscr{S} \mapsto \mathscr{R} : (w_1, \dots, w_{d-1}, w_d) \mapsto (w_1, \dots, w_{d-1})$. Then, for any $\eta \in \mathscr{E}$, there exist masses $p_j \in [0, 1/d]$, $j = 1, \dots, d$, 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 $G_{\theta,\eta}(x) = C_\eta(G_{\theta_1}(x_1), \dots, G_{\theta_d}(x_d))$ 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 $M_{m,i} = \max(X_{(i-1)m+1}, \dots, X_{im})$, $i = 1, \dots, n$, are computed componentwise from disjoint blocks of m r.v.s, extracted from a simple random sample $X_1, \dots, X_{nm}$, 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

$$\begin{aligned}
& w_k := \{w_q, q \in \mathscr{Q}_k\} \sim \text{Dirichlet}(\alpha_k) \\
& E_k | w_k \sim \text{Uniform}\left(0, \min\left\{1, \left[\max_{1 \le j \le d} \sum_{l=1}^{k-1} dk^{-1} \sum_{q \in \mathscr{Q}_k : q_j = l} w_q\right]^{-1}\right\}\right) \\
& \{\gamma_q, q \in \mathscr{Q}_k\} | E_k, w_k \sim \delta_{E_k w_k} \\
& \gamma_{ke_j} | E_k, w_k \sim \delta_{d^{-1} - \sum_{l=1}^{k-1} \sum_{q \in \mathscr{Q}_k : q_j = l} E_k w_k}, \quad \forall j \in [d]
\end{aligned}$$

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 $M_{m,1},\dots,M_{m,n}$ 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

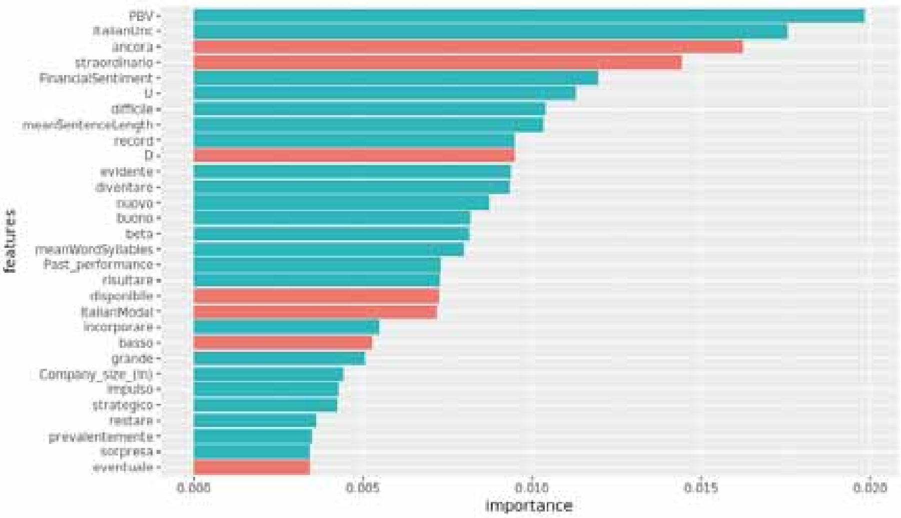


Figure 1: Most important features - RAKE PCR model (blue bars indicate a positive effect, red bars a negative effect).

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).

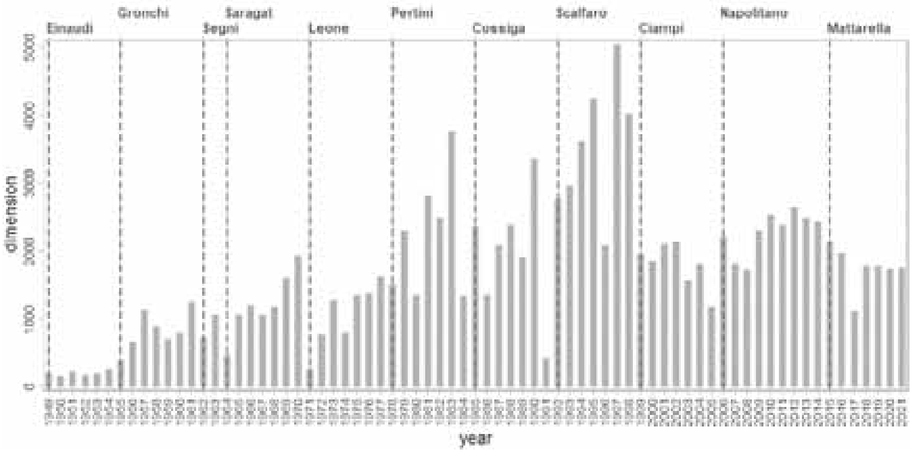


Figure 1: Subcorpora dimension: total number of word-tokens in end-of-year speeches

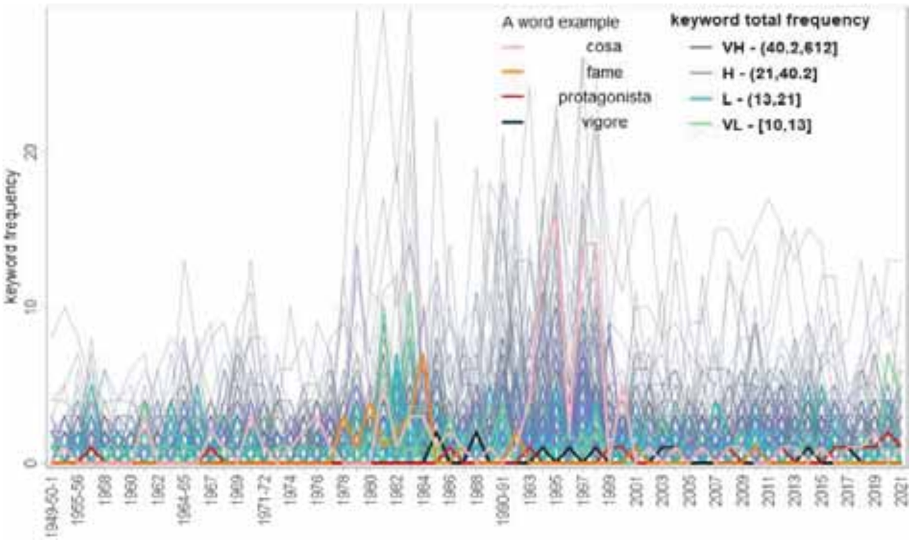


Figure 2: Keyword trajectories (original data): word frequency classes (VH=very high, H=high, L=low, VL=very low) are color-highlighted and a word from each frequency class is drawn

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: $y_{ij}= n_{ij} /(n_{i.} \surd n_{.j} / n)$ (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).

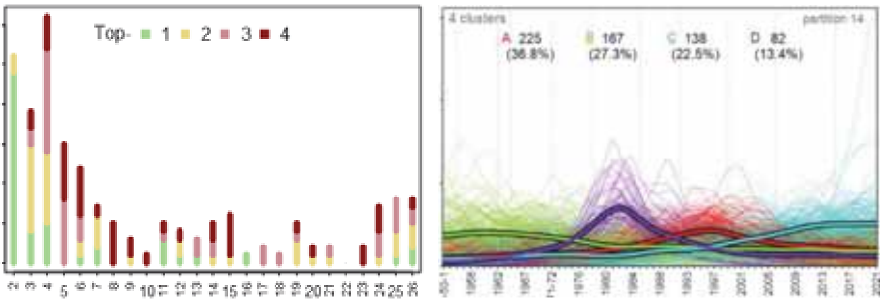


Figure 3: (a) Top-1 to top-4 cluster numbers (on the left) and (b) Best clustering solution: 4 clusters as in the best partition (14) according to the mean Rand index (on the right)

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.

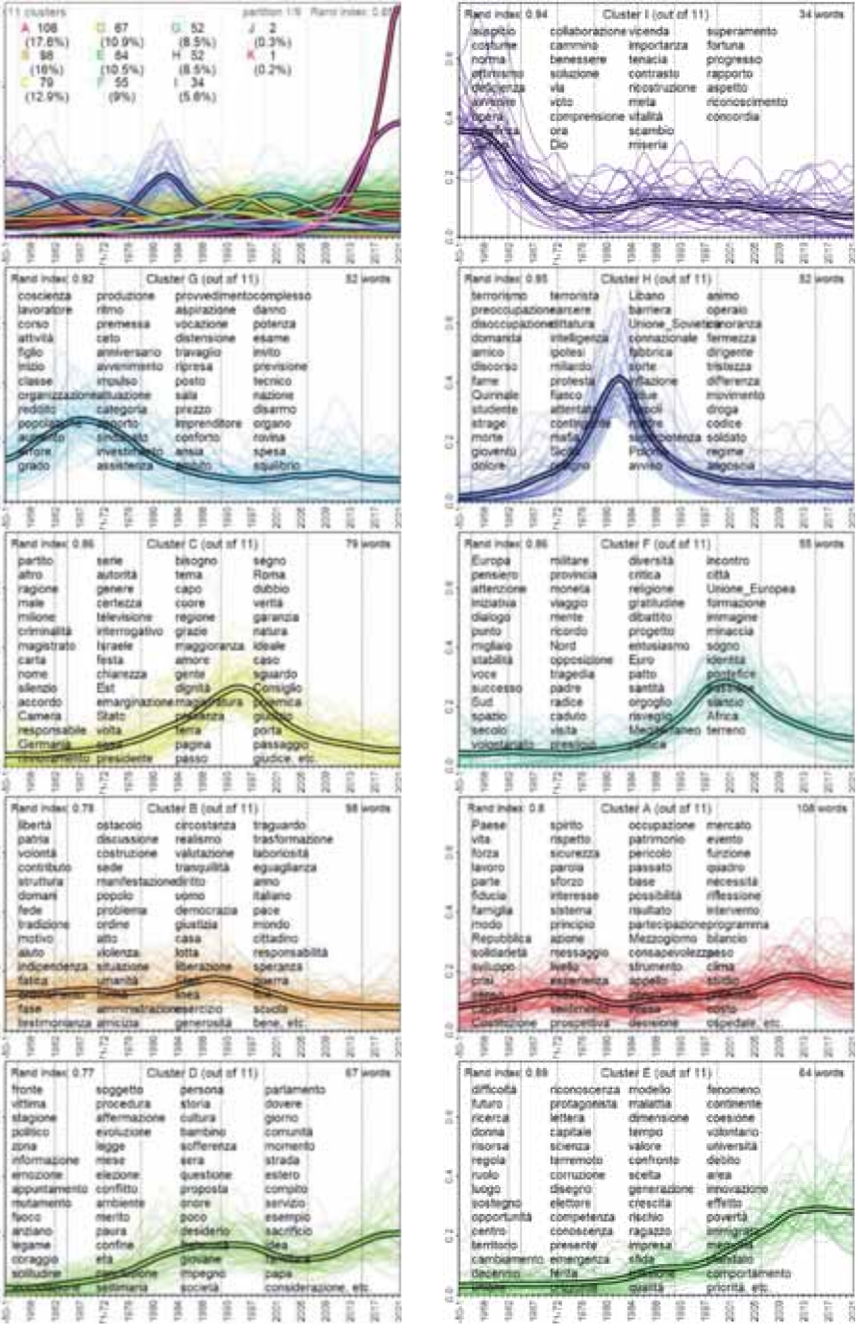


Figure 4: Final selected clustering: the overall 11 clusters and 9 clusters

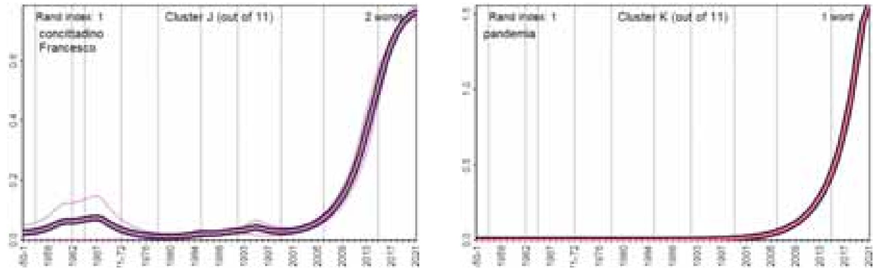


Figure 5: Clusters of sporadic and exceptional (‘black swan’ events) words

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:

$$\mathbf{d_i} = (w_{i1}, \dots, w_{ij}, \dots, w_{iq}), \tag{1}$$

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* ⟶ "I stay at home", *#andratuttobene* ⟶ "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.

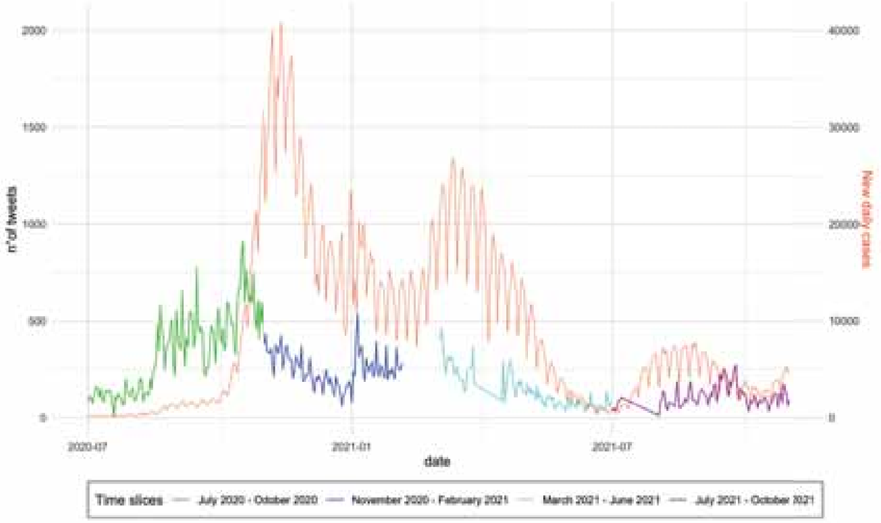


Fig. 1 Daily tweets and COVID-19 new cases in Italy (July 2020 - October 2021)

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 UD-pipe [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.

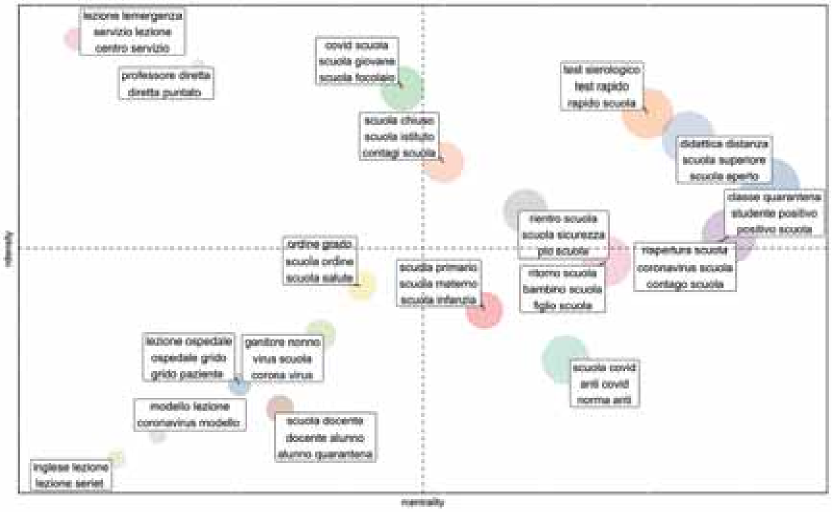


Fig. 2 Thematic map of topics discussed in the period July 2020 - October 2020

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*).

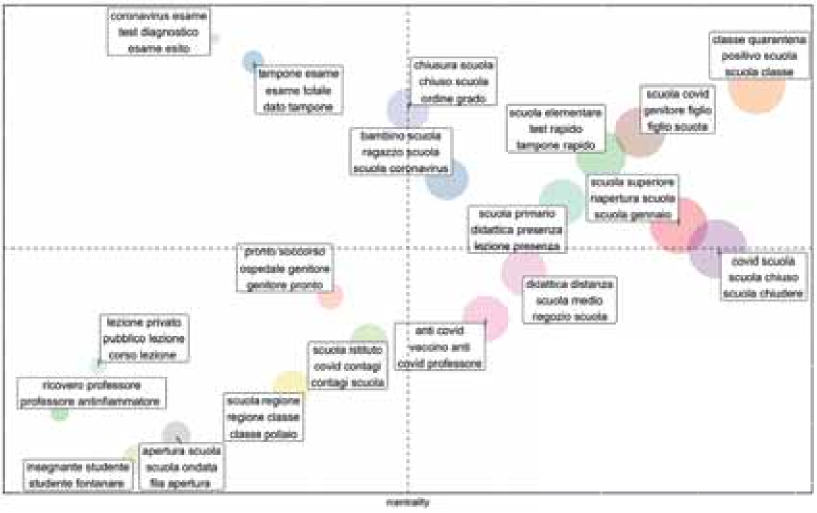


Fig. 3 Thematic map of topics discussed in the period November 2020 - February 2021

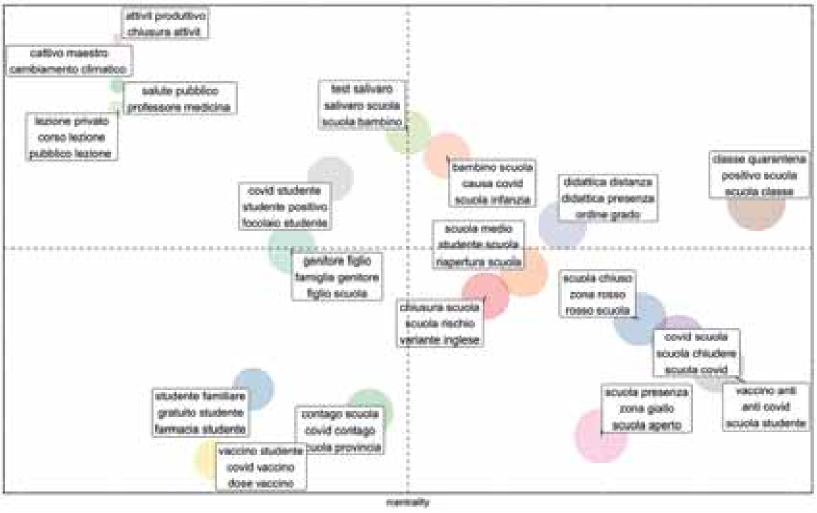


Fig. 4 Thematic map of topics discussed in the period March 2021 - June 2021

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.

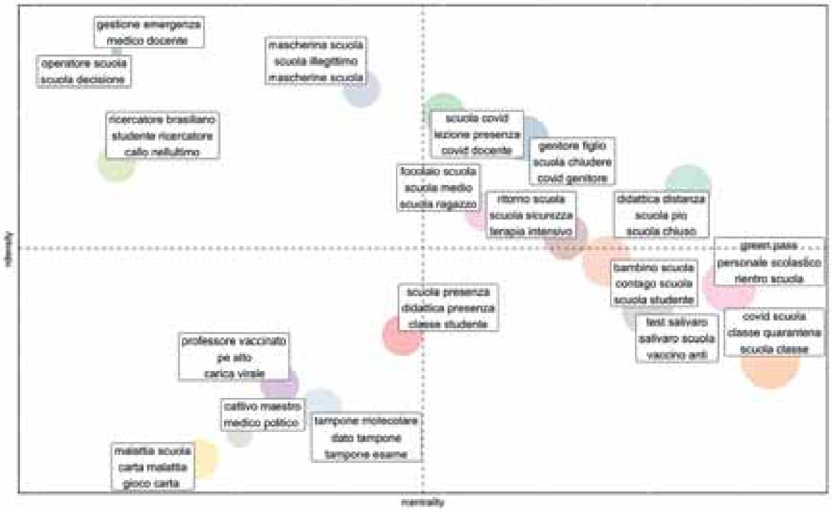


Fig. 5 Thematic map of topics discussed in the period July 2021 - October 2021

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.

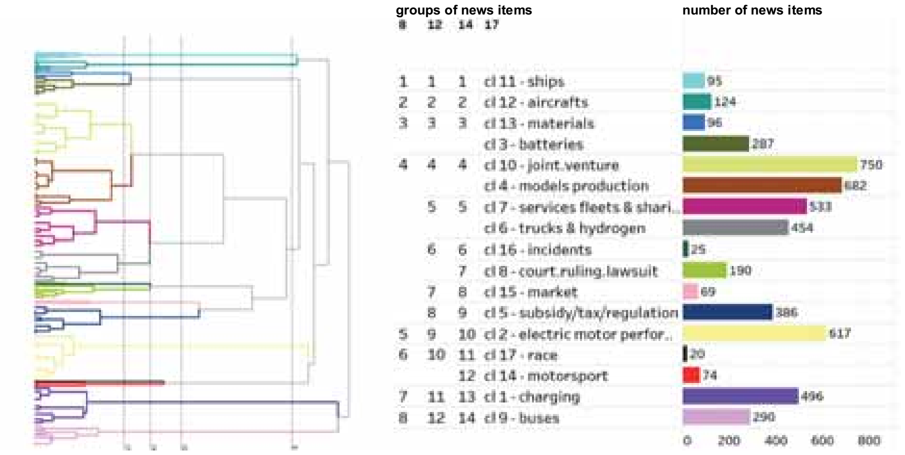
When considering the CA method, Figure 1 shows the dendrogram of the hierarchical clustering on the 10 factors of the correspondence analysis. The several cuts shown in the figure highlight the results from optimal number of clusters according to several methods (detailed results upon request). Our interpretation 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.

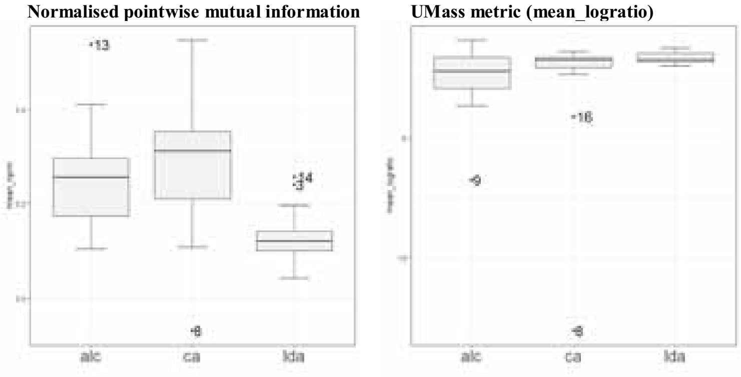
[2] The analysis in R uses the text2vec library. The selection of the terms identifying the topics is specific to the different methods used, respectively: test-value for CA; a chi-square test for ALC and the terms with the highest probability in the LDA.

Figure 1: Dendrogram: results of the CA method, with 8, 12, 14 and 17 groups



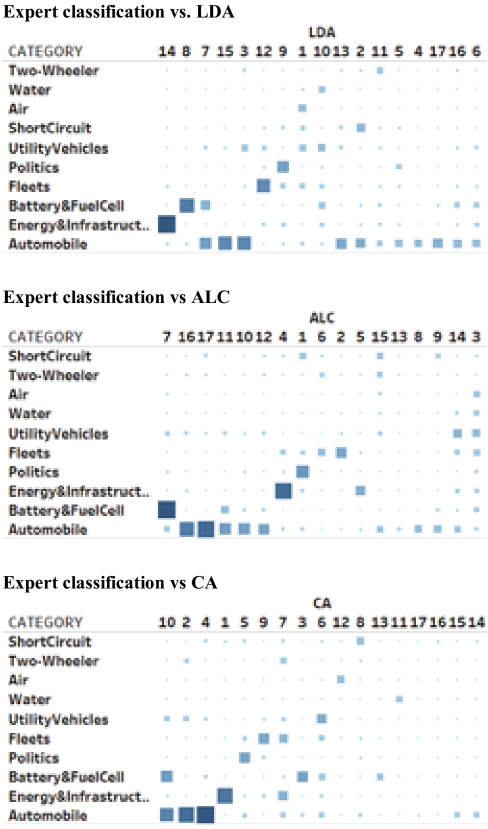
The first challenge concerns the ability of each algorithm to express semantically coherent topics. In this perspective, we implement two coherence indexes, both refer to a given number of topics that is defined with the CA method. Box plots in Figure 2 show the results of measures based, respectively, 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


Normalised pointwise mutual information **UMass metric (mean_logratio)**

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



Expert classification vs. LDA

Expert classification vs ALC

Expert classification vs CA

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:

$$EB^*_{int,k,i} = \frac{EB_{int,k,i}}{\sum_k EB_{int,k,i}}$$

(1)

$$EB^*_{ext,k,i} = \frac{EB_{ext,k,i}}{\sum_k EB_{ext,k,i}}$$

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):

$$W_k = \sum_{k=1}^{K} \sum_{j=1}^{m} |L_{kj}| \cdot \lambda_j \qquad (2)$$

[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 $[(x_i, y_i): i = 1,2, \dots N]$ 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):

$$\mu_{jh} = \mu_h + \lambda^B_h f^B_j + u_j \qquad (3)$$

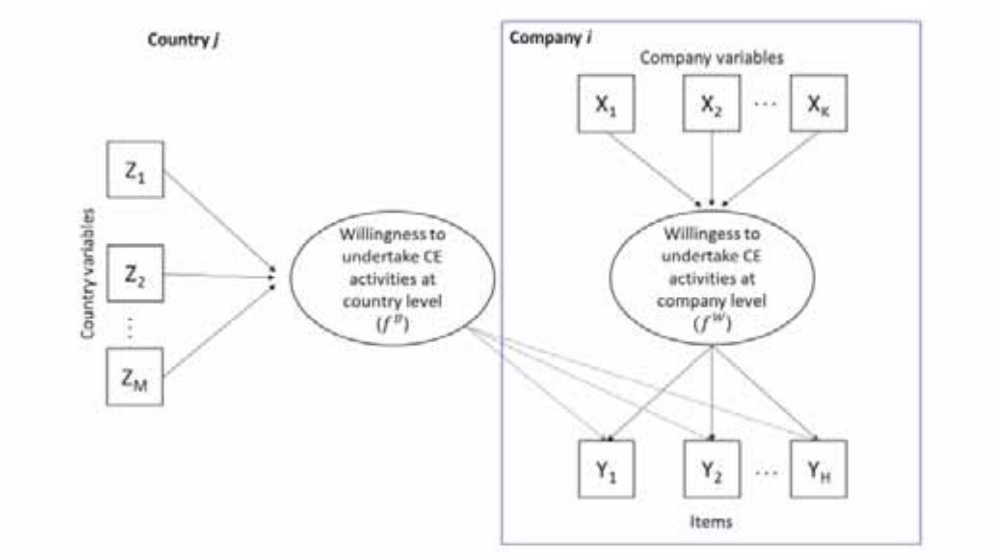

Country *j*
Company *i*
Company variables
X₁
X₂
X_K
Z₁
Z₂
Z_M
Country variables
Willingness to undertake CE activities at country level (f^B)
Willingess to undertake CE activities at company level (f^W)
Y₁
Y₂
Y_H
Items

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) $$E\{y^1_{i,t+s} - y^0_{i,t+s} | TC_{it} = 1\} = E\{y^1_{i,t+s} | TC_{it} = 1\} \text{-} E\{y^0_{i,t+s} | TC_{it} = 1\}$$

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) $$p(TC_{it} = 1) = \beta_0 + \beta_1\, prod + \beta_2\, capital - labor\ ratio + \beta_3\, age + \beta_4\, average\ labor\ cost\ + \beta_4\, flexible\ intensity\ + \beta_5\, total\ assets + \delta_1 Industry + \delta_2 Territory$$

TCTC_{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 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:

$$(5)\ \ \Delta Empl = \beta_0 + \beta_1 TC_i + \beta_2 After_{t+s} + \ \beta_3 TCa_i * After_{t+s} + \beta_4 X_i + \varepsilon$$
$$(5')\qquad \Delta Flex\ \ = \beta_0 + \beta_1 TC_i + \beta_2 After_{t+s} + \ \beta_3 TCa_i * After_{t+s} + \beta_4 X_i + \varepsilon$$

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 0.02	***	0.19 0.04	***	yes	yes	yes	yes		yes	yes

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

$$\begin{aligned}\pi_{j,t} = {} & \alpha_j + \beta_{1,j}\mathscr{J}_{j,t-1} + \beta_{5,j}\sum_{i=1}^{5}\mathscr{J}_{j,t-i} + \beta_{30,j}\sum_{i=1}^{30}\mathscr{J}_{j,t-i} + \beta_{60,j}\sum_{i=1}^{60}\mathscr{J}_{j,t-i} + \\ & + \gamma_{1,j}\mathscr{T}\mathscr{R}^{l}_{j,t-1} + \gamma_{5,j}\sum_{i=1}^{5}\mathscr{T}\mathscr{R}^{l}_{j,t-i} + \gamma_{30,j}\sum_{i=1}^{30}\mathscr{T}\mathscr{R}^{l}_{j,t-i} + \gamma_{60,j}\sum_{i=1}^{60}\mathscr{T}\mathscr{R}^{l}_{j,t-i}\end{aligned} \tag{1}$$

2. Absolute size of jumps and absolute value of TRMI

$$\begin{aligned}\pi_{j,t} = {} & \alpha_j + \beta_{1,j}J_{j,t-1} + \beta_{5,j}\sum_{i=1}^{5}J_{j,t-i} + \beta_{30,j}\sum_{i=1}^{30}J_{j,t-i} + \beta_{60,j}\sum_{i=1}^{60}J_{j,t-i} + \\ & + \gamma_{1,j}TR^{l}_{j,t-1} + \gamma_{5,j}\sum_{i=1}^{5}TR^{l}_{j,t-i} + \gamma_{30,j}\sum_{i=1}^{30}TR^{l}_{j,t-i} + \gamma_{60,j}\sum_{i=1}^{60}TR^{l}_{j,t-i}\end{aligned} \tag{2}$$

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, $\mathscr{T}\mathscr{R}^{l}_{j,t}$ 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 $\mathscr{J}$

	Regressors: TRMIs occurrence				*Regressors: TRMIs abs. value*			
	$\mathscr{TR}_1$	$\mathscr{TR}_5$	$\mathscr{TR}_{30}$	$\mathscr{TR}_{60}$	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 $\mathscr{TR}$

	Regressors: jumps occurrence				*Regressors: jumps abs. size*			
	$\mathscr{J}_1$	$\mathscr{J}_5$	$\mathscr{J}_{30}$	$\mathscr{J}_{60}$	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:

$$\begin{aligned} x_t &= g_t \tau \varepsilon_t, \\ \varepsilon_t &\sim Gamma(a_1, 1/a_1) \quad \forall t \\ g_t &= (1-\alpha_1-\beta_1-\gamma_1/2)+\alpha_1 \tfrac{x_{t-1}}{\tau}+\beta_1 g_{t-1}+\gamma_1 D_{(r_{t-1}<0)} \tfrac{x_{t-1}}{\tau}. \end{aligned} \qquad (1)$$

[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. $E\left(x_t|\mathscr{I}_{t-1}\right) = \mu_t$, 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:

$$\begin{aligned}
&x_{i,t} = g_{i,t}\tau_t\varepsilon_{i,t}\\
&\varepsilon_{i,t} \sim Gamma\left(a, 1/a\right) \quad \forall\, i = 1,...,N_t \quad and \quad t = 1,...,T\\
&g_{i,t} = (1-\alpha_1-\beta_1-\gamma_1/2) + \alpha_1\tfrac{x_{i-1,t}}{\tau_t} + \beta_1 g_{i-1,t} + \gamma_1 D_{(r_{i-1,t}<0)}\tfrac{x_{i-1,t}}{\tau_t}\\
&\tau_t = exp\left\{\omega_1 + \theta\textstyle\sum_{k=1}^{K}\varphi_k(\lambda_1,\lambda_2)X_{t-k}\right\}\\
&\varphi_k(\lambda_1,\lambda_2) = \tfrac{(k/K)^{\lambda_1-1}(1-k/K)^{\lambda_2-1}}{\sum_{j=1}^{K}(j/K)^{\lambda_1-1}(1-j/K)^{\lambda_2-1}}
\end{aligned} \tag{2}$$

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.

$$\begin{aligned}
&x_{i,t} = g_{i,t,s_{i,t}} \tau_{i,t} \varepsilon_{i,t} \\
&\varepsilon_{i,t} | s_{i,t} \sim Gamma\left(a_{s_{i,t}}, 1/a_{s_{i,t}}\right) \quad \forall\, i = 1,...,N_t \quad and \quad t = 1,...,T \\
&g_{i,t,s_{i,t}} = (1 - \alpha_{s_{i,t}} - \beta_{s_{i,t}} - \gamma_{s_{i,t}}/2) + \alpha_{s_{i,t}} \tfrac{x_{i-1,t}}{\tau_{i-1,t}} + \beta_{s_{i,t}} g_{i-1,t,s_{i-1,t}} + \gamma_{s_{i,t}} D_{(r_{i-1,t}<0)} \tfrac{x_{i-1,t}}{\tau_{i-1,t}} \\
&\tau_{i,t} = exp\left\{\omega_{s_{i,t}} + \theta \textstyle\sum_{k=1}^{K} \varphi_k(\lambda_1, \lambda_2) X_{t-k}\right\} \\
&\varphi_k(\lambda_1, \lambda_2) = \tfrac{(k/K)^{\lambda_1 - 1}(1-k/K)^{\lambda_2 - 1}}{\sum_{j=1}^{K}(j/K)^{\lambda_1 - 1}(1-j/K)^{\lambda_2 - 1}}.
\end{aligned} \tag{3}$$

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 $exp(\omega_1)$.

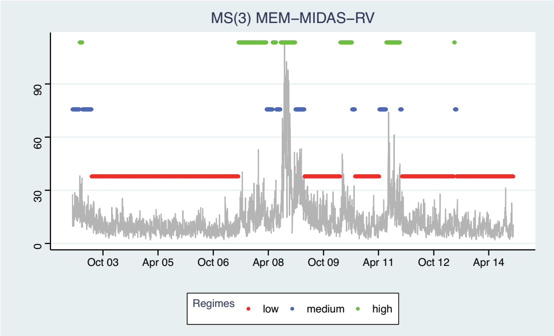


Fig. 1 Smoothed inference and annualized realized kernel volatility (gray line). Each day is assigned to a volatility regime based on the value of the smoothed probability.

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.

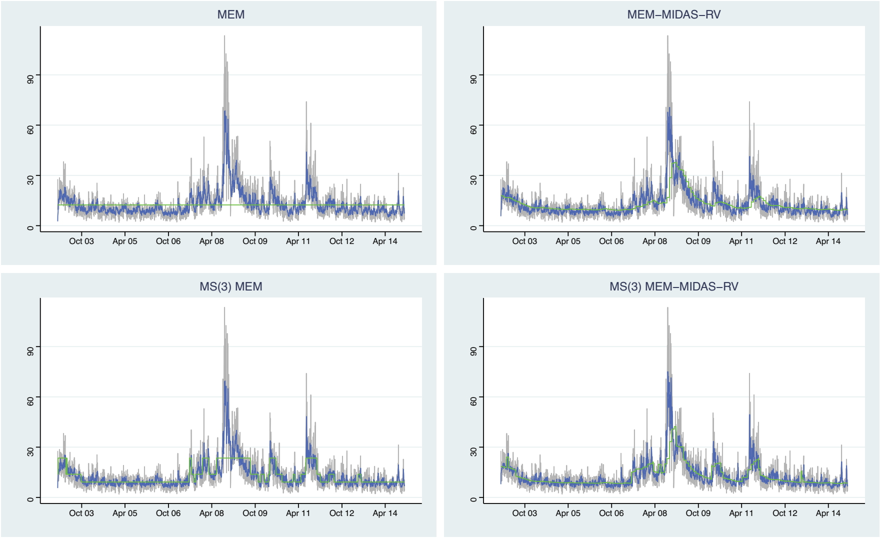


Table 2 In–sample performance of the estimated models: Sample: Jan. 2, 2003 – Dec. 31, 2014.

	MEM	MEM	MEM	MEM	MS(3)	MS(3)	MS(3)	MS(3)
		MIDAS	MIDAS	MIDAS	MEM	MEM	MEM	MEM
		IP	EMV	RV		MIDAS	MIDAS	MIDAS
						IP	EMV	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	MEM	MEM	MS(3)	MS(3)	MS(3)
		MIDAS	MIDAS	MIDAS	MEM	MEM	MEM
		IP	EMV	RV		MIDAS	MIDAS
						IP	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, $\widetilde{\mathbf{D}}_s^{(j)} = \mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}/\|\mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}\|,\ \ j = 1,\ldots,J$

5. Assign each particle a weight, $m_s^{\star(j)} = \|\mathbf{A}_s^{(j)}\mathbf{D}_{s-1}^{\star(j)}\|^{\kappa}$ for $j = 1,\dots,J$, 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-

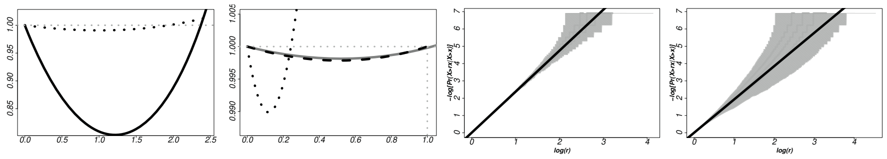


Fig. 1 Plots of κ in some GARCH models: far left, two GARCH(2,2) models (— and $\cdots$) which are second order stationary with $\phi < 1$; middle left for models an IGARCH(1,1) (black dashed) an IGARCH(2,2) (grey solid), that both have $\phi = 1$ and an ARCH(2) taken with $\phi > 1$ (black dotted). 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.

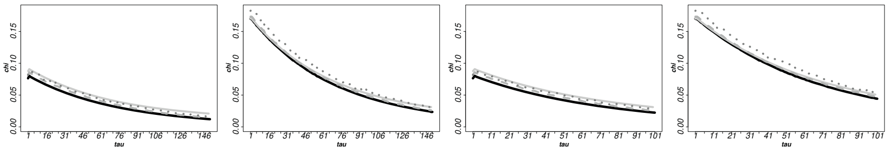


Fig. 2 Extremograms $(\tau, \chi_{X^U}(\tau))$ and $(\tau, \chi_{X^L}(\tau))$ (black solid lines) with $Z_t \sim ST(0,1,1,3)$ and empirical estimates (grey lines) for for an IGARCH(2,2) and an ARCH(2) with $\phi > 1$.

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

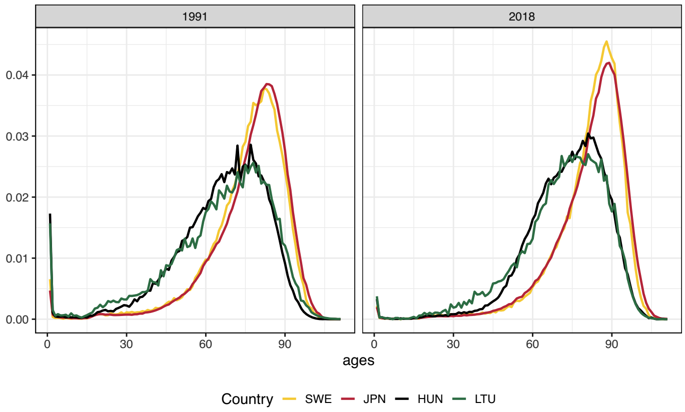


Fig. 1: Age-at-death distribution across four illustrative countries (Sweden, Japan, Hungary and Lithuania) in 1991 and 2018.

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 allows 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

$$y_i(t) = \eta_i(t) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathscr{N}(0, \sigma^2), \tag{1}$$

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

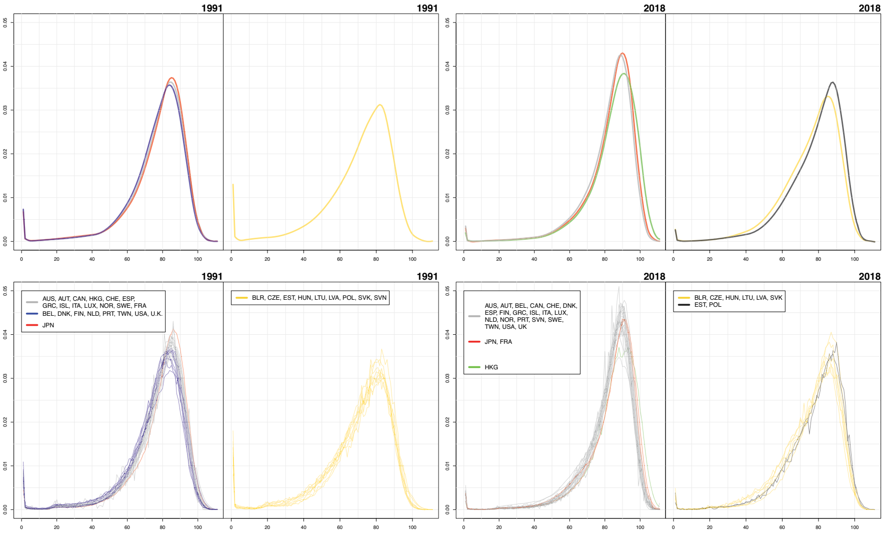


Fig. 2: Top panels: estimated functional centroids for the non-empty groups. Bottom panels: observed data and cluster allocation. Data from the same calendar year have been divided into two panels – according to the cluster allocation – to improve the graphical visualization.

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 $\boldsymbol{s}_i = \{s_{i,1}, \dots, s_{i,T}\}$ be the spike train of neuron $i = 1, \dots, n$, meaning that for all time points $t = 1, \dots, T$, $s_{i,t} \in \{0, 1\}$ 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 $\tilde{\boldsymbol{s}}_i = \{\tilde{s}_{i,1}, \dots, \tilde{s}_{i,T}\}$ 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 $\{\tilde{\boldsymbol{s}}^*_1, \dots, \tilde{\boldsymbol{s}}^*_K\}$ 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 $k = 1, \dots, K$, 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-

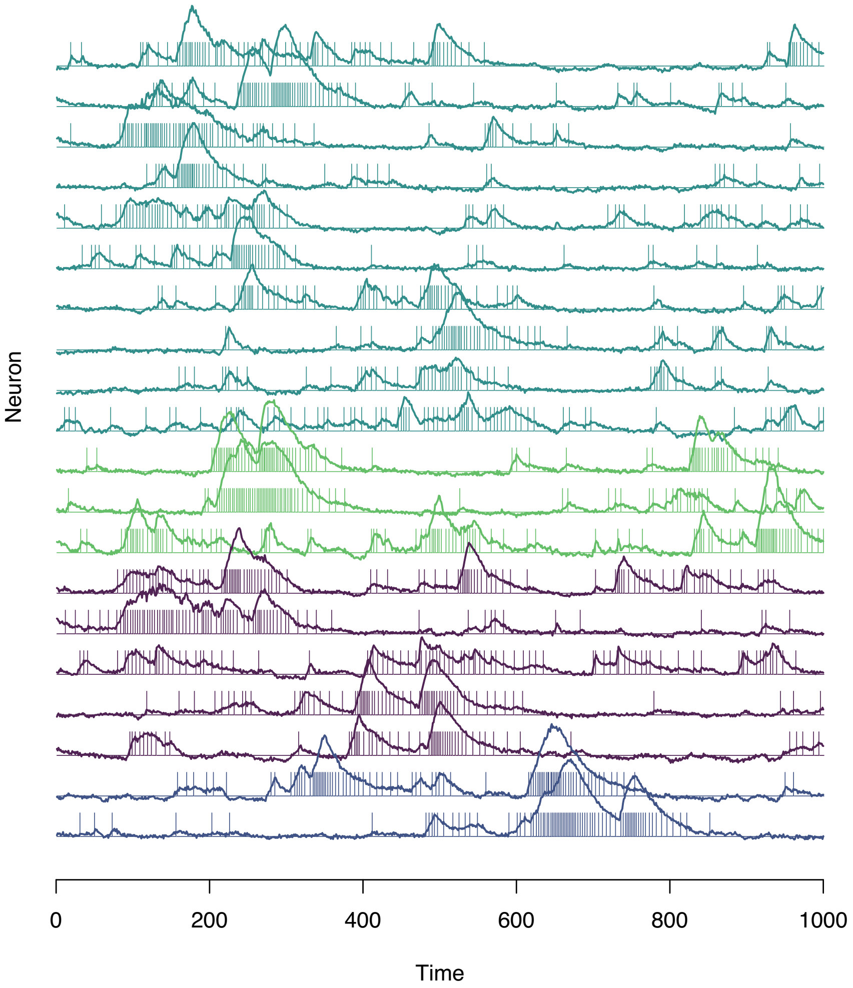


Fig. 1 Observed fluorescence traces (continuous lines) and extracted spike trains (0/1 vertical segments). The colors correspond to the estimated cluster.

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.

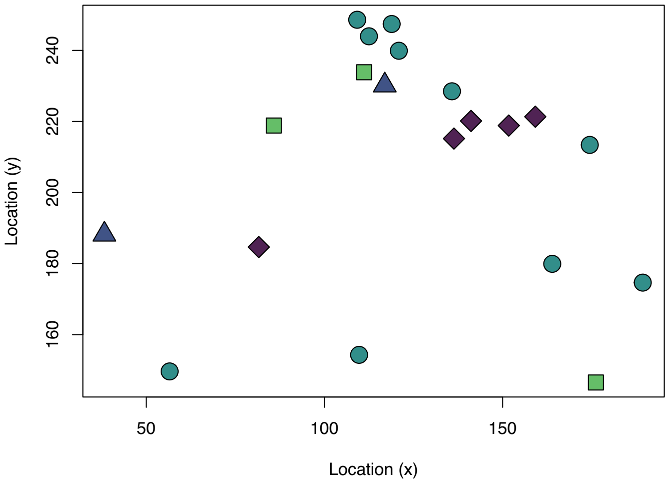


Fig. 2 Location of the neurons in the hippocampus. The colors correspond to the estimated cluster.

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.

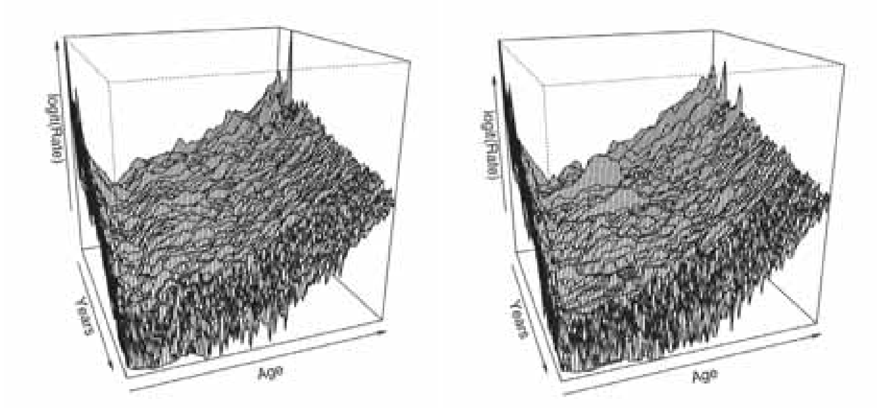


Fig. 1: Mortality rates in Iceland from 1838 to 2018, for males (left) and females (right); ages range from 0 to 80.

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 $\boldsymbol{m}_t = (m_{t1}, \dots, m_{tZ})^T$ 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 $\boldsymbol{u}_t = (u_{t0}, \dots, u_{tK})^T$. The first component u_{t0} models mortality within the first year of life, while the remaining latent variables $\{u_{t1}, \dots, u_{tK}\}$ 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

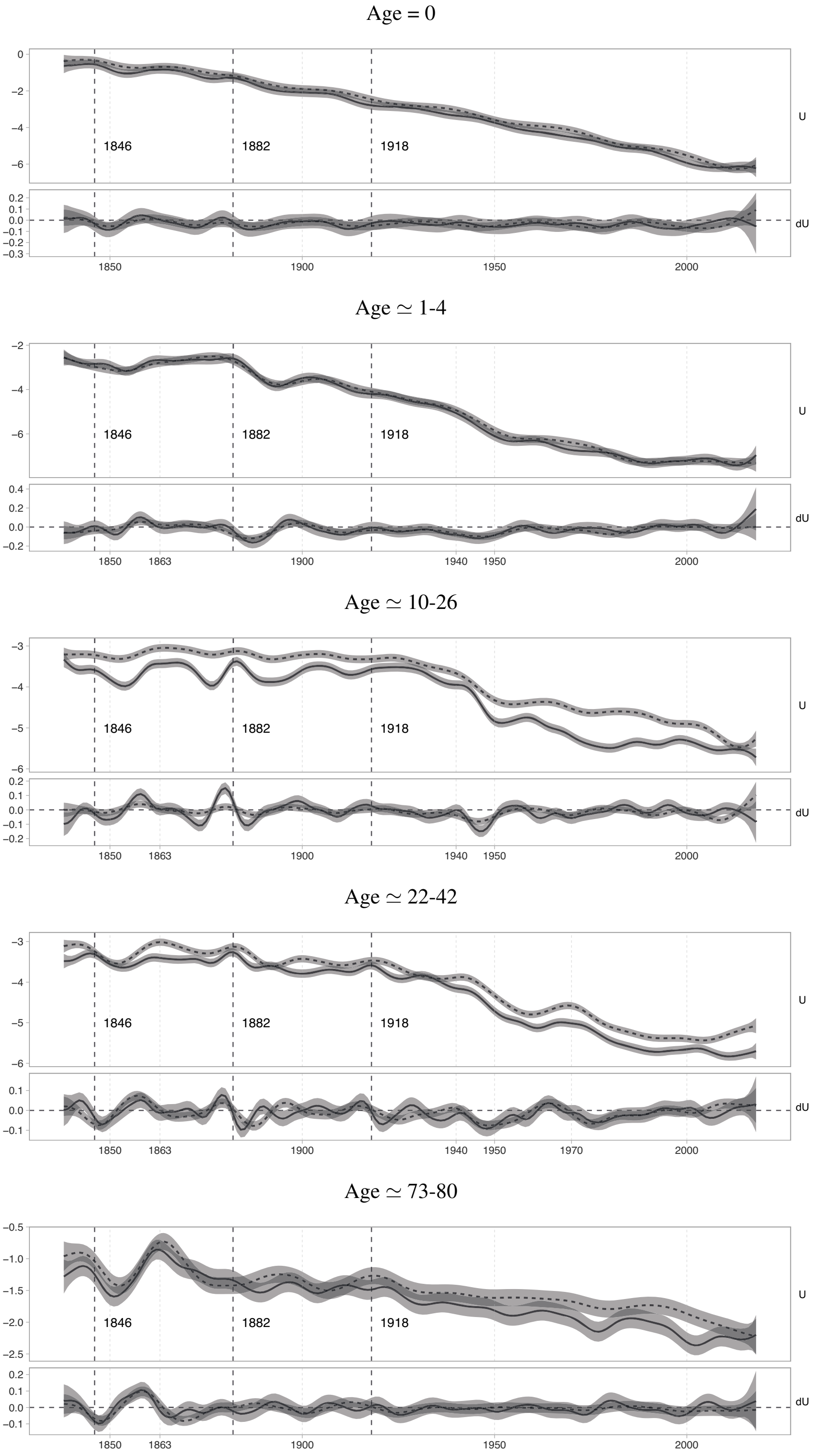


Fig. 2: Posterior smoothing distribution of u_0, u_1, u_4, u_5, and u_{10}. Female and male mortality respectively in solid and dashed line. Two standard deviation uncertainty is reported. The age range refers approximately to the splines support.

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 $i = 1, \ldots, N$ be the index of respondents, $j = 1, \ldots, J$ be the index of items and $s = 1, \ldots, S$ 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:

$$P(\boldsymbol{Y}_i\,|X_i) = \sum_{s=1}^{S} P(X_i = s\,)\prod_{j=1}^{J} P(Y_{ij}|X_i = s) = \sum_{s=1}^{S} P(X_i = s\,)P(\boldsymbol{Y}_i|\,X_i = s)\ (1)$$

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 :

$$P\big(Y_j\,\big|\,\boldsymbol{X} = \boldsymbol{s}\big) = \frac{\exp(\alpha_{js})}{1+\exp(\alpha_{js})} \quad (2)$$

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:

$$P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) = \sum_{s=1}^{S} P(X_i = s \mid \boldsymbol{Z}_i) P(\boldsymbol{Y}_i | X_i = s) \; (3)$$

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:

$$P(\boldsymbol{Y}_i \mid \boldsymbol{Z}_i) = \sum_{s=1}^{S} P(X_i = s \mid \boldsymbol{Z}_i) P(\boldsymbol{Y}_i | X_i = s, \boldsymbol{Z}_i) \; (5)$$

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:

$$\boldsymbol{\ell}(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1 = \widehat{\boldsymbol{\theta}}_2) = \sum_{i=1}^{N} \log P(\boldsymbol{Y}_i | \boldsymbol{Z}_i) \, .$$

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].

Two step LCA with measurement equivalence

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.

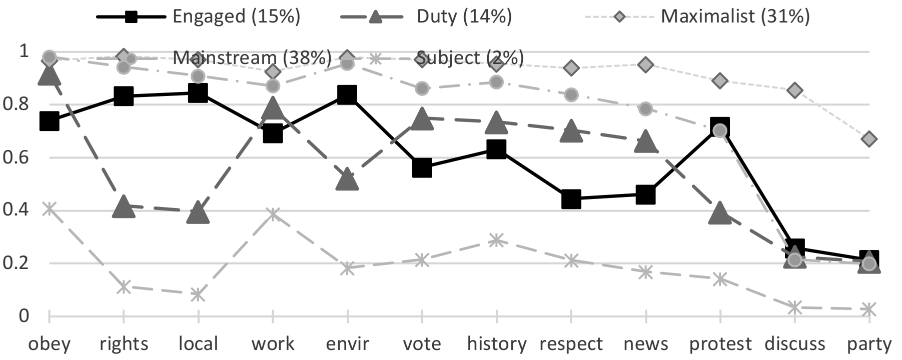
Our analyses focus on the 14 countries in which the IEA survey was conducted in all three observation periods[1]. Due to the heterogeneity of the included countries, it is imperative to test for measurement equivalence [5] to allow for a valid regression analysis that accounts for the multi-level structure 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	L²	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 $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$, 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 $\hat{\Omega}_1^{(0)}, \dots, \hat{\Omega}_K^{(0)}$, 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:

$$\text{median}_{k=1,\ldots,K}\left(||\hat{\Omega}_k - \bar{\Omega}_k||_F\right)$$

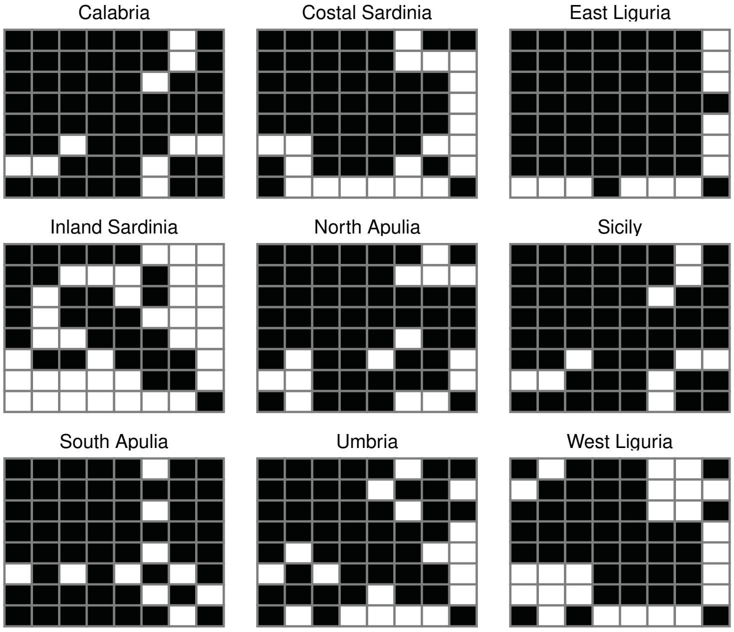


Fig. 1 Estimated precision matrices, with $\mathbf{P}_k$ defined as in (2). Black squares denote the presence of an edge between the two variables.

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 precision 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 $\mathbf{z}_i = (z_{i1}, \dots, z_{iG})' \sim \text{Multinomial}(1, (\eta_1, \dots, \eta_G))$, 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 $g = 1, \dots, G$, 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 $\mathbf{y}_i = (y_{i1}, \dots, y_{iR})$, given that node n_i belongs to the g-th group, is specified by a latent trait model with parameters b_{gk} and $\mathbf{w}_{gk}$, $g = 1, \dots, G$, and $k = 1, \dots, R$. 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 $\pi_{gk}(\mathbf{u}_i)$, $k = 1, \dots, R$, 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

     $$\hat{\mathbf{C}}_{ig}^{(t+1)} = \left[\mathbf{I} - 2\sum_{k=1}^{R} \lambda(\hat{\xi}_{igk}^{(t)}) \hat{\mathbf{w}}_{gk}^{(t)} \hat{\mathbf{w}}_{gk}^{(t)\prime}\right]^{-1},$$

     $$\hat{\mu}_{ig}^{(t+1)} = \hat{\mathbf{C}}_{ig}^{(t+1)} \left[\sum_{k=1}^{R} \left(y_{ik} - \frac{1}{2} + 2\lambda(\hat{\xi}_{igk}^{(t)}) \hat{b}_{gk}^{(t)}\right) \hat{\mathbf{w}}_{gk}^{(t)}\right].$$

   - 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 $\xi_{igk} \in \xi_{ig}$, $k = 1, \ldots, R$:

     $$Q(\hat{\xi}_{igk} \mid \hat{\xi}_{igk}^{(t)}) = E[\log \tilde{p}(y_{ik} \mid \mathbf{u}_i, \hat{z}_{ig} = 1, \hat{\xi}_{igk}) p(\mathbf{u}_i)]. \qquad (3)$$

     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:

$$\begin{aligned}\mathscr{L}(\hat{\xi}_{ig}^{(t+1)}) = \textstyle\sum_{k=1}^{R} & \Big[\log(\sigma(\hat{\xi}_{igk}^{(t+1)})) - \frac{\hat{\xi}_{igk}^{(t+1)}}{2} - \lambda(\hat{\xi}_{igk}^{(t+1)})\hat{\xi}_{igk}^{(t+1)2} + \Big(y_{ik}-\frac{1}{2}\Big)\hat{b}_{gk}^{(t+1)} \\ + & \quad \lambda(\hat{\xi}_{igk}^{(t+1)})\hat{b}_{gk}^{(t+1)2}\Big] + \frac{\log|\hat{\mathbf{C}}_{ig}^{(t+1)}|}{2} + \frac{\hat{\mu}_{ig}^{(t+1)\prime}[\hat{\mathbf{C}}_{ig}^{(t+1)}]^{-1}\hat{\mu}_{ig}^{(t+1)}}{2}, \\ \ell(\hat{\theta}^{(t+1)}) \quad \approx & \quad \sum_{i=1}^{N}\log\Big(\sum_{g=1}^{G}\hat{\eta}_{ig}^{(t+1)}\exp(\mathscr{L}(\hat{\xi}_{ig}^{(t+1)}))\Big).\end{aligned}$$

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:

$$\mathrm{logit}(\eta_{ig}) = \beta_{0g} + \beta_{1g}x_i, \quad g = 2, \dots, G.$$

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.

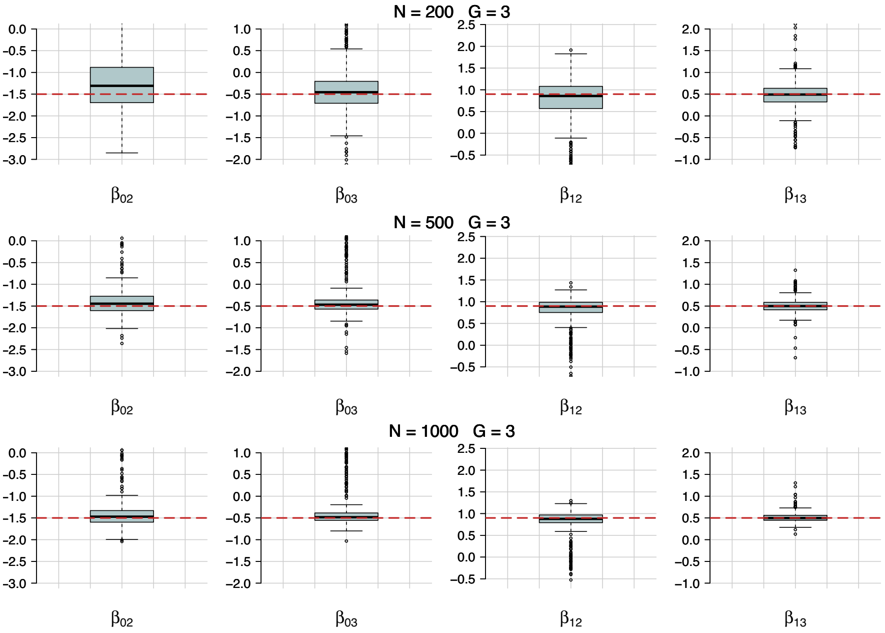


Fig. 1 Distribution across samples of the parameter estimates $\hat{\beta}_g$ for varying N and $G = 3$. The red lines correspond to the true values of the parameters.

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.

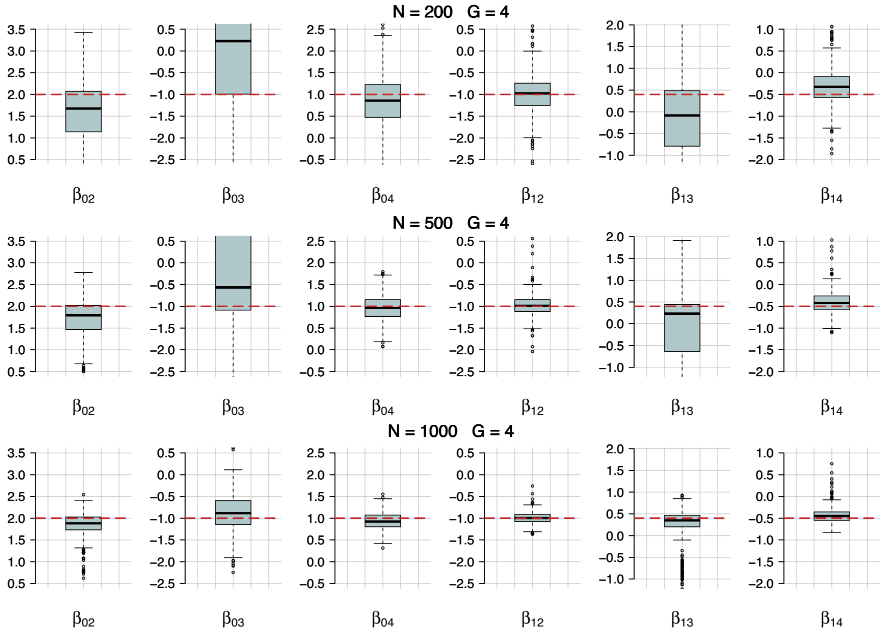


Fig. 2 Distribution across samples of the parameter estimates $\hat{\beta}_g$ for varying N and $G = 4$. The red lines correspond to the true values of the parameters.

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 $j \in \{1,\dots,n\} \setminus \{i\}$, 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

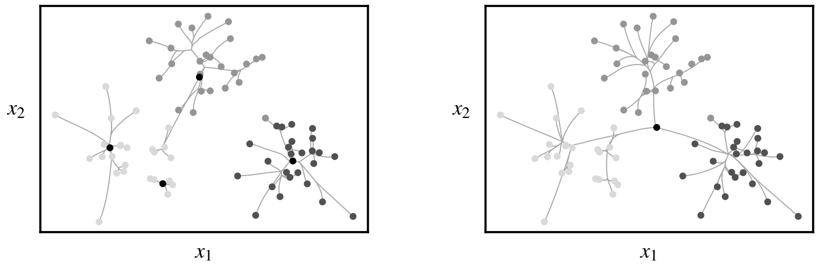
$$w_{ij} = w_{ji} = \begin{cases} \exp\left(-\phi \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\text{mean}_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2}\right) & \text{if } (i,j) \in \mathscr{S} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

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

$$L(\mathbf{M}) = \frac{\|\mathbf{UM} - \mathbf{X}\|_2^2}{2\|\mathbf{X}\|_2^2} + \lambda \frac{\sum_{k<l} (\mathbf{U}^\top \mathbf{W} \mathbf{U})_{kl} \|\mathbf{m}_k - \mathbf{m}_l\|_2}{\|\mathbf{X}\|_2 \sum_{i<j} w_{ij}}, \tag{3}$$

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 $c \ll n$.

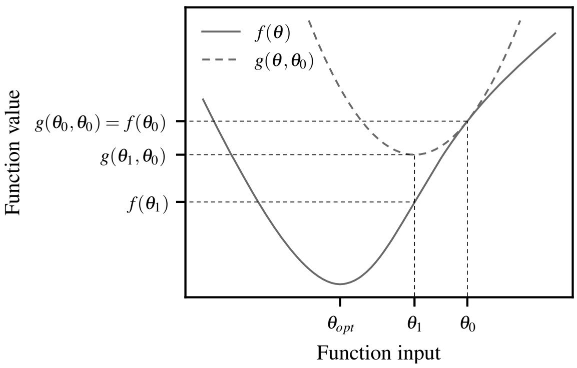


Fig. 2 Example of minimization by an MM algorithm. The majorization function $g(\theta, \theta_0)$ is greater than the target function $f(\theta)$ on the entirety of its domain, except for the supporting point θ_0 where they are equal. Minimizing $g(\theta, \theta_0)$ yields the new supporting point 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

$$f(\theta_1) \leq g(\theta_1, \theta_0) \leq g(\theta_0, \theta_0) = f(\theta_0),$$

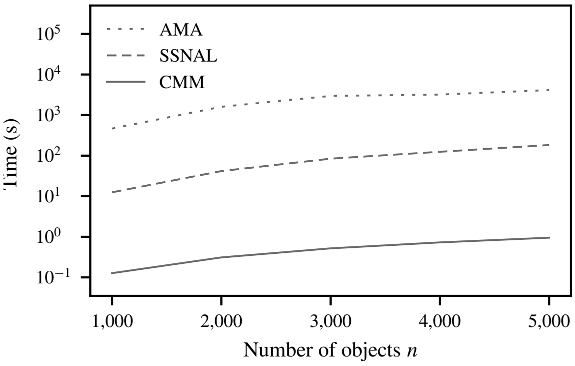
which is also illustrated by Fig. 2. Setting θ_0 to θ_1 and repeating these steps causes the supporting point to converge towards a local minimum. If the target function is convex and coercive, this is the global minimum [1]. For further reading on MM, which is also known as the *concave-convex procedure*, 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 $\lambda \in \{0.0, 0.2, \dots, 110.0\}$ 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).

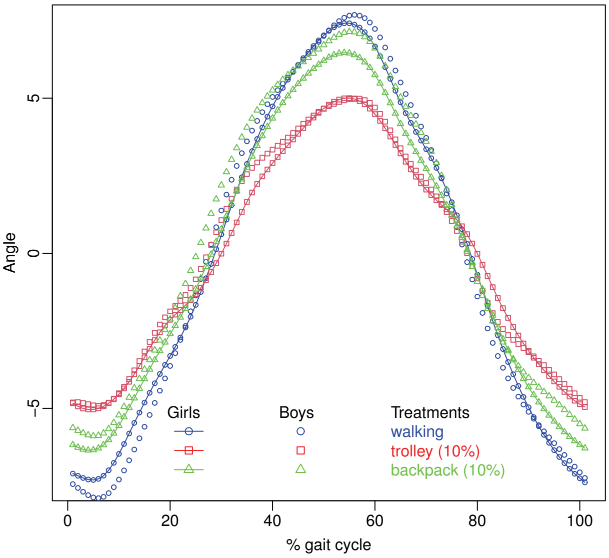


Fig. 1 Sample group mean functions depending on gender and condition.

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.

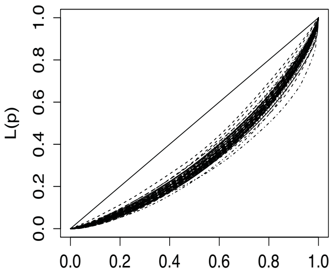


Fig. 1 Each curve illustrates the concentration of family income in a given year (2000, 2006, 2010) and region for a total of 60 empirical LCs.

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 $j \in \{1, \dots, n\}$.

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.

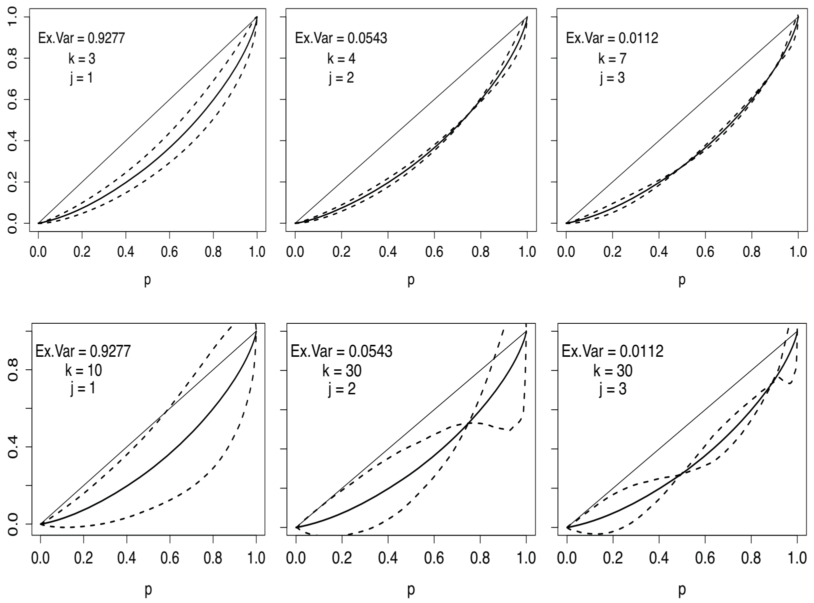


Fig. 2 Fraction of explained variance of the j-th PC, mean curve (solid line) and modes of variation for $j = 1, 2, 3$ and different k (dashed lines) for the sample of original LCs.

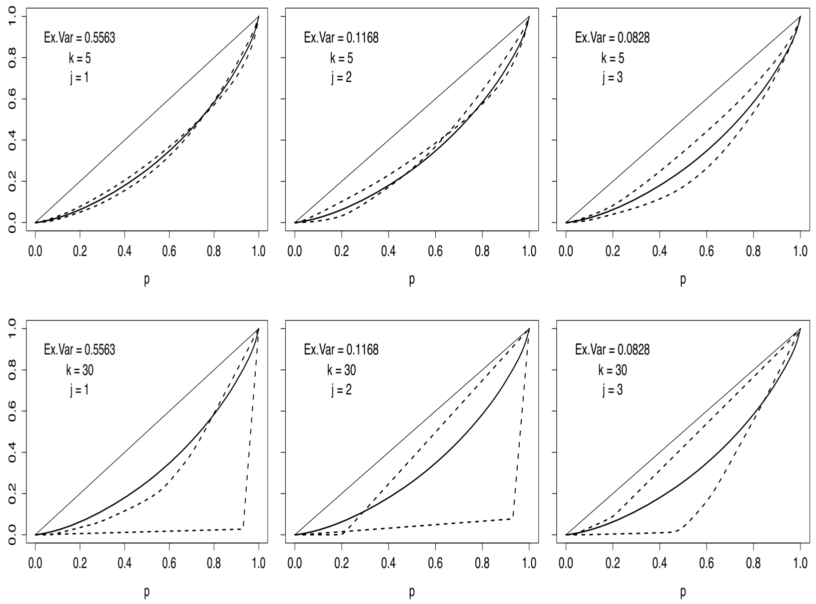


Fig. 3 Fraction of explained variance of the j-th PC, mean curve $\widehat{M}_j(0)$ (solid line) and modes of variation $\widehat{M}_j(k)$ for different values of j and k (dashed lines) for the sample of LCs in Fig. 1.

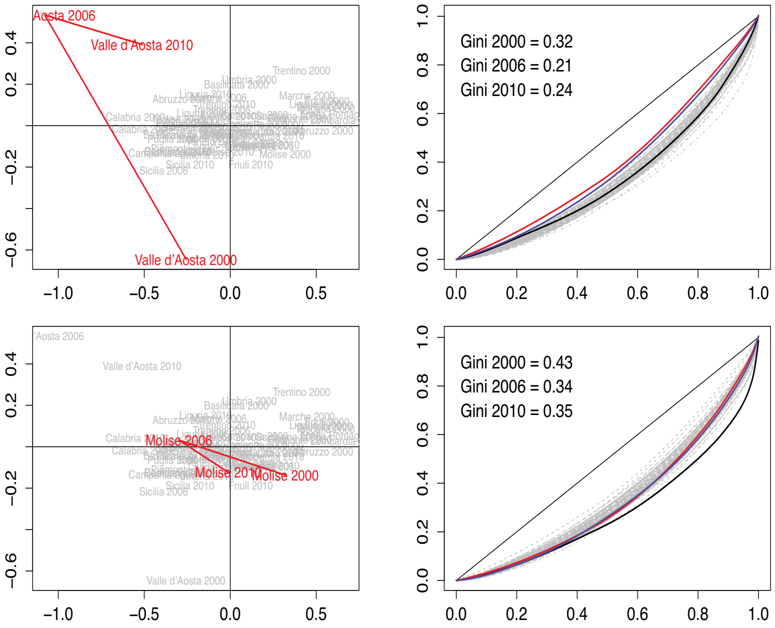


Fig. 4 (Left) Track-plots in the factorial plane of the first two PCs. (Right) Assosiated LCs and Gini indexes.

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.

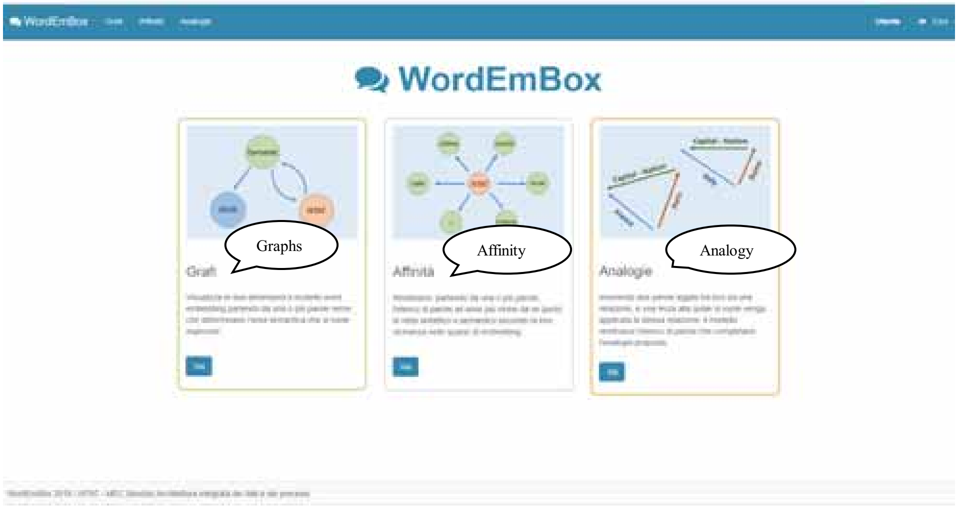


Figure 1: WordEmBox Homepage

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

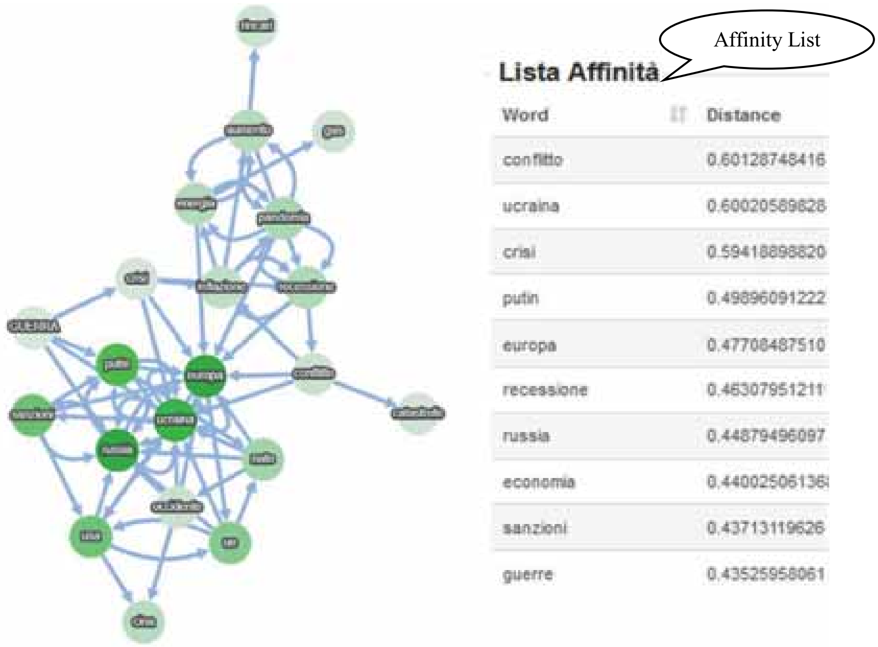
In this paragraph we investigate a sample of one month period tweets collected by using the Istat[1] economic filter consisting of a list of relevant keywords. The dataset is composed by 855,865 tweets. The aim of the following analysis is to understand the impact of the Ukrainian crisis on the economic 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



Affinity List

Lista Affinità

Word	Distance
conflitto	0.60128748416
ucraina	0.60020589828
crisi	0.59418898820
putin	0.49896091222
europa	0.47708487510
recessione	0.46307951211
russia	0.44879496097
economia	0.44002506136
sanzioni	0.43713119626
guerre	0.43525958061

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.

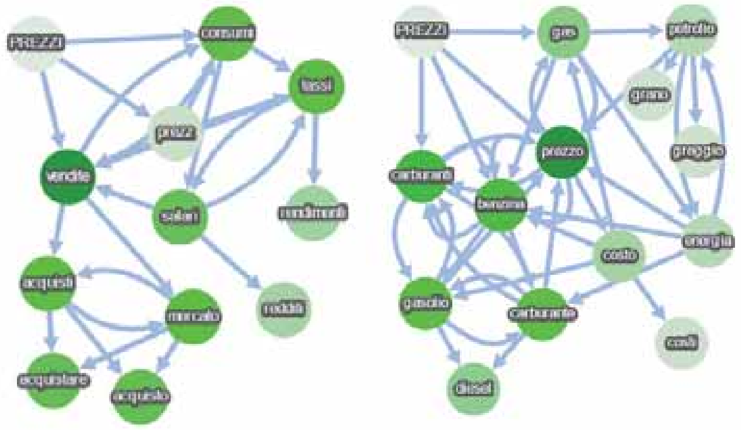


Figure 3: Graph analysis for word “PREZZI” in the two models 2017SME and UkraineWar

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.

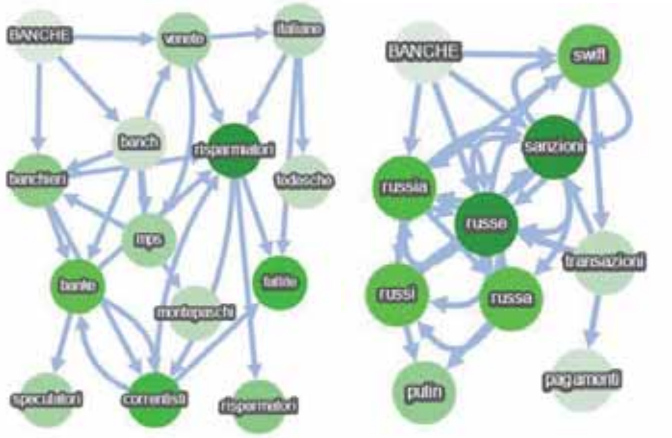


Figure 4: Graph analysis for word "BANCHE" in the two models 2017SME and UkraineWar

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.

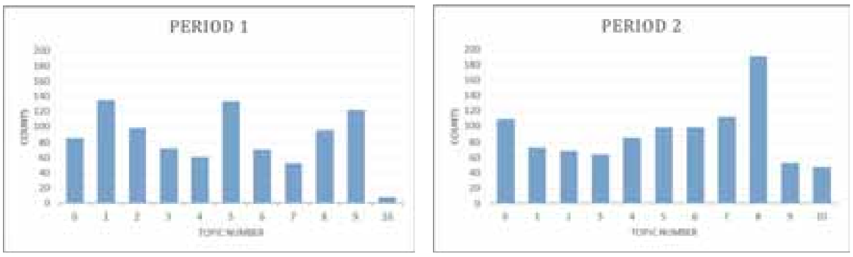


Figure 5: Counts of words in each topic in the two periods

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.

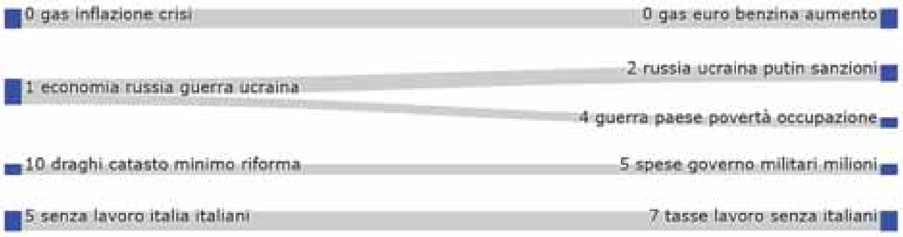

0 gas inflazione crisi
0 gas euro benzina aumento
1 economia russia guerra ucraina
2 russia ucraina putin sanzioni
4 guerra paese povertà occupazione
10 draghi catasto minimo riforma
5 spese governo militari milioni
5 senza lavoro italia italiani
7 tasse lavoro senza italiani

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.

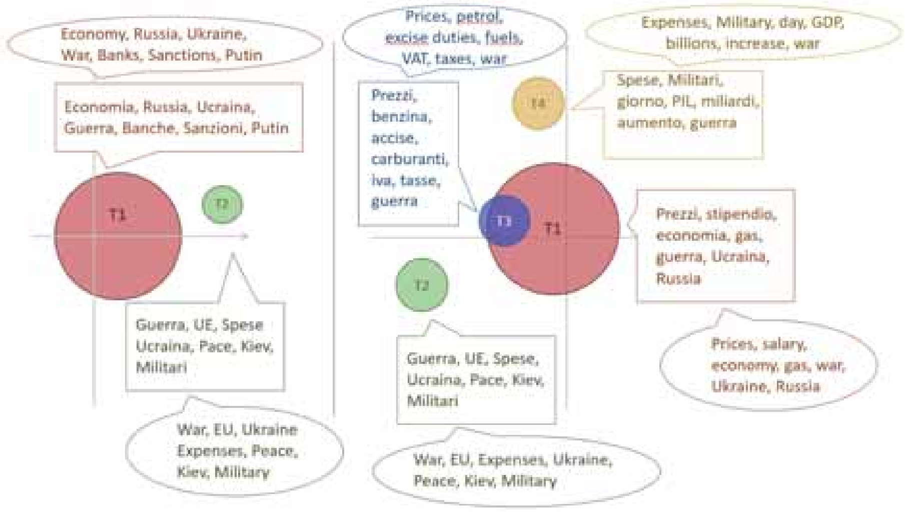

Economy, Russia, Ukraine, War, Banks, Sanctions, Putin
Economia, Russia, Ucraina, Guerra, Banche, Sanzioni, Putin
T1
T2
Guerra, UE, Spese Ucraina, Pace, Kiev, Militari
War, EU, Ukraine Expenses, Peace, Kiev, Military
Prices, petrol, excise duties, fuels, VAT, taxes, war
Prezzi, benzina, accise, carburanti, iva, tasse, guerra
Expenses, Military, day, GDP, billions, increase, war
T4
Spese, Militari, giorno, PIL, miliardi, aumento, guerra
T3
T1
Prezzi, stipendio, economia, gas, guerra, Ucraina, Russia
T2
Prices, salary, economy, gas, war, Ukraine, Russia
Guerra, UE, Spese, Ucraina, Pace, Kiev, Militari
War, EU, Expenses, Ukraine, Peace, Kiev, Military

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:

$$\sqrt{n}\left\{g\left(\overline{w}_{HT}(p)\right)-g\left(\overline{w}(p)\right)\right\}\overrightarrow{d}\mathscr{N}\left(0,\left[g'\left(\overline{w}(p)\right)\right]^2\gamma(p,p')\right),\quad \forall p\in[0,1]. \tag{6}$$

From this result it is possible to derive the pointwise confidence interval as follows:

$$P\left(\overline{f}_{HT}(p)\in\left[\overline{f}^*_{HT}(p)\pm q_\alpha\sqrt{Var\left(\overline{f}^*_{HT}(p)\right)}\right]\right)=1-\alpha,\quad \forall p\in[0,1], \tag{7}$$

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:

$$P\left(\overline{f}_{HT}(p)\in\left[\overline{f}^*_{HT}(p)\pm c_\alpha\sqrt{Var\left(\overline{f}^*_{HT}(p)\right)}\right],\ \forall p\in[0,1]\right)=1-\alpha, \tag{8}$$

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.

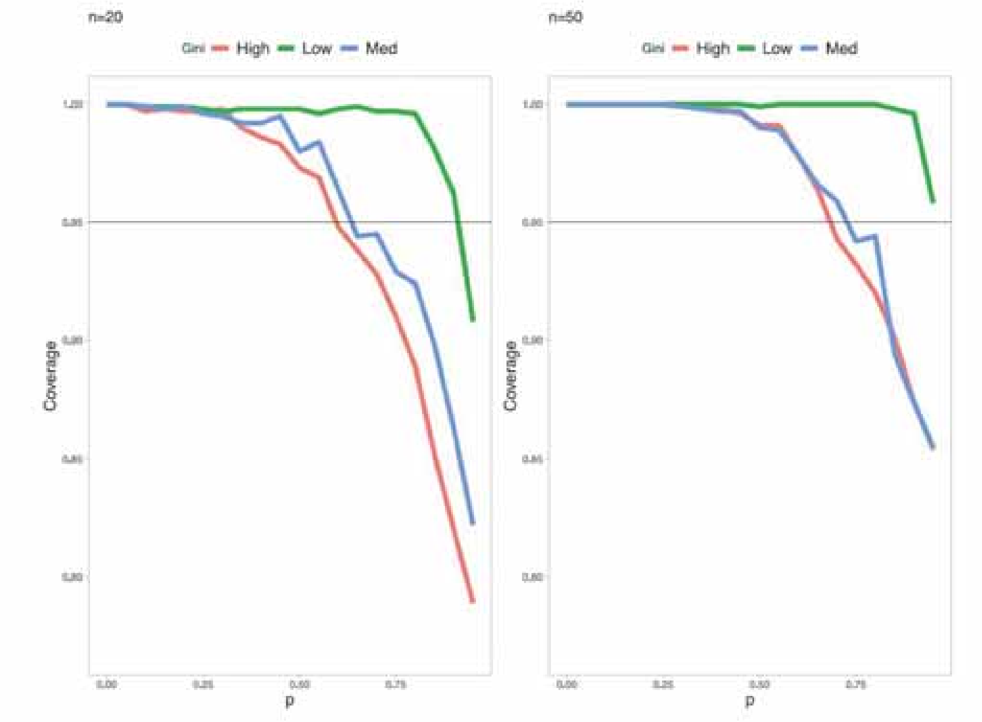


Fig. 1 Pointwise empirical coverage. Nominal level $1-\alpha = 0.95$.

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.

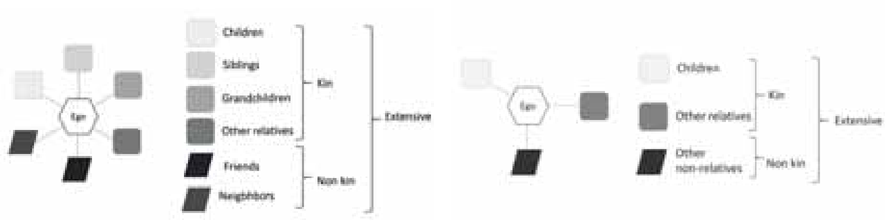
Disregarding contacts with grandparents (in FSS) and with parents (both in FSS and in SHARE Corona 1) because of the age of the target respondents, the potential alters in the ego–network of f-t-f contacts for each elder (ego) are represented by a maximum of six different alter's categories (alter roles) 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.

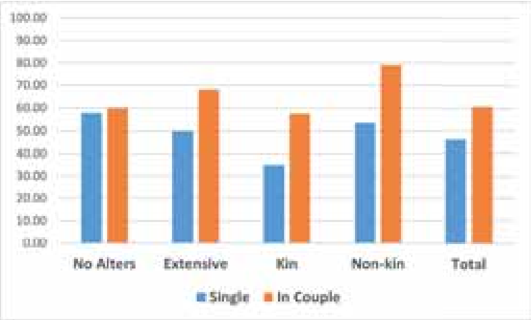
Ties with children have been strengthened also through virtual contacts with daily calls or digital means for more than one single out of two and for two partners in a couple out of three. This result is in line with studies referring to a pre-pandemic period, which highlighted the positive association 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.

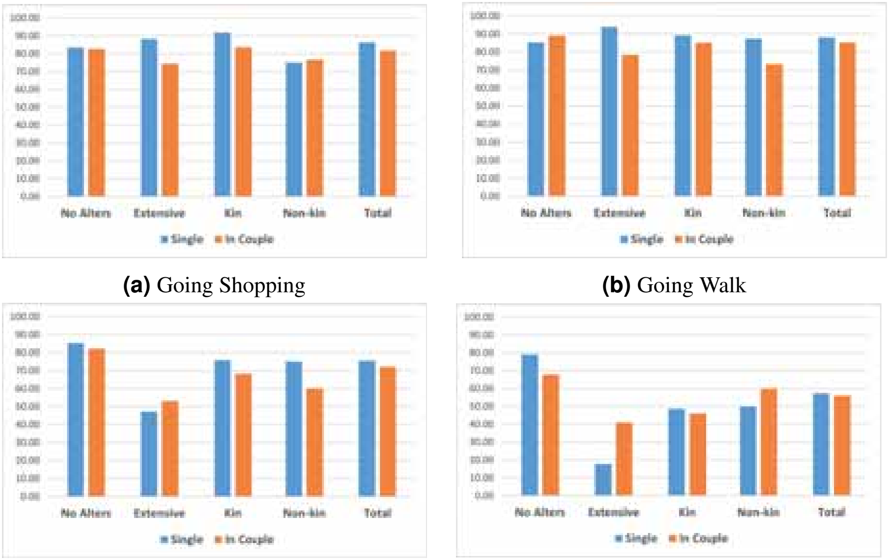
Considering the type of network in which older people were embedded some differences can be noted.

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

Bruno Arpino

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).

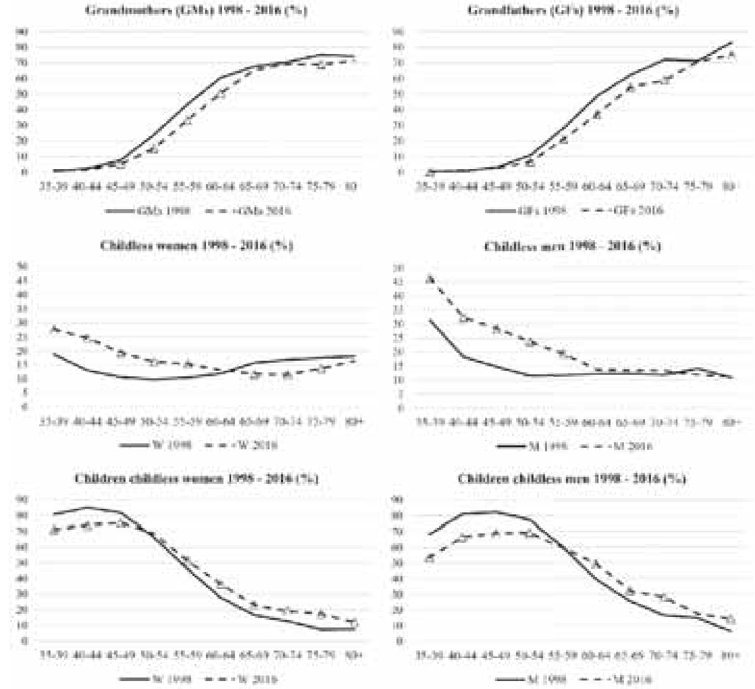


Figure 1: Prevalence of grandparents, childless and children-childless population by age, sex and year[1].

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.

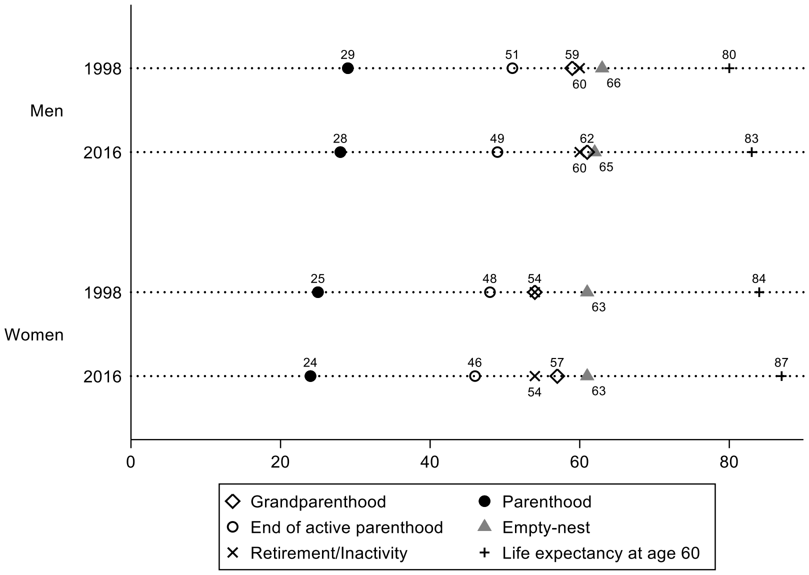


Figure 2: Timing of grandparenthood and other four life transitions, median age by sex and year.

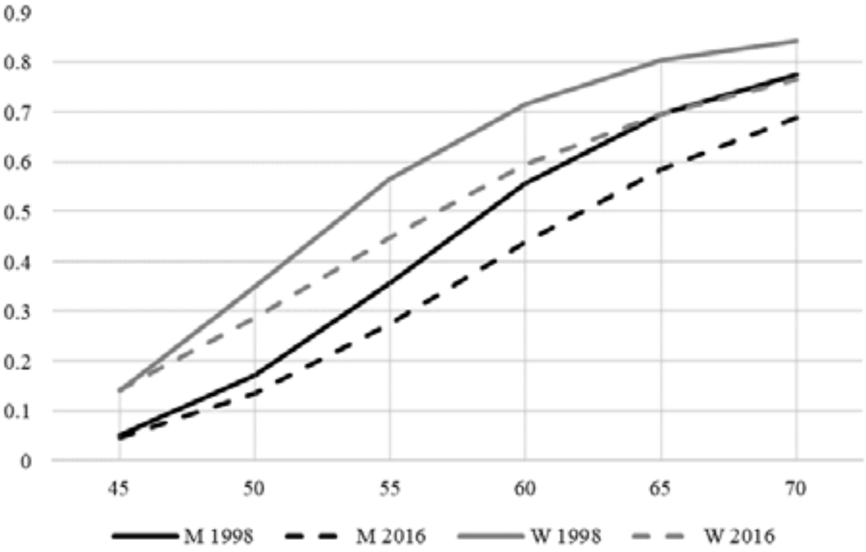


Figure 3: Probability of becoming grandparent at different ages by sex and survey year.

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 $i = 1, \dots, n$ 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 $k = 1, \dots, K$. 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 $k = 1, \dots, K$.

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.

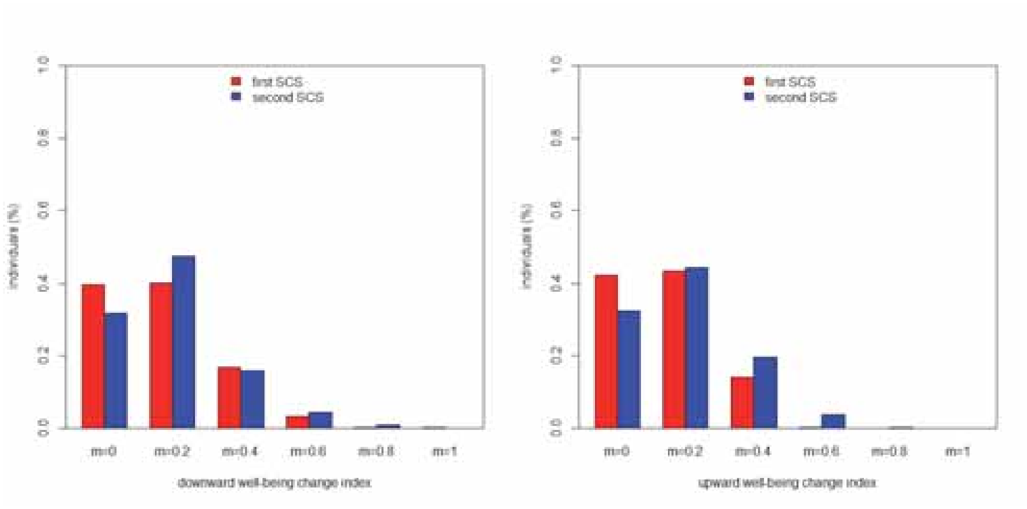


Fig. 1: Percentage of individuals who suffer worsening (left panel) or improvement (right panel) in multidimensional well-being for different index cutoff values: $m = 0$ refers to individuals who worsen in none of the well-being dimensions, $m = 0.2$ to individuals who worsen in one dimension, and so 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 $(e^{\beta} - 1) * 100$ 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.

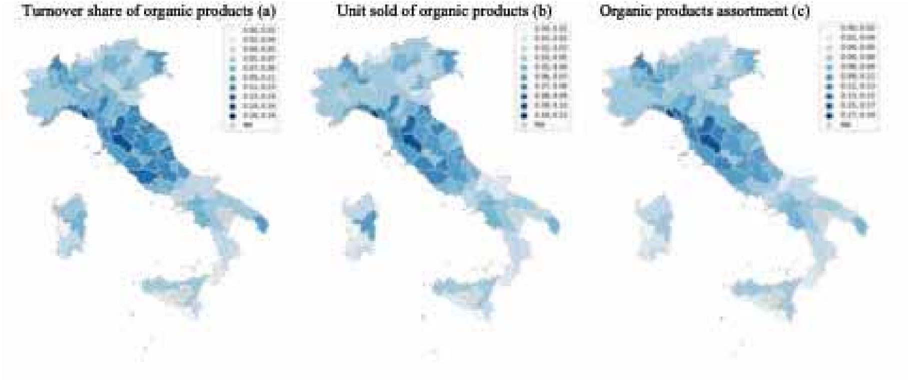

Turnover share of organic products (a)
Unit sold of organic products (b)
Organic products assortment (c)

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.

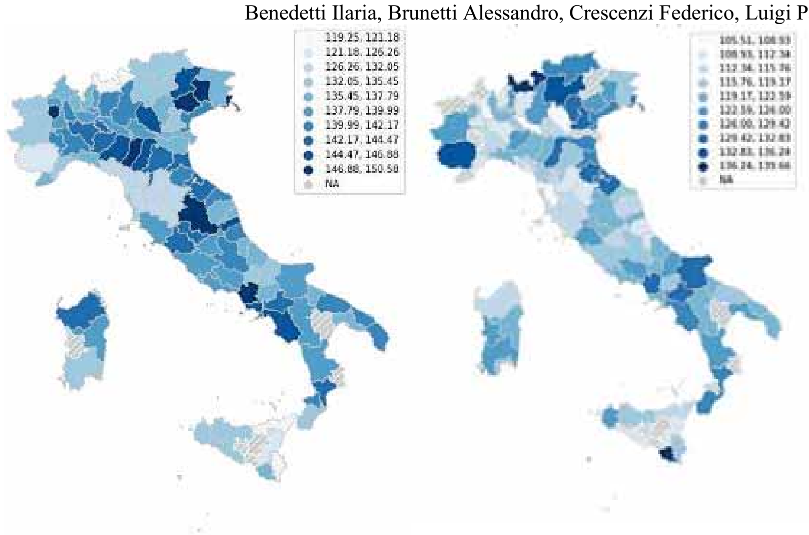


Figure 2: Organic products premium price at provincial level for yoghurt (left) and rice (right).

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 night-time 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

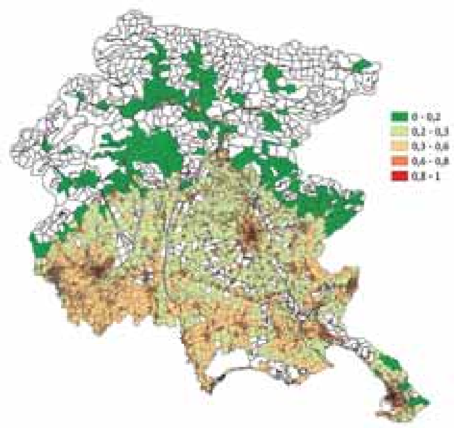


Fig. 1: Heat vulnerability map for FVG

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.

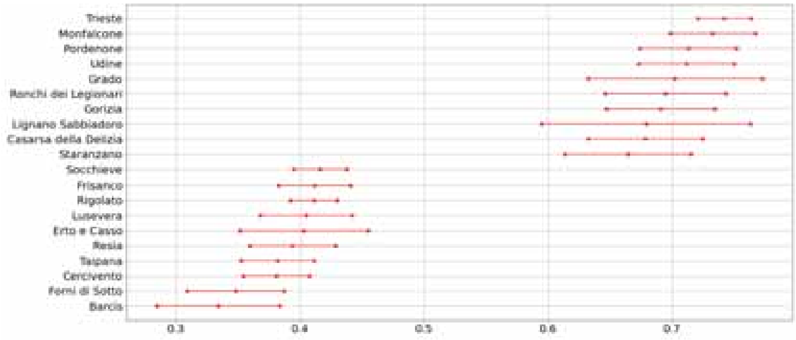


Fig. 2: Performance intervals - Top and bottom 10 municipalities

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)

Understanding School-to-Work Transition in Italy: First Results from a Newly Integrated Database

Le determinanti della transizione scuola-lavoro in Italia: primi risultati da una nuova base dati integrata

Stefano De Santis, Anita Guelfi, Valentina Talucci e Francesco Maria Truglia (Istat)

Abstract This paper aims at analysing the main drivers of school-to-work transition in Italy in the last decade - a topic often at the centre stage of both the policy and research debate for its key social and economic relevance. To this aim a new and innovative dataset was recently put together at the Italian National Institute of Statistics by integrating several different administrative data sources, thereby allowing to follow the complex path of individuals along their education and labour market choices, taking into account, on the one hand, the individual's socio-demographic characteristics and education choices, on the other – for those to appear to be employed – their employment characteristics, as well as those of the corresponding employer.

Abstract *Il paper mira ad analizzare le principali determinanti sottostanti la delicata transizione che si osserva per i giovani tra la fase formativa e l'entrata nel mondo del lavoro: un tema frequentemente al centro del dibattito scientifico e politico ma sui cui esiti ha spesso pesato la carenza di informazioni sufficientemente dettagliate e disponibili in forma longitudinale. A tale scopo si presentano le potenzialità analitiche di una nuova base dati, frutto dell'integrazione di diversi archivi amministrativi, grazie alla quale è possibile analizzare in modo nuovo e più efficace tematiche rilevanti nel campo economico e sociale.*

Key words: School-to-work transition, social mobility, integrated administrative data, geo-referenced data

[1] Stefano De Santis, Istat; stdesan@istat.it:
Anita Guelfi, Istat, anita.guelfi@istat.it:
Valentina Talucci, Istat; talucci@istat.it:
Francesco Maria Truglia, Istat; truglia@istat.it:

Summary

School-to-work transition represents a key time-span in people's life cycle, The long and complex process leading individuals from educational to employment choices turns out indeed to deeply affect their future social and economic status and wellbeing. These choices are in turn the result of the interaction of many different concurrent variables, which must therefore be taken into account in order to truly understand the underlying forces. In this respect, our paper aims at offering some new insight in that it exploits a new and innovative dataset which, being the result of the integration of several different administrative data sources, allows to analyse the determinants of school-to-work transition taking into account, on the one hand, the individual's socio-demographic characteristics and educational choices, on the other – for those to appear to be employed – their employment characteristics, as well as those of the corresponding employer. In particular, individual data could also be linked to the information on the cultural and economic background of each person's family of origin, making it thus possible to tackle the issue of social mobility in school-to-work transition.

The extensive use of administrative data treated for statistical purposes and integrated with survey data offered indeed the opportunity to build a new integrated database providing a detailed and multi-dimensional mapping of individuals, families and firms. In particular, to study the determinants and main characteristics of school-to-work transition, we chose to focus on a specific cohort of people, namely those who were born in 1992, whom we were able to follow starting from the year 2011 (when, turning 19, they were supposed to have completed or be about to complete secondary school and eventually choose to enrol to tertiary education) to 2019, the last available year for most of the selected data sources so far.

This selection led us to the observation of a universe of about 2 million individuals, consisting of all people belonging to the family unit of our target group (i.e., about 600,000 persons born in 1992).

Our aim is analyzing both the educational and professional choices made by this cohort of young people during the 9 years following the highest educational degree obtained in 2011. To this aim, we focused on the situation we were able to observe in 2019 in terms of their condition on the labor market, taking into account their previous educational choices, as well as the interaction with their socio-demographic characteristics, including their family cultural and economic background in order to shed some light on the issue of social mobility.

Focusing first on educational attainment, data show that in 2019 about 48.1% of our 1992 cohort completed at most upper secondary school, followed by 26.8% with a tertiary degree (3-year degree: 14,2%; 5-year degree: 12,6%), and 18,2% not going beyond lower secondary education. In this context, women tend to achieve more often a tertiary degree compared with men (+13.1 p.p.), though there appear to be a high degree of segmentation in terms of specific areas of study, with females more often enrolled in areas of study such as teaching (+7.2 p.p.), medicine (+4.5 p.p.) and socio-political sciences (+4.3 p.p.), whereas men are more often found in areas such as statistics & economics (+7.7 p.p.) as well as engineering (+8.6 p.p.).

In this respect, the final educational outcome of our selected cohort of youths looked strongly correlated with their families' cultural background. Looking at graduates having completed a second-level (5-years) tertiary degree, data show that about one third (33,8%) come from families with at least one parent of same educational attainment, while 22,5% have at least one parent having completed a first-level (3 years) tertiary degree; the share of graduates whose parents did not go beyond lower secondary school amounted instead to only 4.8%, while the remaining 38,9% come from families where at least one parent completed high-school. Low mobility is confirmed at the other end of the educational distribution with parents with low educational attainment increasing the probability for their children to achieve a low educational degree.

The employment picture we observed in 2019 for our target group allowed us to zoom into the specific path of school-to-work transition followed by each member of the 1992 cohort. Data show indeed that in 2019, about 61,4% of them turn out to be actually employed, while 7,1% is both studying and working, and 6,9% are full-time students. As for the remaining 24,5% who is neither studying, nor working (the so-called NEETs), most of them (22,4%) had at least one job experience between 2012 and 2018. The probability of being employed turns out to be 6.6 percentage points higher for males as compared with females; this reflects a significantly higher permanence of girls in the education system as a full-time student but also a higher female incidence among the NEETs with at least one past(and now interrupted) job experience (23,9% as compared with 21,0% for males). Breaking down data by geographical area, data confirm how youths tend to enter relatively more easily the labor market in the Northern regions.

Geo-referenced residence data also allowed us to analyze education- and work-related mobility across the country. In this respect, the evidence shows that: (i) education-related average trajectories mainly develop from the Central regions towards the Northern ones, with no major differences by gender, except for the largest flows (i.e. over 500 individuals), for which women tend to move mainly from the Eats to the West (mainly towards Rome); (ii) job-related average trajectories are mainly positioned in the Central region from east to west for low-intensity flows (<101 individuals), while higher density flows mainly show a North direction.

Using a random forest methodology, we finally analyzed the link between family background, education choices and first job experiences. Results emphasize employment experience (i.e. being employed for at least 4 years in the selected period) and family background as the main determinants of the employment outcome in 2019.

Citations and References

1. Breiman L., Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324.
2. Cressie,Noel. Statistics for Spatial Data, Revised Edition, Wiley,2015 e Mitchell, Andy. The ESRI Guide to GIS Analysis, Volume 2. ESRI Press, 2005.

3. Istituto Nazionale di Statistica - Istat. 2020. “Livelli di istruzione e ritorni occupazionali. Anno 2019”. Statistiche Report. Roma: Istat. https://www.istat.it/it/archivio/245736.
4. Istituto Nazionale di Statistica - Istat. 2018. “Risorse, regolarità degli studi e mobilità nel sistema universitario”. In “Rapporto sulla Conoscenza 2018. Economia e Società”. Capitolo 6: 108-109. Letture Statistiche - Temi. Roma: Istat. https://www4.istat.it/it/archivio/209513.
5. Istituto Nazionale di Statistica - Istat. 2012. “Le disparità nei percorsi formativi e lavorativi”. In Rapporto Annuale 2012. La situazione del Paese. Capitolo 4: 247-256. Roma: Istat. https://www.istat.it/it/archivio/61203.
6. Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

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

$$X \approx \mu + \sum_{j=1}^{d} \xi_j v_j \qquad\qquad \mathbb{E}\left[\xi_j \xi_k\right] = \lambda_j \delta_{jk} \tag{1}$$

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 nonparametric 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:

$$C_d\,(x,h) = \mathbb{E}\left[\left(\left(1 - \frac{\|\Pi_d^{\perp}(X-x)\|^2}{h^2}\right)\mathbb{I}_{\left\{\|\Pi_d^{\perp}(X-x)\|^2 \le h^2\right\}}\right)^{d/2}\middle|\, \Pi_d x\right] \tag{3}$$

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. $((1 - \|\Pi_d^{\perp}\,(X-x)\,\|^2/h^2)\mathbb{I}_{\left\{\|\Pi_d^{\perp}(X-x)\|^2 \le h^2\right\}})^{d/2}$ 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

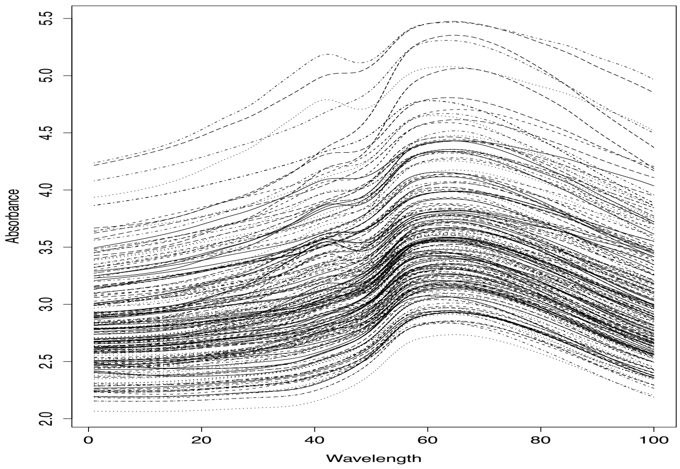


Fig. 1 The Tecator dataset.

$$\widehat{C}_{d,n}(x,h) = \sum_{i=1}^{n}\left(\left(1-\frac{\|\widehat{\Pi}_d^{\perp}(X_i-x)\|^2}{h^2}\right)\mathbb{I}_{\left\{\|\widehat{\Pi}_d^{\perp}(X_i-x)\|^2\le h^2\right\}}\right)^{d/2}\times$$
$$\times\frac{K(\|\widehat{\Pi}_d(X_i-x)\|/b)}{\sum_j K(\|\widehat{\Pi}_d(X_j-x)\|/b)}$$

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 $\left\{\chi_j, j=1,\ldots,N\right\}$, 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).

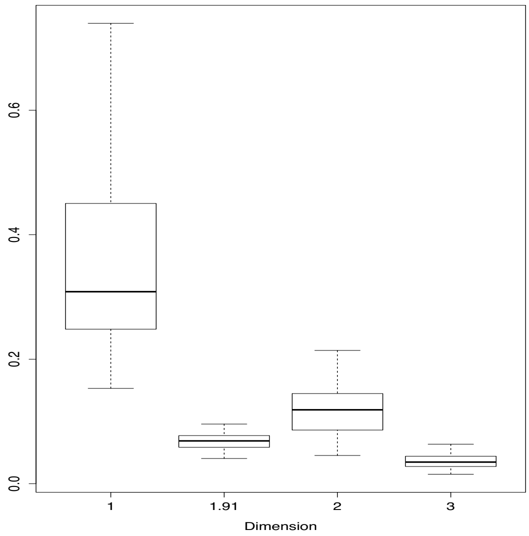


Fig. 2 Empirical distributions of the means of the ISEs, varying the dimension. The second (from the left) boxplot corresponds to the ISE computed when local dimensions are used.

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 $d = 1, \dots, 5$, whereas the remaining part $\left\{\chi_j, j = 1, \dots, 15\right\}$ 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):

$$\operatorname{vec}(\boldsymbol{\Lambda}_j) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad \operatorname{vec}(\boldsymbol{E}_j) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}), \quad j = 1, \dots, J$$

with 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:

$$\operatorname{vec}(\boldsymbol{Y}_j) \sim N\left((\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right), \quad j = 1, \dots, J.$$

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:

$$\begin{aligned}\ell_{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \log\phi\left(\operatorname{vec}(\boldsymbol{Y}_j), (\boldsymbol{I}_r \otimes \boldsymbol{X}_j)\operatorname{vec}(\boldsymbol{B}), (\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)\boldsymbol{\Psi}(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j}\right) + \\ -\lambda\left[(1-\alpha)\sum_{c=1}^{r}\sum_{l=2}^{p} b_{lc}^2 + \alpha\sum_{l=2}^{p} ||\boldsymbol{b}_{l.}||_2\right],\end{aligned} \tag{2}$$

where b_{lc} and $\boldsymbol{b}_{l.}$ 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

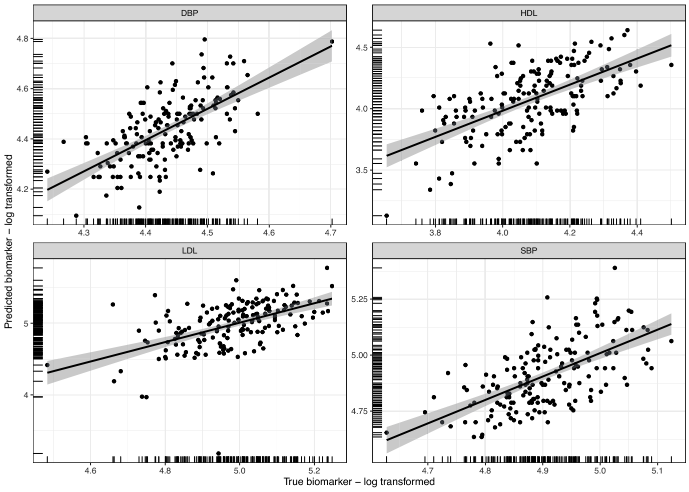


Fig. 1 Observed vs fitted scatterplots for the estimated biomarkers, namely log-transformed Diastolic Blood Pressure (DBP), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL) and Systolic Blood Pressure (SBP), EPIC Italy test set. Linearly smoothed conditional means and associated standard 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%).

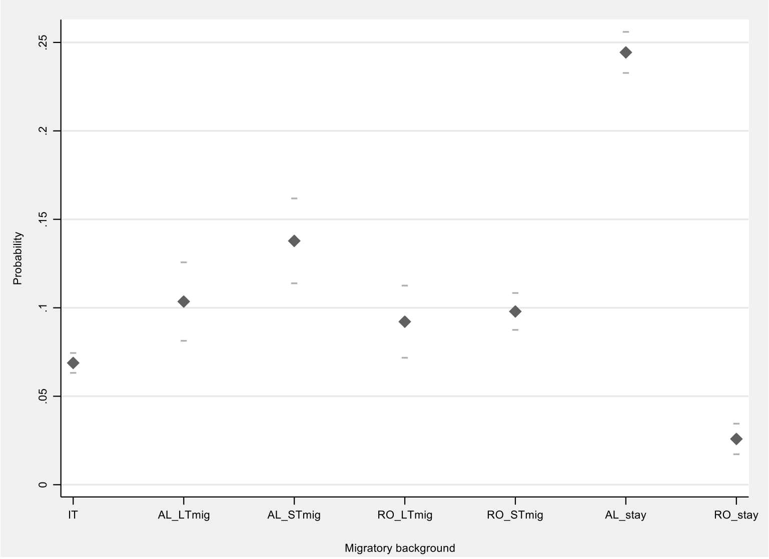


Figure 1: *Adjusted predictions for being intended to have a (another) child by migratory background (95% confidence intervals) while holding control variables at their mean values.*

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.

Does self-employment in the origin-country affect self-employment after migration?

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 $\mathbf{y}_N = (y_1, \dots, y_N)$.

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 $(D_1, \dots, D_N)$. 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 ($i = 1, \dots, N$). 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 $X_1^d, \dots, X_k^d$ 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 $\mathbf{y_s}$ the y-values corresponding to sampled units. In symbols: $\mathbf{y_s} = \{y_i;\ i \in \mathbf{s}\}$. *Sample data* are composed by the pair $(\mathbf{s}, \mathbf{y_s})$, namely by the observed y-values together with the corresponding (sampled) units. The knowledge of $(\mathbf{s}, \mathbf{y_s})$ 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, $\mathbf{Y_s}$ 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* $X_1, \dots, X_p$ 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, $\mathbf{X_s}$ 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 $X_1, \dots, X_p$, 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 $X_1, \dots, X_p$, 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:

$$f(\mathbf{s}, \mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta})p(\mathbf{s}|\mathbf{X}_N, \mathbf{y}_N; \boldsymbol{\psi}). \tag{1}$$

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

$$f(\mathbf{s}, \mathbf{y}_N|\mathbf{X}_N; \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_N|\mathbf{X}_N^d; \boldsymbol{\theta})p(\mathbf{s}|\mathbf{X}_N^d). \tag{2}$$

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

$$\bar{y}_\mathbf{s} - \bar{y}_N = \sqrt{f_N^{-1} - 1}\sqrt{S_{yN}^2}\,Corr_N(y, D)$$

where

$$\begin{aligned}
S_{yN}^2 &= \frac{1}{N-1} = \sum_{i=1}^{N} (y_i - \overline{y}_N)^2 \ \text{(finite population variance of } y_i\text{s)} \\
S_{DN}^2 &= \frac{1}{N-1} = \sum_{i=1}^{N} (D_i - f_N)^2 \ \text{(finite population variance of } D_i\text{s)} \\
Corr_N(y, D) &= \frac{1}{N-1} \sum_{i=1}^{N} y_i (D_i - f_N) \bigg/ (S_{yN} S_{DN}) \ \text{(finite population correlation of } y_i\text{s and } D_i\text{s)}.
\end{aligned}$$

Consider next simple random sampling without replacement (SRS, for short). As well known, the variance of the sample mean is

$$V_{SRS}(\overline{y}_{\mathbf{s}}) = \frac{1-f_N}{f_N} S_{yN}^2 / N$$

and hence, if $MSE_{NP}(\overline{y}_{\mathbf{s}}) = E_{NP}[(\overline{y}_{\mathbf{s}} - \overline{y}_N)^2]$ denotes the Mean Squared Error w.r.t. the non-probabilistic design actually used, we get

$$\frac{MSE_{NP}(\overline{y}_{\mathbf{s}})}{V_{SRS}(\overline{y}_{\mathbf{s}})} = N E_{NP}[Corr_N(y, D)^2] = N D_I. \tag{3}$$

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 $\overline{y}_{\mathbf{s}}$. Under non-probabilistic sampling, D_I is generally $O(1)$, and hence $\frac{MSE_{NP}(\overline{y}_{\mathbf{s}})}{V_{SRS}(\overline{y}_{\mathbf{s}})}$ 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

$$T_H = \sum_{i=1}^{N} w_i D_i y_i \bigg/ \sum_{i=1}^{N} w_i D_i.$$

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 important 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 $X_1, \dots, X_m$ 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 $\mathbf{y_s}$ the vector of sample units, and by $\mathbf{X_s}$ 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_\mathbf{s}^T \mathbf{X_s})^{-1} \mathbf{X}_\mathbf{s}^T \mathbf{y_s}$, 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 $\mathbf{Y_s}$ given $\mathbf{X_s}$ and $\mathbf{s}$ coincides with the conditional distribution of $\mathbf{Y_s}$ given $\mathbf{X_s}$. 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 $X_1, \dots, X_p$ 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 $\mathbf{s}_c$, $\mathbf{s}_c^*$; $c = 1, \dots, C$, 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

$$f_c = \frac{\sum_{i \in \mathbf{s}_c^*} w_i^* \big/ \sum_{i \in \mathbf{s}^*} w_i^*}{\sum_{i \in \mathbf{s}_c} w_i \big/ \sum_{i \in \mathbf{s}} w_i},\ \ c = 1, \dots, C$$

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 $\widetilde{y}_i^P = g(y_{j_1}^{NP}, \ldots, y_{j_k}^{NP})$. Estimation of finite population parameters (such as $\bar{y}_N = \sum_i y_i/N$), or superpopulation parameters, is based on imputed values $\widetilde{y}_i^P$, $i \in \mathbf{s}^*$. 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 under-coverage.

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

$$l(\boldsymbol{\lambda}) = \sum_{i \in \mathbf{s}^*} \frac{1}{\pi_i} \left\{ D_i \log p_i(\boldsymbol{\lambda}) + (1 - D_i) \log(1 - p_i(\boldsymbol{\lambda})) \right\}.$$

If $\widehat{\boldsymbol{\lambda}}$ denotes the maximum pseudo-likelihood estimator of $\boldsymbol{\lambda}$, the population mean $\bar{y}_N$ 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

$$\hat{\hat{Y}} = \sum_{k \in s_V} \frac{y_k}{n_v} \tag{1}$$

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:

$$E(\hat{\hat{Y}}) = \sum_{k \in U_V} \frac{y_k \pi_i^v}{\bar{\pi}^v}$$

where $\bar{\pi}^v$ is the mean of all response propensities. The bias of this estimator is:

$$B(\hat{\hat{Y}}) = E(\hat{\hat{Y}}) - \overline{Y} = \frac{C_{\phi y}}{\bar{\pi}^v} \frac{N_v}{N} (\overline{Y}_v - \overline{Y}_{NV})$$

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, **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 **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:

$$\hat{\theta}_{NPSAS2} = \sum_g \sum_{k\in s_{V\,g}} a_k y_k / \bar{\pi}_g. \tag{10}$$

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:

$$\hat{\theta}_{NKW} = \sum_{k \in s_V} a_k y_k w_k^{KW} \tag{13}$$

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.

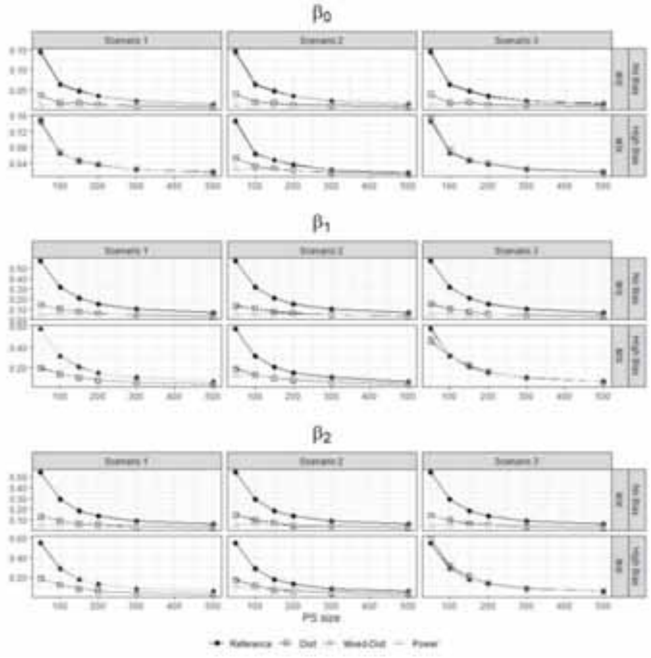


Figure 1: Median MSEs for regression coefficients over 100 iterations in alternative scenarios

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 $\mathcal{Y}$. 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 $\mathcal{Y}$ 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,

$$\hat{Y}_{PC,B}^{P} = \sum_U \delta_k \tilde{y}_k w_k, \tag{4.4}$$

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

$$\hat{Y}_{PC,B}^{D} = \hat{Y}_{PC,B}^{P} + \sum_{s \subset U} \delta_k (y_k - \tilde{y}_k) f_k, \tag{4.5}$$

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

$$\tilde{\mathbf{a}}_k = \left[1_{\left\{\tau_{h-1<\xi(\mathbf{z}_k)\le\tau_h}\right\}}\right]_{h=1}^{H}$$

where the $\{\tau_h\}_{h=0}^{H}$ are known breakpoints; *iii*) $\tilde{y}_k$=$\tilde{\mathbf{a}}'_k\widehat{\boldsymbol{B}}$ is computed, where $\widehat{\boldsymbol{B}}' = \left(\frac{N_1}{\widehat{N}_1}, \dots, \frac{N_h}{\widehat{N}_h}, \dots, \frac{N_H}{\widehat{N}_H}\right)$ 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 $\mathbf{X}_{(\boldsymbol{d})} = \sum_U \mathbf{x}_k\, \lambda_{k(d)}$ 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 $\mathbf{X}_{(\boldsymbol{d})}$, $N_{B(d)}$ 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 $\mathbf{X}_{(\boldsymbol{d})}$, $N_{B(d)}$ 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 $\mathbf{X}_{(\boldsymbol{d})}$, $N_{B(d)}$, $\sum_{U_B} \tilde{y}_k\lambda_{k(d)}$ 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).

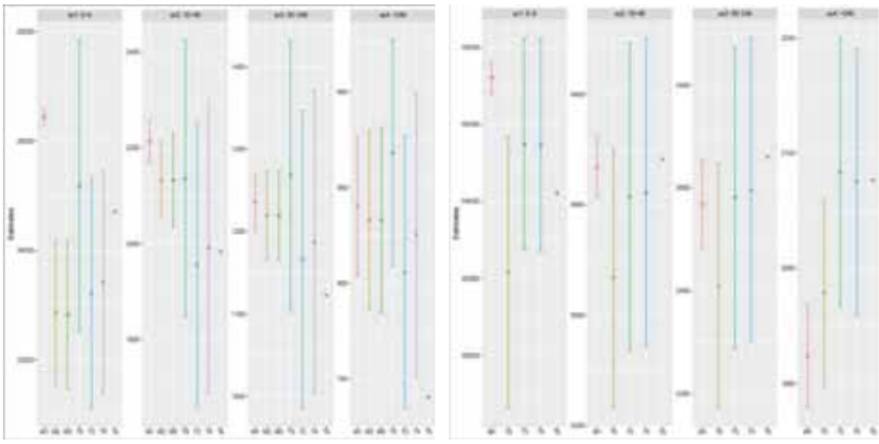


Figure 6.2: Estimator CIs (95%) by macro-regions for WEBORD total (right) and WEBSM total (left).

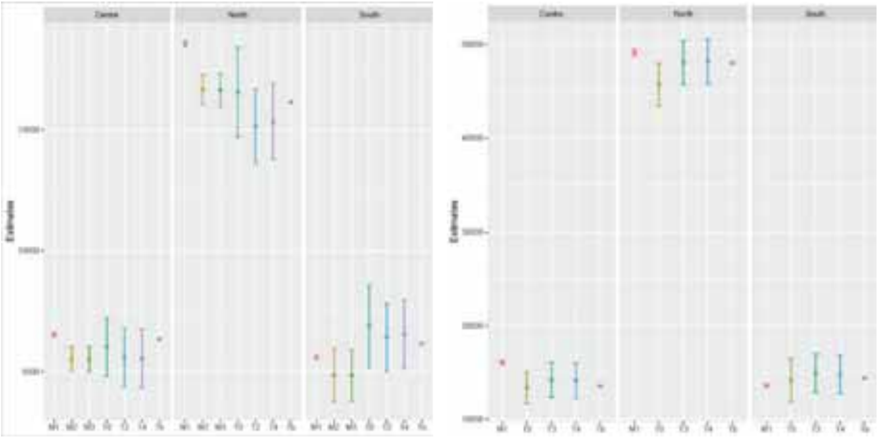


Table 6.2: CV of the estimators for size classes by macro region domain of WEBORD total

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						

Centre	10,962 *(17.6%)*	10,793 *(17.8%)*	10,392 *(17.4%)*	10,500 *(18.1%)*	11,014 *(18.3%)*	11,629 *(18.2%)*
South	12,746 *(20.5%)*	12,975 *(21.4%)*	12,682 *(21.2%)*	12,015 *(20.7%)*	12,685 *(21.1%)*	13,429 *(20.9%)*
	17,659 *(28.4%)*	16,993 *(28%)*	16,557 *(27.8%)*	16,010 *(27.6%)*	16,043 *(26.8%)*	17,062 *(26.6%)*
Island	6,308 *(10.1%)*	5,793 *(9.6%)*	6,076 *(10.2%)*	5,819 *(10.1%)*	6,042 *(10.1%)*	6,768 *(10.6%)*
High School						
Classical	5,527 *(8.9%)*	5,463 *(9%)*	5,318 *(9%)*	4,779 *(8.2%)*	5,138 *(8.6%)*	5,117 *(8%)*
Scientific	33,659 *(54%)*	33,710 *(55.6%)*	33,622 *(56.3%)*	31,936 *(55.1%)*	33,183 *(55.3%)*	36,230 *(56.6%)*
Foreign Language	1,233 *(2%)*	1,182 *(1.9%)*	1,148 *(1.9%)*	1,077 *(1.9%)*	1,207 *(2%)*	1,677 *(2.6%)*
Technical Institute	11,036 *(17.7%)*	10,325 *(17%)*	10,319 *(17.3%)*	10,897 *(18.8%)*	11,459 *(19.1%)*	12,016 *(18.8%)*
Professional Institute	5,893 *(9.5%)*	5,332 *(8.8%)*	4,877 *(8.2%)*	5,136 *(8.8%)*	5,289 *(8.8%)*	5,001 *(7.8%)*
Other	3,922 *(6.3%)*	3,607 *(6%)*	3,420 *(5.7%)*	3,353 *(5.8%)*	3,185 *(5.3%)*	3,384 *(5.3%)*
Missing	988 *(1.6%)*	1,021 *(1.7%)*	984 *(1.6%)*	802 *(1.4%)*	528 *(0.9%)*	628 *(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.*

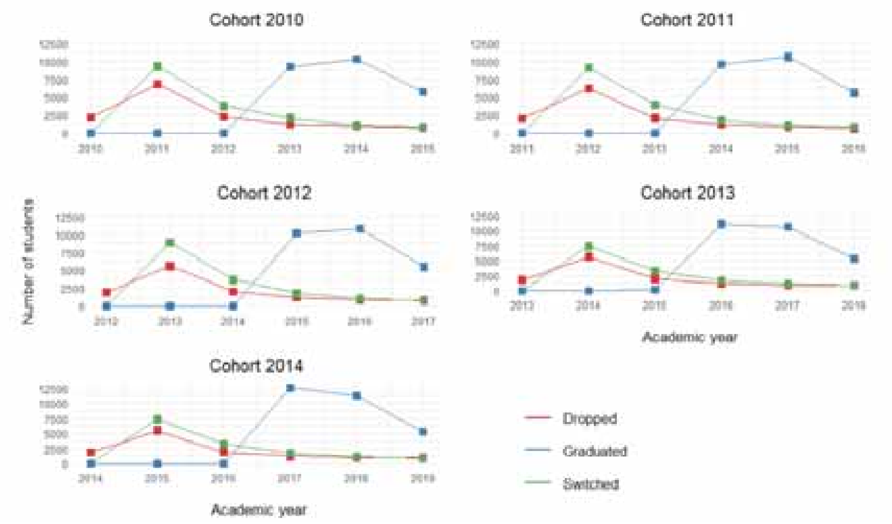


Table 2 shows the relative risk ratios for the estimation of the multinomial logistic regression. As for individual characteristics, the relative risk of dropping out over graduation is lower for women than for men, whereas the opposite is true for course switches. Italian students have a relative risk 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*** (0.303)	46.998*** (2.333)	70.927*** (3.544)

Female	0.862***	1.050***	0.994
	(0.017)	(0.013)	(0.012)
Italian citizenship	0.629***	0.667***	0.546***
	(0.028)	(0.024)	(0.020)
High school final mark (ref: 60 – 69)			
70 – 79	0.514***	0.620***	0.650***
	(0.011)	(0. 011)	(0.012)
80 – 89	0.275***	0.386***	0.436***
	(0.007)	(0. 007)	(0.008)
90 – 99	0.166***	0.251***	0.313***
	(0.005)	(0. 005)	(0.006)
100 and 100 cum laude	0.091***	0.167***	0.215***
	(0.004)	(0. 004)	(0.005)
Missing	0.283***	0.189***	0.276***
	(0.018)	(0. 010)	(0.014)
Credits (ref: 0 – 24)			
25 – 40	0.016***	0.115***	0.169***
	(0.001)	(0. 003)	(0.004)
41 – 56	0.002***	0.052***	0.031***
	(0.0002)	(0.001)	(0.001)
57 – 72	0.001***	0.027***	0.005***
	(0.0001)	(0. 001)	(0.0001)
Missing	5.399***	0.578***	0.090***
	(0.170)	(0. 015)	(0.002)
Student outside region	0.948**	1.333***	1.259***
	(0.024)	(0.021)	(0.020)
ISCED code (ref: Natural sciences, mathematics and statistics)			
Information and Communication Technologies	0.977	0.542***	1.373***
	(0.032)	(0.014)	(0.034)
Engineering, manufacturing and construction	0.698***	0.796***	1.194***
	(0.013)	(0.009)	(0.015)
High School (ref: Classical lyceum)			
Scientific lyceum	0.946	0.661***	0.874***
	(0.033)	(0.012)	(0.017)
Technical Institute	4.506***	0.924***	1.448***
	(0.169)	(0.021)	(0.034)
Foreign Language lyceum	2.520***	1.344***	1.235***
	(0.161)	(0.054)	(0.052)
Professional Institute	5.168***	0.854***	1.555***
	(0.214)	(0.023)	(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.

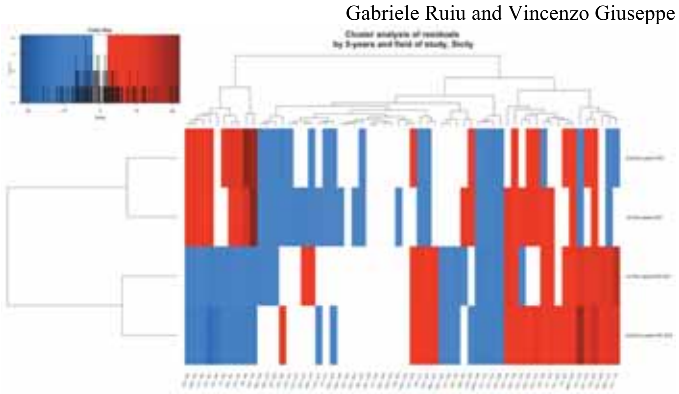


Figure 1: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Sicily

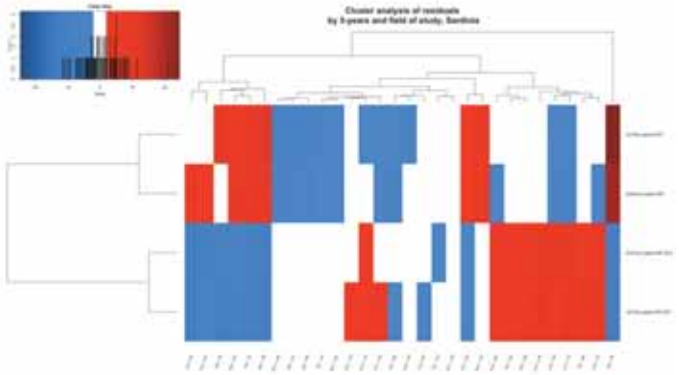


Figure 2: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Sardinia

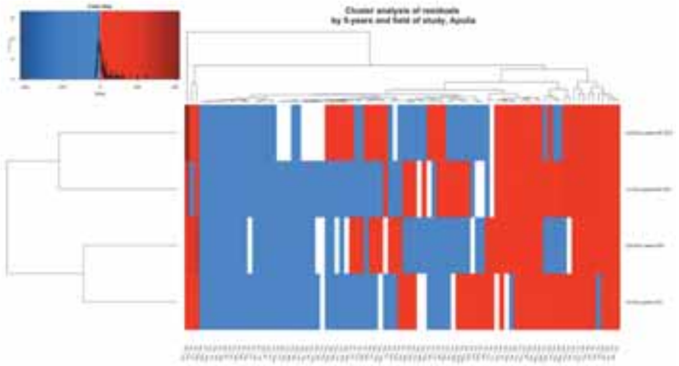


Figure 3: Heatmap associated with cluster analysis carried out for the five-year periods and the disciplinary field (STEM/non-STEM), Apulia

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 number) 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_{ij}^k 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 $\lambda_{ij} = \frac{K}{\sum_{k=1}^{K} \mathbb{1}(\gamma_{ij}^k = 0)}$ 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 **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 context 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].

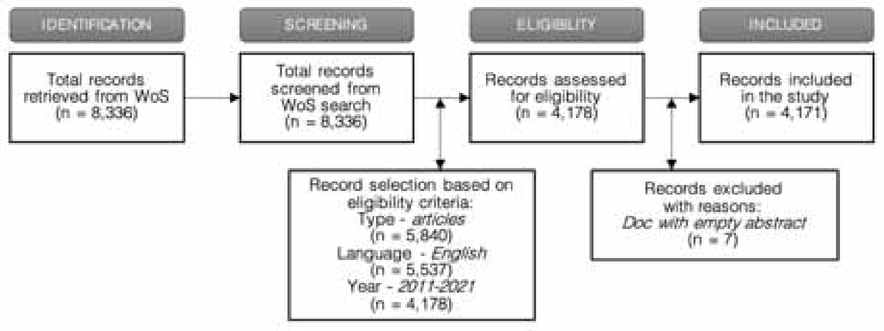

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Total records retrieved from WoS (n = 8,336)
Total records screened from WoS search (n = 8,336)
Records assessed for eligibility (n = 4,178)
Records included in the study (n = 4,171)
Record selection based on eligibility criteria: Type - *articles* (n = 5,840) Language - *English* (n = 5,537) Year - *2011-2021* (n = 4,178)
Records excluded with reasons: *Doc with empty abstract* (n = 7)

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).

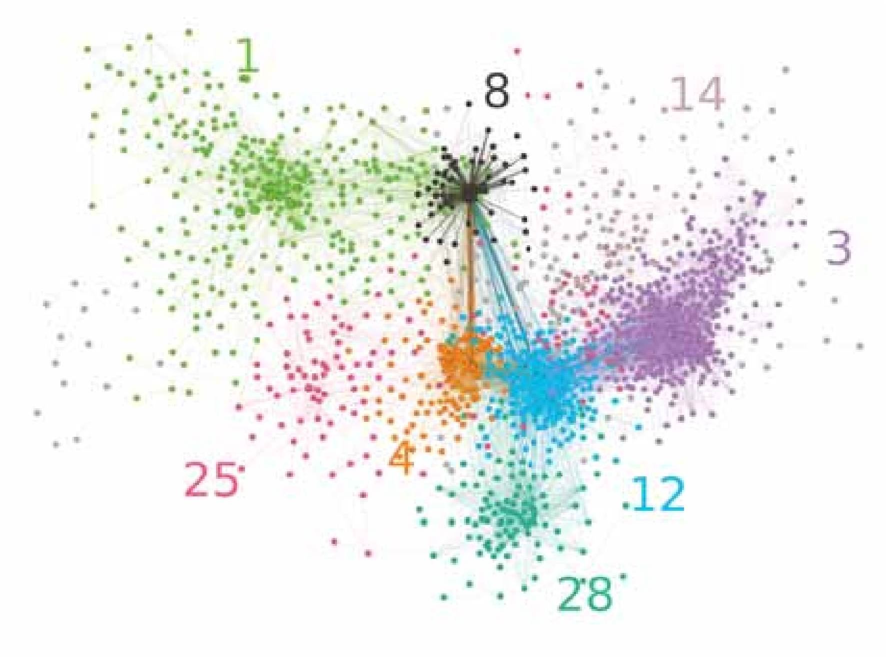


Fig. 2 Community detection on the abstract network (2011–2021).

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

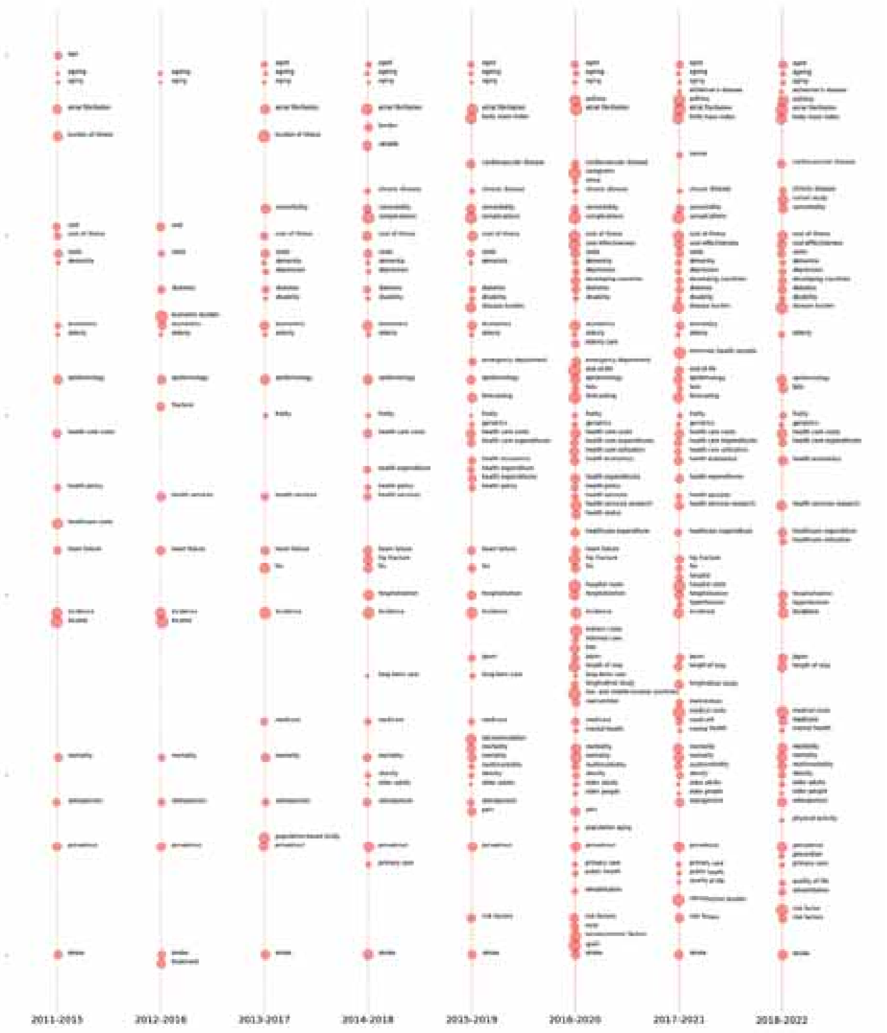


Fig. 3 Temporal evolution of publications included in community 14.

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 health-care 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, Scopus 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 $(k = 1,\dots,K)$ 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	⋆
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	⋆
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	⋆
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	⋆
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

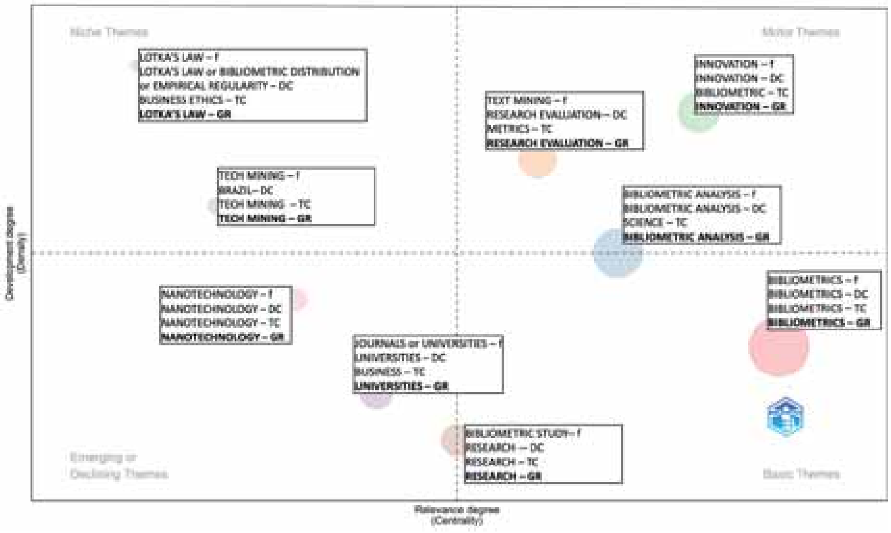


Fig. 1 Thematic diagram with the different AK candidates for labelling

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 $n{\times}p$ 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).

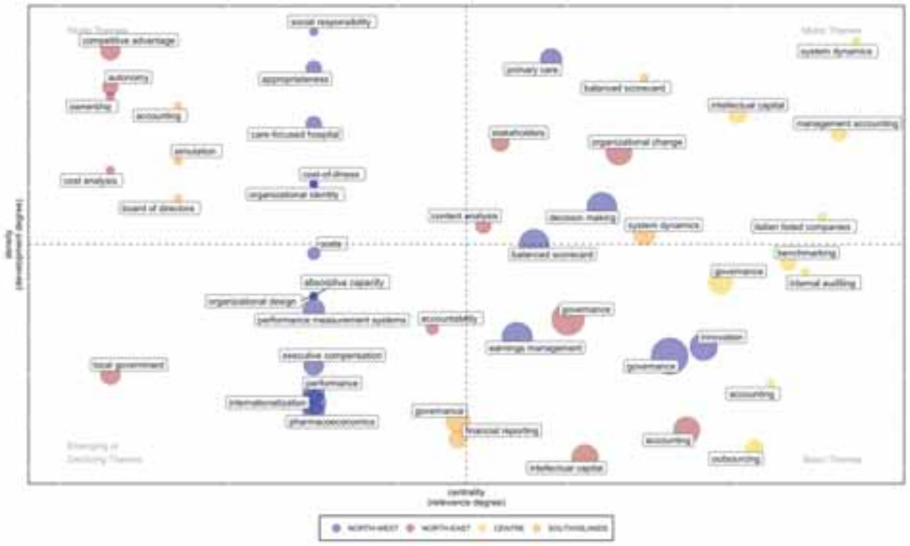


Figure 1: Thematic map. Focus on 2000-2010 timespan

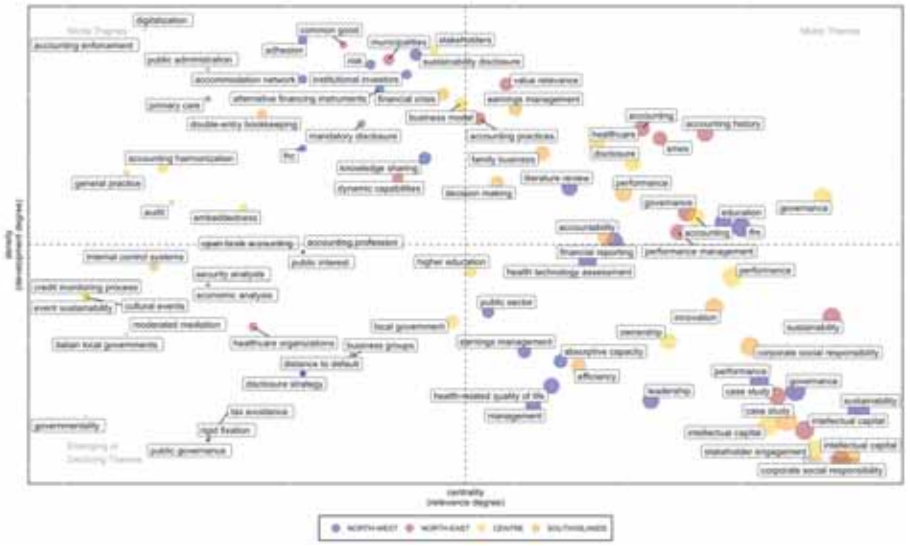


Figure 2: Thematic map. Focus on 2011-2021 timespan

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

$$\begin{aligned} P(Y_i = 0|X_i) &= (1-\theta_i) + \theta_i e^{-\lambda_i} && (1) \\ P(Y_i = y_i|X_i) &= \theta_i \frac{e^{\lambda_i} \lambda^{y_i}}{y_i!} && (2) \\ log(\lambda_i) &= \beta_0 + \beta^T X_i + u_i && (3) \\ logit(\theta_i) &= \gamma_0 + \gamma^T X_i && (4) \end{aligned}$$

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

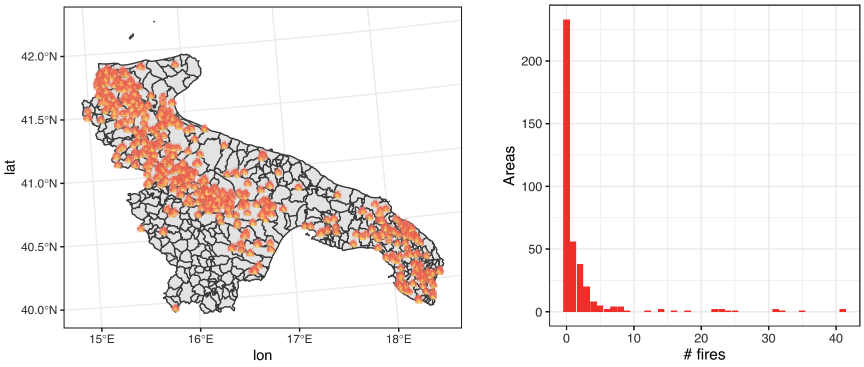


Fig. 1 Areas affected by fires (left) and fire counts (right) for Apulia and Basilicata regions during summer season 2021

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

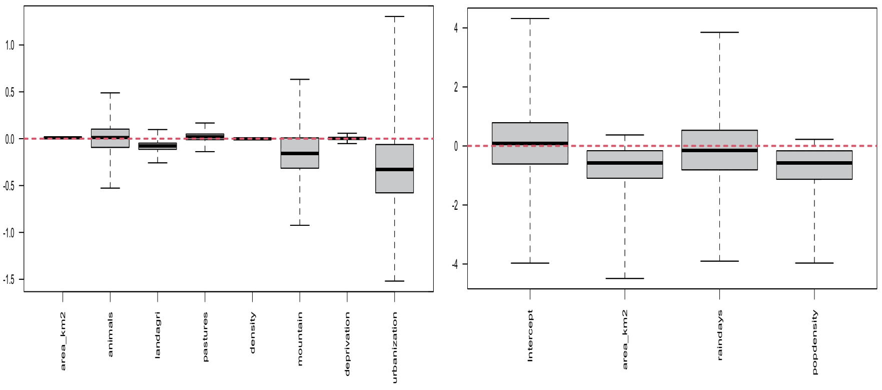


Fig. 2 Left panel: posterior distribution of regression parameters in (3). Right panel: posterior distribution of the regression parameters related to the zero component (4).

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.

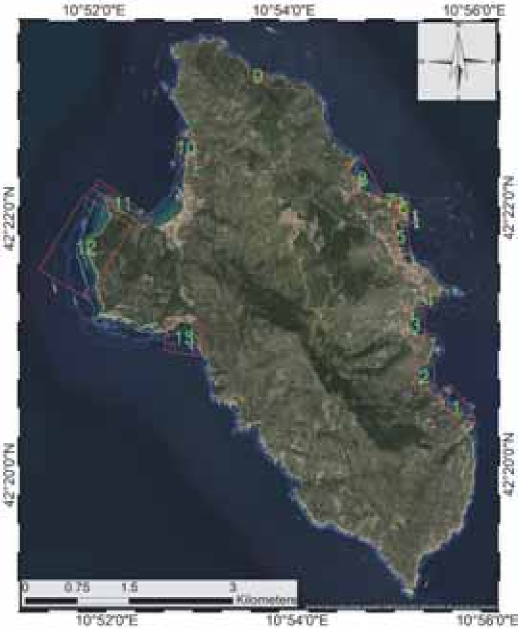


Fig. 1: Study area with the 13 zones highlighted.

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

$$Y_{it} \sim \text{Beta}(\mu_{it}, \tau_{it})$$

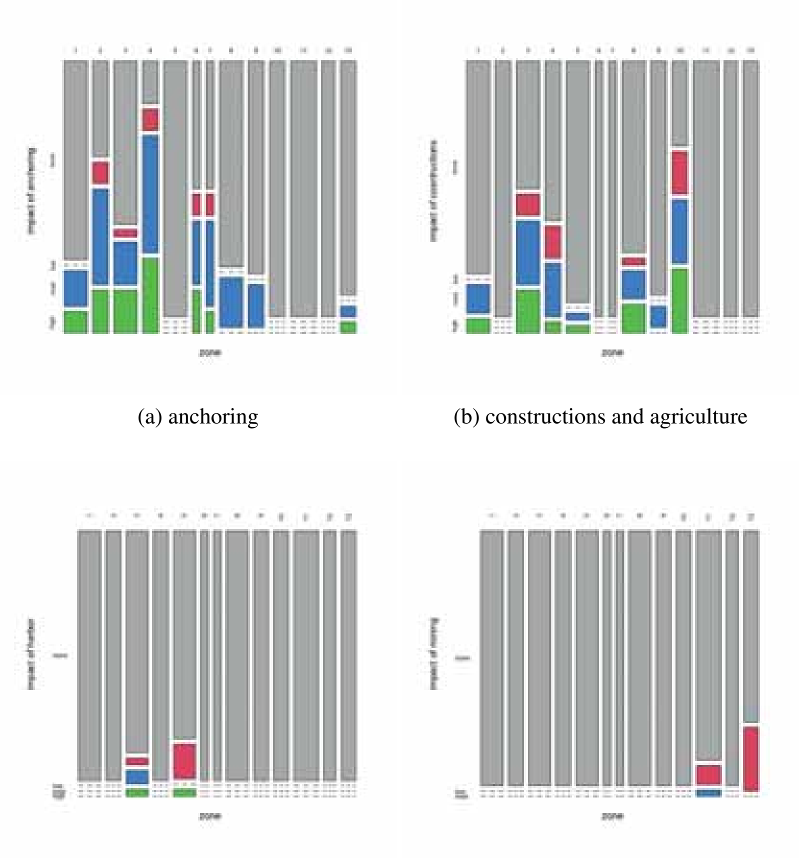
where Y_{it} is the i-th observation of Posidonia coverage at time t, μ_{it} is the mean of the distribution and τ_{it} the precision. Further

$$\text{logit}(\mu_{it}) = \beta_{0\mu}^{z_{k_i}} + \sum_{h=1}^{p} x_{hti}\beta_{h\mu}$$

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 $\beta_\mu, \beta_\tau \sim N(0, 1000)$, $z_j \sim$ Multinomial(π), with $j = 1, 2, \ldots, 13$, and $\pi \sim$ 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.

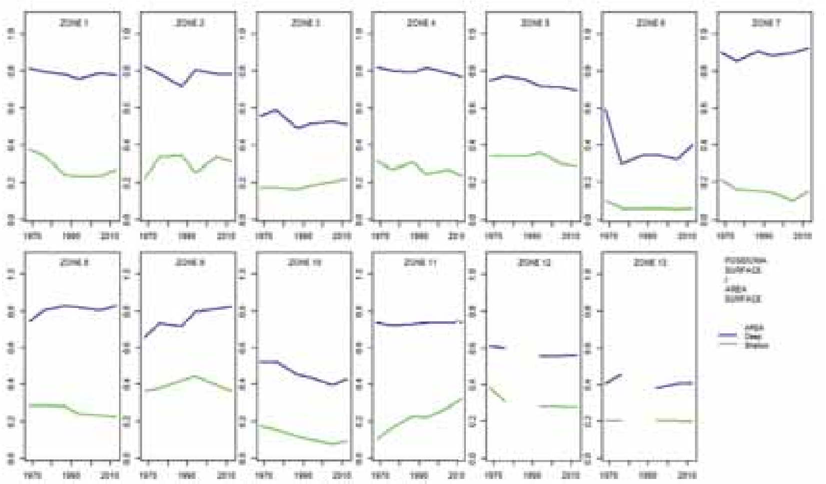


Fig. 3: Posidonia o. meadows coverage: time trend by zone and depth.

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:

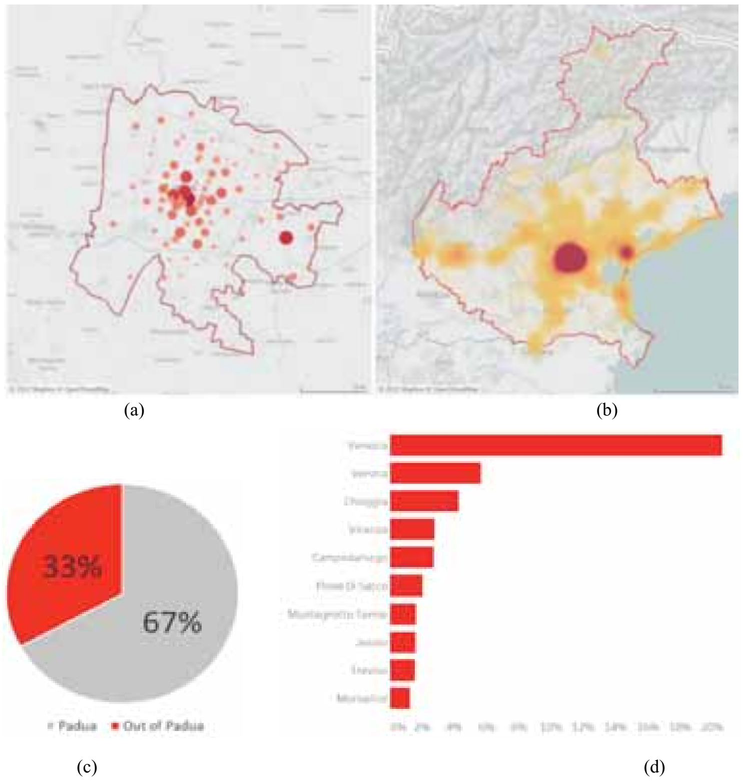
- prevailing night area: this information approximates the location of the overnight stay of the tourist and corresponds to the coverage of the prevailing network cells where the Italian SIM is mainly registered during the night;
- telco home location (or phone residence): corresponds to the municipality 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.



(a) (b)

(c) (d)

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

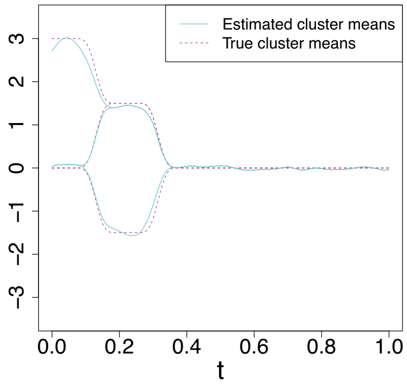
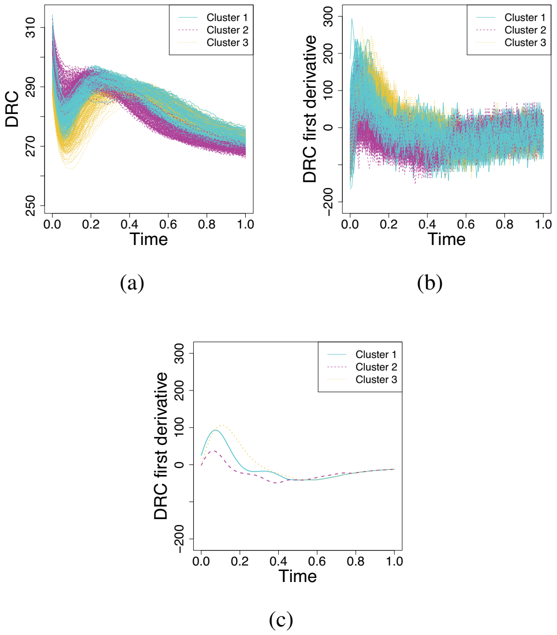


Fig. 1: True and estimated cluster means obtained through the SaS-Funclust method for a simulated data set in which the real number of clusters is $G = 3$.

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 practitioners 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.

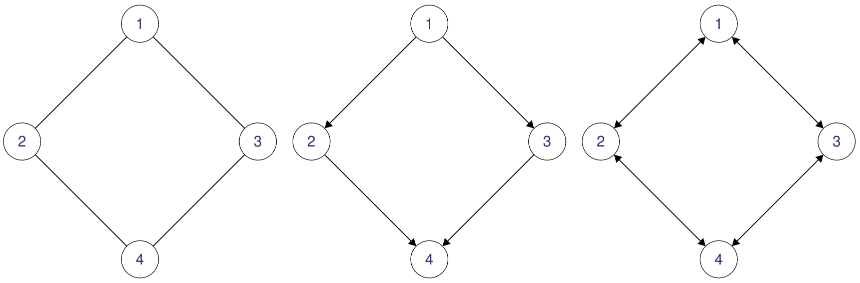


Fig. 1 Three example graphs for four variables: an undirected graph (left), an acyclic directed graph (center), and a bidirected graph (right).

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 $\mathbf{Y}$.

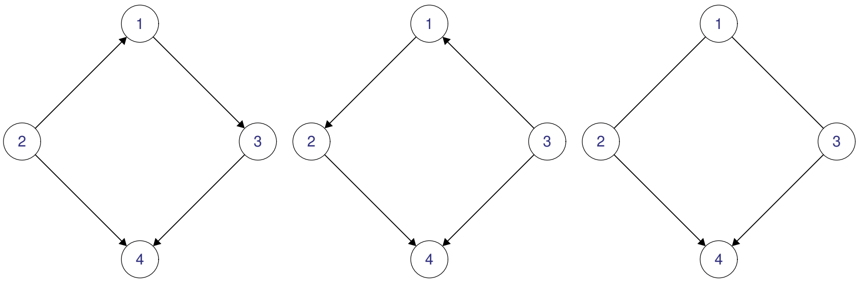


Fig. 2 Two equivalent acyclic directed graphs (left, center) and their essential graph (right); the latter has $\{1,2,3\}$ and $\{4\}$ as chain components.

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 $i \rightarrow k \leftarrow j$ 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 $2 \rightarrow 4 \leftarrow 3$ 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, $\mathrm{pa}(\tau) = \{i \in V \mid \mathbb{G} \text{ contains } i \rightarrow j \text{ for some } j \in \tau\}$ 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 $k = 1, \dots, K$, and obtain the full data matrix $\mathbf{Y}_{[1:K]}$ by stacking $\mathbf{Y}_{[1]}, \dots, \mathbf{Y}_{[K]}$ on top of one another. Then, conditional on a multiple essential graph $\mathbb{G}_{[1:K]} = (\mathbb{G}_{[1]}, \dots, \mathbb{G}_{[K]})$, assuming independence across groups (as well as within groups) and Gaussian observations, the likelihood, constrained by (4), can be written as

$$f_{\mathbb{G}_{[1:K]}}(\mathbf{Y}_{[1:K]} \mid \boldsymbol{B}_{[1:K]}, \boldsymbol{\Omega}_{[1:K]}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} \prod_{\tau \in \mathcal{T}_k} f_{\mathbb{G}_{[k]}}(\mathbf{y}_{[k]i\tau} \mid \mathbf{y}_{[k]i\mathrm{pa}(\tau)}, \boldsymbol{B}_{[k]\tau}, \boldsymbol{\Omega}_{[k]\tau}) \qquad (5)$$

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

$$\Pr(\mathbb{G}_{[1:K]} \mid \mathbf{Y}_{[1:K]}) \propto \Pr(\mathbb{G}_{[1:K]}) \prod_{k=1}^{K} \prod_{\tau \in \mathcal{T}_k} m_{\mathbb{G}_k}(\mathbf{Y}_{[k]\tau} \mid \mathbf{Y}_{[k]\mathrm{pa}(\tau)}) \qquad (6)$$

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 $\mathbb{G}_1, \dots, \mathbb{G}_K$, 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 ν 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. Studenỳ, 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)