Luigi D'Ambra, Pietro Amenta, Antonio Lucadamo, Anna Crisci
EDITORS

Statistical Methods for Service Quality Evaluation

Proceedings of IES 2019, Rome, Italy, July 4-5

Pearson

Luigi D'Ambra, Pietro Amenta, Antonio Lucadamo, Anna Crisci
EDITORS

Statistical Methods for Service Quality Evaluation

Proceedings of IES 2019, Rome, Italy, July 4-5

9th International Conference **IES 2019** - Innovation & Society - **Statistical evaluation systems at 360°: techniques, technologies and new frontiers** organized by Statistics for the Evaluation and Quality in Services Group of the Italian Statistical Society and European University of Rome

Published by Pearson

www.pearson.com

Giugno 2020 ISBN 9788891921239

On the acceptable level of inconsistency of Pairwise Comparison Matrices: a proposal of thresholds based on the percentile rule....1
Livelli accettabili di inconsistenza per le matrici di confronto a coppie: una proposta di calcolo delle soglie basata sulla regola del percentile
Pietro Amenta, Antonio Lucadamo and Gabriella Marcarelli

Mines and quarries: An analysis of withdrawals determinants in Italy 9
Cave e Miniere: un'Analisi delle Determinanti dell'Estrazione in Italia
Sabrina Auci and Donatella Vignani

Evaluating people's behaviour towards risk: a multidimensional problem....15
Valutazione del comportamento individuale verso il rischio: un problema multidimensionale
Luigi Bollani, Guido Antonio Rossi and Ivan Sciascia

Inferential versus descriptive statistical approach in the analysis of Delphi performance: A case study. 23
Analisi delle performance del Delphi. Approccio inferenziale verso descrittivo: studio di un caso
Bolzan M., Auciello M., Pesarin F

Biplot and unfolding models for evaluation of teacher behaviour in the classroom....33
Utilizzo dei modelli biplot e unfolding per la valutazione dell'atteggiamento dell'insegnante in classe
Giuseppe Bove, Maria Gaetana Catalano, Paola Perucchini, Alessio Serafini, Giovanni Maria Vecchio

Student satisfaction for Teaching and its impact on drop-out rate....41
La soddisfazione degli studenti per la didattica e il suo legame con il tasso di abbandono
Barbara Cafarelli, Angela Maria D'Uggento and Alessandra Petrucci

Space-time nonparametric analysis for the Italian real estate market....49
Analisi non parametrica spazio-temporale per il mercato immobiliare italiano
Cappello C. and De Iaco S. and Palma M. and Pellegrino D. and Posa D.

A Multidimensional Approach for Classifying Italian University Students by Mobility 57
Un approccio multidimensionale per la classificazione degli studenti universitari italiani secondo la mobilità
Sara Casacci

Robust analysis of the labor market 65
Analisi robusta del mercato del lavoro
Aldo Corbellini, Marco Magnani and Gianluca Morelli

Analysis of the financial performance in Italian football championship clubs via longitudinal count data 73
Analisi della performance finanziaria delle squadre di calcio Italiane attraverso longitudinal count data
Anna Crisci and Luigi D'Ambra

A Poset perspective for the evaluation of self-reported health of the elderly in Italy 81
Un metodo basato sul POSET per valutare lo stato di salute auto-percepita della popolazione anziana in Italia
E. Furfaro, L. Pagani and M. C. Zanarotti

Reduced K-means Principal Component Multinomial Regression with external information to evaluate consumer's preferences 89
Reduced K-means Principal Component Regression con informazioni esterne per la valutazione delle preferenze dei consumatori
Antonio Lucadamo and Pietro Amenta

A two-part finite mixture quantile regression model for semi-continuous longitudinal data ...97
Modello di regressione quantilica mistura a due parti per dati longitudinali
Maruotti Antonello, Merlo Luca and Petrella Lea

New developments in the evaluation of goodness of fit for multidimensional IRT models based on posterior predictive assessment: Results from the INVALSI data....103
Nuovi sviluppi nello studio della bontà di adattamento per i modelli di IRT multidimensionali basati sulla valutazione predittiva a posteriori: Alcuni risultati sui dati INVALSI
Mariagiulia Matteucci and Stefania Mignani

Models for measuring well-being. Reflective or Formative? 111
Modelli per misurare il benessere. Riflessivi o formativi?
Matteo Mazziotta and Adriano Pareto

Automated Content Analysis of Destination Image: a Case Study 119
Automated Content Analysis dell'immagine della destinazione: studio di un caso
Antonino Mario Oliveri and Gabriella Polizzi

Evaluating the determinants of environmentally significant behaviour in Europe.....127
Le determinanti dei comportamenti sostenibili in Europa
Gennaro Punzo, Demetrio Panarello, Margherita Maria Pagliuca, and Rosalia Castellano

Financial transaction data for early estimates of macroeconomic indicators for Services in Italy: value added and turnover index 135
I dati di pagamenti elettronici per le stime anticipate di indicatori macroeconomici dei Servizi in Italia: valore aggiunto e indice del fatturato
Alessandra Righi, Guerino Ardizzi, Alessandro Gambini, Filippo Moauro, Nazzareno Renzi

Education, Women and Empowerment: the case of India 143
Istruzione, donne ed empowerment: il caso indiano
Azzurra Rinaldi and Fabiana Sciarelli

LOCAL AUTHORITIES AND TOURIST USE OF THE TERRITORY.....151
ENTI LOCALI E USO TURISTICO DEL TERRITORIO
Sara Sergio

An index for crowdsourced data on multipoint scales in tourism services evaluation.....159
Un indice per dati crowdsourced su scale multipoint nella valutazione dei servizi turistici
Venera Tomaselli and Giulio Giacomo Cantone

Testing consistency in preference orderings.....167
Un test sulla consistenza tra ordinamenti di preferenze
Amalia Vanacore, Maria Sole Pellegrino, Yariv N. Marmor and Emil Bashkansky

A simulation study to investigate the paradoxical behaviour of inter-rater agreement coefficients in non-asymptotic conditions 177
Studio in simulazione del comportamento paradossale dei coefficienti di accordo inter-valutatore in condizioni non asintotiche
Amalia Vanacore and Maria Sole Pellegrino

The effect of noise factors in experimental studies on aircraft comfort.....187
L'effetto dei fattori di disturbo sulla valutazione sperimentale del comfort aereo
Amalia Vanacore and Chiara Percuoco

Climate Change and Italian Cities: evidence from Meteo-climatic Statistics and Indices on Extreme Events.....197
Cambiamenti Climatici e Città Italiane: le Statistiche Meteoclimatiche e gli Indici di Estremi Climatici
Donatella Vignani, Francesca Budano, Claudia Busetti

Preface

This Book of Proceedings includes a selection of 25 peer-reviewed papers submitted to the **Innovation & Society 2019** (IES 2019) - *Statistical evaluation systems at* 360°: *techniques, technologies and new frontiers* held at the European University of Roma from July 4^{th} to 5^{th} , 2019. With 80 contributions organized in solicited and contributed sessions, two plenary talks and 220 authors, underlying a very strong interest around the evaluation topics, IES2019 has been the 9^{th} meeting of a two-year initiative proposed by the permanent group Statistics for the Evaluation and Quality in Services (SVQS).

SVQS group was born in 2004 thanks to the collaborative work of some members of the Italian Statistical Society (SIS) with a focus on national research programs and applied research activities, on statistical methods and methodologies for the evaluation of the quality of services in the public and private fields.

SVQS organizes IES conference every two years. Recent debate, indeed, has shown that public and private service quality measurement is the basic prerequisite for quality planning and strongly improve public policies that highly impact our societies. This topic is also strongly interconnected with the satisfaction measurement, which is one of the most important tools for firms and public institutions to fully capture consumers and citizens needs. Recently, new challenges have emerged in particular from study designs, large and heterogeneous data sources availability and complex treatment assignment mechanisms. Big data also provide a complement to traditional data sources, such as survey and census data, to create a complete analysis of a service process. In this field, numerical taxonomy, classification, multidimensional scaling and other ordination techniques, clustering, tree structures and other network models, as well as other statistical models (e.g., multilevel or latent variable models) for the analysis of data of a different nature (e.g., ranking or ordinal categorical data) play a crucial role together with related inferential methods that may depart from traditional methods (e.g., methods based on composite likelihood or generalized estimating equations).

IES2019 has been sponsored by the Italian National Institute of Statistics (ISTAT), the European Network for Business and Industrial Statistics (ENBIS), and by two groups of the Italian Statistical Society: Statistics and Data Science (SDS) and Enhancement of Public Statistics (VSP).

This conference aimed at

- shedding light on the main statistical approaches and methodologies for evaluation, currently in use in different contexts, of public utility services;
- fostering advanced methodological research supporting the dissemination of ideas related to several fields of interest;
- contributing to the discussion on the innovative statistical evaluation systems impact of services, involving several economic and social policies actors;
- being a platform where statisticians, data analysts, machine learning researchers meet to understand and analyze service phenomena with data.

Previous editions of IES conference were:

- **IES2009** held at University of Brescia (June 24-26, 2009) with selected papers published in special issues of *Electronic Journal of Applied Statistical Analysis* (*EJASA*) and *Statistica & Applicazioni*;
- **IES2011** held at University of Florence (May 30-June 1, 2011) with selected papers published in a special issue of the *Journal of Applied Quantitative Methods* (ADSE);
- **IES2013** held at University of Milan "Bicocca" (December 9-13, 2013) with selected papers published in the *Procedia Economics & Finance* (Elsevier);
- **IES2015** held at University of Bari "Aldo Moro" (June 8-9, 2015) with selected papers published in a special issue of *Quality & Quantity* (Springer);
- **IES2017** held at University of Naples "Federico II" (September 6-7, 2017) with selected papers published in special issues of *Social Indicator Research* (Springer), *Quality & Quantity* (Springer) and *EJASA* (ESE).

Next IES conference *Innovation and Society 5.0: Statistical and Economic Models and Techniques for Quality Assessment* (**IES2021**) will take place from July 1^{th} to 2^{th} 2021 at the Department of Economics of the University of Campania "Luigi Vanvitelli" (Capua, Italy).

Moreover, special issues of the international journals *Socio-Economic Planning Science* (Elsevier) and *Metron* (Springer) published a selection of full papers. This is a strategic way to disseminate recent developments and critical discussion in statistics to a wider community who did not participate directly to the event.

The Scientific Program Committee, the Chair (Prof. Matilde Bini) and the Local Organizing Committee, with the support of the European University of Rome, have all contributed to a productive and stimulating IES2019 conference. We acknowledge their precious work.

Luigi D'Ambra, Pietro Amenta, Antonio Lucadamo, Anna Crisci

Editors

IES2019 Scientific and Program Committee

Giorgio Alleva (University of Rome ”La Sapienza”)
Pietro Amenta (University of Sannio)
Francesco Bartolucci (University of Perugia)
Eric Beh (University of Newcastle, Australia)
Wicher Bergsma (The London School of Economics and Political Science, UK)
Matilde Bini (European University of Rome) - **Chair**
Mario Bolzan (University of Padova)
Eugenio Brentari (University of Brescia)
Maurizio Carpita (University of Brescia)
Rosalia Castellano (University of Naples ”Parthenope”)
Paola Cerchiello (University of Pavia)
Vartan Choulakian (University of Moncton, Canada)
Corrado Crocetta (University of Foggia)
Antonello D’Ambra (University of Campania ”L.Vanvitelli”)
Luigi DAmbra (University of Naples ”Federico II”)
Tonio Di Battista (University of Chieti ”G.DAnnunzio”)
Michele Gallo (University of Naples ”Orientale”)
Anna Giraldo (University of Padova)
Luca Greco (University of Sannio)
P.M. Kroonenberg (University of Leiden, Netherlands)
Michele La Rocca (University of Salerno)
Rosaria Lombardo (University of Campania ”L.Vanvitelli”)
Filomena Maggino (University di Florence)
Paolo Mariani (University of Milan ”Bicocca”)
Lucio Masserini (University of Pisa)
Stefania Mignani (University of Bologna)
Francesco Palumbo (University of Naples ”Federico II”)
Lea Petrella (University of Rome ”Sapienza”)
Alessandra Petrucci (University of Florence)
Jean Michel Poggi (Paris-Descartes University & Lab. Maths Orsay, France)
Emilio Porcu (University of Newcastle, UK)
Donato Posa (University of Salento)
Gennaro Punzo (University of Naples ”Parthenope”)
Carla Rampichini (University of Florence)
Alessandra Righi (ISTAT)
Pasquale Sarnacchiaro (University of Rome ”Unitelma Sapienza”)
Nicola Tedesco (University of Cagliari)
Amalia Vanacore (University of Naples ”Federico II”)
Grazia Vicario (Polytechnic University of Turin)
Maurizio Vichi (University of Rome ”Sapienza”)

IES2019 Organising Committee

Antonio Lucadamo (University of Sannio)
Pietro Amenta (University of Sannio)
Matilde Bini (European University of Rome)
Lucio Masserini (University of Pisa)
Pasquale Sarnacchiaro (University of Rome ”Unitelma Sapienza”)
Margherita Velucchi (University of Rome ”Unitelma Sapienza”)

On the acceptable level of inconsistency of Pairwise Comparison Matrices: a proposal of thresholds based on the percentile rule

Livelli accettabili di inconsistenza per le matrici di confronto a coppie: una proposta di calcolo delle soglie basata sulla regola del percentile

Pietro Amenta, Antonio Lucadamo and Gabriella Marcarelli

Abstract Several measures have been proposed to verify the rationality and the consistency of judgements expressed by means of pairwise comparisons. To verify if a matrix can be considered consistent or not some thresholds for these indices have been introduced. However when the number of alternatives is high, the thresholds proposed until now lead to consider as inconsistent almost all the matrices, forcing to revise the judgments. In this paper, using a 10th percentile rule, we propose to derive new approximated thresholds for the Consistency Ratio proposed by Saaty, for the Geometric Consistency Index introduced by Aguaron-Moreno Jimenez, for the Salo-Hamalainen index and for the Consistency Measure proposed by Koczkodaj. The new bounds can be used to check if the level of inconsistency can be considered acceptable or not.

Abstract *Diverse misure sono state proposte per verificare la razionalitá e la consistenza di giudizi espressi attraverso i confronti a coppie. Per verificare se una matrice puó essere considerata consistente o meno, alcune soglie per questi indici sono stati introdotti. Tuttavia quando il numero di alternative é alto, le soglie proposte fino ad ora portano a considerare come inconsistenti quasi tutte le matrici, forzando a rivedere i giudizi. In questo articolo, usando la regola del decimo percentile, proponiamo di derivare nuove soglie approssimate per il Consistency Ratio proposto da Saaty, per il Geometric Consistency Index introdotto da Aguaron-Moreno Jimenez, per l'indice di Salo-Hamalainen e per la misura di Consistenza proposta da Koczkodaj. I nuovi valori soglia possono essere usati per controllare se il livello di inconsistenza pu essere considerato accettabile o no.*

Pietro Amenta
University of Sannio, Benevento, Italy, e-mail: amenta@unisannio.it

Antonio Lucadamo
University of Sannio, Benevento, Italy, e-mail: antonio.lucadamo@unisannio.it

Gabriella Marcarelli
University of Sannio, Benevento, Italy, e-mail: gabriella.marcarelli@unisannio.it

Key words: Pairwise comparisons, Inconsistency indices, Consistency thresholds, Analytic Hierarchy Process

1 Introduction

Pairwise comparison matrix is one of the most common tool used for representing the preferences of Decision Makers (DMs) in multi-criteria decision problems. They are used, among others, in Analytic Hierarchy Process (AHP) and its generalisation [16, 17]. However, pairwise comparisons have two main issues: the consistency of the judgements and the reliability of the preferences. The judgements expressed by pairwise comparisons may be not consistent (irrational). The reliability of preferences are strongly related to consistency of judgements. When the judgments are less consistent the priority vector estimates are lousy [9, 10]. If judgements are perfectly consistent, then all prioritisation methods give the same result, but if they are not consistent, then each method leads to a different priority vector. Perfect consistency is unattainable in practice and for this reason a degree of inconsistency can be considered acceptable. It is, therefore, obviously necessary to measure the consistency of judgements before deriving a priority vector.

Many consistency indices have been proposed in the literature to measure the level of inconsistency in a set of pairwise judgements. For example, the Consistency Index (CI) and the Consistency Ratio (CR) introduced by Saaty [16], the Consistency Measure (CM_K) proposed by Koczkodaj [12], the Consistency Measure (CM_{SH}) introduced by Salo-Hamalainen [19], and the Geometric Consistency Index (GCI) that was proposed by Crawford and Williams [6]. See also [2, 14, 17] for a deeper analysis on this matter. These indices are based on the distance from the perfect consistency condition. The range of the values that indices may assume varies depending on the matrix size and the index we consider. Thresholds defining the level of consistency/inconsistency have been introduced for some of them [2, 7, 16]. If the index has a value lower than the threshold, then the judgements are at an acceptable level of inconsistency; otherwise, the inconsistent judgements should be revised to avoid invalid decision. In this paper we introduce new bounds, based on the 10th percentile rule, that define a tolerable level of inconsistency, on varying the matrix size n for CR, GCI, CM_{SH} and CM_K; they may provide useful information when jointly used to classical consistency thresholds proposed in literature. The rest of the paper is organised as follows: section 2 illustrates some indices proposed in literature to measure the consistency of pairwise comparisons; in section 3 the new thresholds, defining the level of tolerable inconsistency, for CR, GCI, CM_{SH} and CM_K are proposed; in section 4 the use of inconsistency thresholds in tourism services is described; finally, some concluding remarks are provided in section 5.

2 Background

Given a set X of n elements, in order to derive the ranking of preferences, a positive number a_{ij} is assigned to each pair of elements (x_i, x_j). The number measures how much x_i is preferred to x_j respect to a given criterion. By comparing all the elements, a positive square matrix $A = (a_{ij})$ of order n is then obtained. The value $a_{ij} > 1$ means that x_i is strictly preferred to x_j, whereas $a_{ij} < 1$ expresses the opposite preference, and $a_{ij} = 1$ implies that x_i and x_j are indifferent [16, 17]. The matrix A is at the heart of many methods that have been proposed in the literature to derive a priority vector, $w = (w_1 \dots w_n)$, representing the ranking of preferences [2, 8, 13, 14, 16]. Regardless of the method used for the prioritisation procedure, before applying any methods, it is necessary to check the consistency of these judgements.

A decision maker is perfectly consistent in making estimates if his or her judgements satisfy the consistency condition $a_{ij} * a_{jk} = a_{ik}$ for each $i, j, k = 1, 2, \dots, n$ [16]. Many factors could affect the judgements and in this case the pairwise comparison matrix may be not consistent. The consistency of judgements is related to the reliability of the preferences: for this reason it has been widely analysed by many authors, that proposed several indices to measure the degree of consistency of the judgements expressed by the decision maker. Following we describe the indices that are most widely used in the AHP literature. Saaty proposed the Consistency Index (CI), given by

$$CI = \frac{\lambda_{max} - n}{n-1}, \tag{1}$$

for $i, j = 1, \dots, n$, where λ_{max} represents the maximum eigenvalue of the pairwise comparison matrix. If the matrix is perfectly consistent, then $CI = 0$. Saaty [16] suggested also to use the Consistency Ratio

$$CR = \frac{CI}{RI}, \tag{2}$$

with RI Random Index obtained as the mean value of the CI derived from randomly generated matrices of order n. The CR value should not be higher than 0.1. If $CR > 0.1$, then the decision maker has to revise his or her judgements to improve the consistency [17]. A measure of inconsistency based on the estimator of the variance of the perturbation, when the Row Geometric Mean Method (RGMM) is used as prioritization procedure was suggested by Crawford and Williams [6]: the Geometric Consistency Index (GCI) index, based on the logarithmic residual mean square, is defined as

$$GCI = \frac{2}{n(n-1)} \sum_{i<j} \log^2 e_{ij}, \tag{3}$$

where $e_{ij} = a_{ij} \times \frac{w_j}{w_i}$ represents the error obtained when the ratio w_j/w_i is approximated by a_{ij} and w is the vector derived by the RGMM. Aguaron and Jimenez [2] proposed consistency thresholds for GCI equivalent to CR. Koczkodaj has defined the following consistency measure:

$$CM_K = \max_{i,j,k} \left[\min \left(\left| 1 - \frac{a_{ik}}{a_{ij}a_{jk}} \right|, \left| 1 - \frac{a_{ij}a_{jk}}{a_{ik}} \right| \right) \right], \tag{4}$$

based on the triplet of the elements of a pairwise comparison matrix, with $1 \leq i < j < k \leq n$. This measure does not depend on the scale and is not connected with any specific prioritisation method [5, 7, 12].

The Salo-Hamalainen Consistency Index [19] is defined as:

$$CM_{SH} = \frac{2}{n(n-1)} \sum_{i>j} \frac{\bar{r}(i,j) - \underline{r}(i,j)}{(1+\bar{r}(i,j))(1+\underline{r}(i,j))} \tag{5}$$

where $\bar{r}(i,j) = \max_k (a_{ik} \cdot a_{kj})$ and $\underline{r}(i,j) = \frac{1}{\bar{r}(j,i)}$.

Threshold values for some consistency indices have been introduced by many authors. These measures differ according to the sizes of the pairwise comparison matrix (PCM). In particular, Saaty suggested at first the value 0.1 as *CR* threshold [16] for all matrix sizes. Then he introduced additional threshold values of 0.05 and 0.08 for matrices of size 3 and 4 respectively [17], respectively. Aguaron and Moreno-Jimenez [2] instead proposed the following threshold values for the GCI index: 0.31 for $n = 3$, 0.35 for $n = 4$, 0.37 for $n > 4$.

Anyway, many questions are still opened about the choice of the right cut-off rule to declare the inconsistency of a matrix. Ishizaka and Labib [11] highlighted that this rule should not depend on the size of the matrix, but if we look at table 1, it is evident that if we consider the percentage of matrices for which the *CR* is less than the corresponding threshold proposed by Saaty, it varies according to matrix size. Indeed, about 13% of random matrices has a *CR* lower than 0.05 for 3×3 matrices, about 2% for matrices of order 4, 0.2 % and 0.006% for 5×5 and 6×6 respectively. When the alternatives to be evaluated by the decision makers are greater than 6, it is evident that all the pairwise comparison matrices have to been considered not consistent. This means that the Decision Makers (DMs) usually have to revise their judgments. The same considerations hold if we look at the GCI thresholds. In this case the proportion of consistent matrices is a little bit higher than before, if the number of alternatives is between 3 and 6 (about 21%, 3% 0.3% and 0.01% respectively), but when $n \geq 7$ all the matrices must be revised.

Table 1 Percentage of matrices with *CR* or *GCI* less than threshold values

	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
CR thresholds	0.05	0.08	0.1	0.1	0.1
% of matrices with $CR \leq$ threshold	12.74	2.219	0.234	0.009	0.000
GCI thresholds	0.31	0.35	0.37	0.37	0.37
% of matrices with $GCI \leq$ threshold	20.617	3.305	0.278	0.012	0.000

To solve this issue Saaty [18] proposed a method based on perturbation theory to find the most inconsistent judgment in a PCM. This method improves the consistency by asking the judges to provide a new judgment within a certain range of

values. If the DMs are not willing to revise the judgments or are not achievable, a possible solution to improve consistency is to consider a linearization technique that provides the closest consistent matrix to a given non-consistent matrix. This procedure provides a closed form for achieving consistency, by using an orthogonal projection in a given linear space [4]. This approach aims to overcome the inconsistency transforming the original PCM in a matrix close to it.
In our paper, we propose a method to introduce new thresholds measuring the inconsistency level for the above mentioned indices. Without contrasting with the classical ones, our bounds may be taken into account before improving the consistency by means of the previous mentioned procedures [4]. If the consistency index of a PCM assumes a value higher than the classical thresholds, the supervisor can ask, if possible, the DMs to revise their judgments or, alternatively, evaluate if the level of inconsistency is acceptable according to the new bounds. Only in last case, he can decide to compel the matrix to be consistent as suggested in [4].

3 Consistency threshold values based on the 10th percentile rule

In this section, we introduce the new threshold values for CR, GCI, CM_{SH} and CM_K, using the 10th percentile rule. In order to perform the analysis, we use the statistical software R [15]. We simulate 500000 reciprocal matrices for each size (n) in the following way: we generate random number for the upper triangular matrices (a_{ij} for $i < j$) from the set of values $(1/9, 1/8, \ldots, 1/2, 1, 2, \ldots, 9)$; then we calculate the reciprocal values for the lower triangular matrices (a_{ij} for $i > j$). Finally we set values on the main diagonal ($a_{ij} = 1$ for $i = j$).

We do not aim to substitute the existing thresholds, but our idea is to define an acceptable level of inconsistency. It can be obtained considering as additional thresholds, the values of the indices that, independently by the number of alternatives, represent the 10th percentile of the empirical sample distribution. Furthermore, for completeness of the analysis we calculated the thresholds also for some other percentiles (5th and 15th). In table 4 we show the threshold values for CR, GCI, CM_{SH}, CM_K and for different number of alternatives (n), by using the percentiles rule.

It is easy to see that we obtain different thresholds according to matrix size, but through our procedure, the percentages of matrices for which a revision of judgements is required do not change. If we instead consider the Saaty thresholds, the percentage of matrices for which the CR index is less than the corresponding threshold, varies according to matrix size (as shown in previous section). This can be particularly problematic when, for example, it is not possible to associate the matrix to the Decision Maker who filled in it. In this case, it is not possible to revise the judgments and so, when the number of alternatives is high, there is the risk of not using many matrices that are considered inconsistent according to classical thresholds. In next section we show how our idea can be useful in AHP applied to tourism services.

Table 2 Threshold values for each index according to several percentiles and different matrix size orders (n). Results are based on 500000 random PCMs for each n

percentile	Index	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
0.05	CR	0.0067	0.1308	0.2757	0.4379	0.5740	0.6622	0.7214
	GCI	0.0211	0.4422	0.9012	1.3125	1.6075	1.7924	1.9229
	CM_{SH}	0.0429	0.2551	0.4362	0.5810	0.6779	0.7434	0.7908
	CM_K	0.2222	0.8000	0.9365	0.9821	0.9917	0.9944	0.9959

percentile	Index	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$
0.05	CR	0.7636	0.7946	0.8186	0.8374	0.8529	0.8656
	GCI	2.0272	2.1027	2.1618	2.2106	2.2503	2.2821
	CM_{SH}	0.8264	0.8532	0.8736	0.8892	0.9019	0.9119
	CM_K	0.9965	0.9970	0.9974	0.9975	0.9977	0.9978

percentile	Index	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
0.10	CR	0.0279	0.1961	0.3734	0.5598	0.6735	0.7407	0.7858
	GCI	0.0870	0.6539	1.1777	1.5942	1.8290	1.9796	2.0836
	CM_{SH}	0.0841	0.3100	0.5004	0.6312	0.7150	0.7718	0.8128
	CM_K	0.4000	0.8571	0.9592	0.9889	0.9938	0.9955	0.9964

percentile	Index	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$
0.10	CR	0.8177	0.8421	0.8595	0.8745	0.8866	0.8956
	GCI	2.1643	2.2225	2.2669	2.3040	2.3359	2.3606
	CM_{SH}	0.8439	0.8669	0.8845	0.8998	0.9092	0.9179
	CM_K	0.9969	0.9974	0.9975	0.9977	0.9979	0.9980

percentile	Index	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
0.15	CR	0.0515	0.2574	0.4639	0.6428	0.7375	0.7928	0.8281
	GCI	0.1602	0.8451	1.4116	1.7826	1.9863	2.1087	2.1889
	CM_{SH}	0.1185	0.3524	0.5464	0.6631	0.7381	0.7898	0.8262
	CM_K	0.5000	0.8889	0.9750	0.9917	0.9948	0.9960	0.9968

percentile	Index	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$
0.15	CR	0.8536	0.8738	0.8872	0.8987	0.9083	0.9163
	GCI	2.2552	2.3035	2.3383	2.3672	2.3938	2.4138
	CM_{SH}	0.8544	0.8751	0.8909	0.9035	0.9135	0.9215
	CM_K	0.9972	0.9974	0.9977	0.9978	0.9980	0.9980

4 AHP and consistency in hotel services

Tourism industry is one of the most rapidly growing sectors in the world. Hotels constitute the main units of tourism sectors. These enterprises consist of the businesses where the production and consumption of touristic goods and services occur simultaneously. One of the factors to have good results is to satisfy customer requirements. Customer satisfaction surveys are then necessary to understand which services are important for the tourists and the use of pairwise comparison matrices can help in this selection. The classification of a matrix as consistent or inconsistent is then essential. Let's consider, i.e. the following matrix in which the preferences of a consumer are expressed according to the Saaty scale.

	Food	Cleanliness	Staff	Price/benefit	Comfort	Position
Food	1	2	6	8	8	9
Cleanliness	1/2	1	3	4	4	9
Staff	1/6	1/3	1	1	4	3
Price/benefit	1/8	1/4	1	1	2	9
Comfort	1/8	1/4	1/4	1/2	1	5
Position	1/9	1/9	1/3	1/9	1/5	1

The *CR* index for this matrix is equal to 0.086, so it can be considered consistent and prioritization methods can be applied to define the preference ranking. If the hotel wants to evaluate more than six services, we can consider the following pairwise comparison matrix, where we simply add alternatives Wi-fi and Parking to the evaluations given in the previous matrix.

	Food	Cleanliness	Staff	Price/benefit	Comfort	Position	Wi-fi	Parking
Food	1	2	6	8	8	9	9	9
Cleanliness	1/2	1	3	4	4	9	9	9
Staff	1/6	1/3	1	1	4	3	6	9
Price/benefit	1/8	1/4	1	1	2	9	4	8
Comfort	1/8	1/4	1/4	1/2	1	5	2	9
Position	1/9	1/9	1/3	1/9	1/5	1	3	6
Wi-fi	1/9	1/9	1/6	1/4	1/2	1/3	1	2
Parking	1/9	1/9	1/9	1/8	1/9	1/6	1/2	1

The decision maker does not change his/her opinion about the first 6 alternatives, but simply integrates the judgments concerning the new ones. The behaviour is rational, but the *CR* index in this case is 0.117, showing as, using the classical thresholds, it is not easy to obtain matrix that can be considered consistent. In this case, according to classical procedures, the decision maker has to revise his/her judgments, but generally he/she is anonymous, so it is impossible to contact him/her. Another solution could be to consider the linearization technique we introduced in section 2. We instead propose to evaluate the level of inconsistency using the new bounds. They can be useful to decide if the level of inconsistency is too high or if the matrix can be used to calculate the prioritization vector. Only if the level of inconsistency is too high the matrix would not be taken into account for the analysis.

5 Concluding remarks

In this paper we propose a procedure to derive new consistency thresholds based on the percentiles rule. When the values of the indices measuring the consistency are higher than the classical thresholds, the decision makers are asked to revise their judgments. This happens frequently when the number of alternatives is higher

than six and the existing bounds should lead to revise almost 100% of the judgments. However sometimes decision makers are not achievable, for example when the questionnaire are anonymous, or they are unwilling to revise their judgments. This happens often in evaluation of tourism services. In this case a possible solution is to force the matrix to be consistent through an orthogonal projection in a given linear space. Our idea instead is to introduce new bounds that can be useful to define the level of inconsistency. The analyst can check if the matrix has a tollerable level of inconsistency and he can decide to keep it to calculate the priority vector. Only when the inconsistency exceeds also these thresholds it is necessary to use some other methods.

References

1. J. Aguaron, M. Escobar, J. Moreno-Jimenez, The precise consistency consensus matrix in a local ahp-group decision making context, *Annals of Operational Research* 245 (2014) 245-259.
2. J. Aguaron, J. Moreno-Jimenez, The geometric consistency index: Approximated threshold, *European Journal of Operational Research* 147 (2003) 137–145.
3. A. Altuzarra, J. Moreno-Jimenez, M. Salvador, Consensus building in ahp-group decision making: A bayesian approach, *Operations Research* 58 (2010) 1755–1773.
4. J. Benitez, X. Delgado-Galvan, J. Izquierdo, R. Perez-Garcia (2011). Achieving matrix consistency in AHP through linearization.*Applied Mathematical Modelling* 35: 4449–4457.
5. S. Bozóki, T. Rapcsák, On Saaty's and Koczkodaj's inconsistencies of pairwise comparison matrices, Journal of Global Optimization 42 (2008) 157–175.
6. G. Crawford, C. Williams, A note on the analysis of subjective judgment matrices, *Journal of Mathematical Psychology* 29 (1985) 387–405.
7. Z. Duszak, W. Koczkodaj, Generalization of a new definition of consistency for pairwise comparisons, *Information Processing Letters* 52 (1994) 273–276.
8. S. Gass, T. Rapcsák, Singular value decomposition in ahp, *European Journal of Operational Research* 154 (2004) 573–584.
9. A. Grzybowski, Note on a new optimization based approach for estimating priority weights and related consistency index, *Expert Systems with Applications* 39 (2012) 11699–11708.
10. A. Grzybowski, New result on inconsistency indices and their relationship with the quality of priority vector estimation, *Expert Systems with Applications* 43 (2016) 197–212.
11. A. Ishizaka, A. Labib, Review of the main developments in the analytic hierarchy process, *Expert Systems with Applications* 38 (2011) 14336–14345.
12. W. Koczkodaj, A new definition of consistency of pairwise comparisons, *Mathematical and Computer Modelling* 18 (1993) 79–84.
13. R. Narasimhan, A geometric averaging procedure for constructing supertransitivity approximation to binary comparison matrices, *Fuzzy Sets and Systems* 8 (1982) 53–61.
14. J. Peláez, M. Lamata, A new measure of consistency for positive reciprocal matrices, *Computer and Mathematics with Applications* 46 (2003) 1839–1845.
15. R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria (2014). url: http://www.R-project.org/
16. T. Saaty, *Multicriteria Decision making: The Analytic Hierarchy Process* (McGraw-Hill, New York, 1980).
17. T. Saaty, *Fundamental of decision making and priority theory with the AHP* (RWS Publications, Pittsburgh, 1994).
18. T.L. Saaty (2003) Decision-making with the AHP: why is the principal eigenvector necessary. European J. Oper. Res. 145: 85–91.
19. A. Salo, R. Hamalainen, On the measurement of preference in the analytic hierarchy process, *Journal of Multi-Criteria Decision Analysis* 6 (1997) 309–319.

Mines and quarries: An analysis of withdrawals determinants in Italy

Cave e Miniere: un'Analisi delle Determinanti dell'Estrazione in Italia

Sabrina Auci and Donatella Vignani

Abstract The demand of no-energy mineral resources is still growing. This paper aims to disentangle the drivers of raw materials extraction in Italy. Using a new dataset of raw resources extracted from mines and quarries (m&q) in the 21 Italian regions for the period 2013-2016, results confirm the relevance of m&q price index as well as manufacturing and construction sectors as the main drivers and a positive relationship between extraction intensity and price index in line with Hotelling (1931)'s theory.

Abstract *La domanda di risorse non energetiche è tuttora crescente. Questo studio si propone di verificare i driver dell'estrazione di materiali da cave e miniere (m&q) in Italia. Utilizzando un nuovo dataset di risorse naturali estratte per le 21 regioni italiane nel periodo 2013-2016, si conferma come drivers sia l'indice dei prezzi alla produzione di m&q che i settori manifatturiero ed edilizio, oltre a una relazione positiva tra l'intensità dell'estrazione e l'indice dei prezzi come in Hotelling (1931).*

Keywords: non-renewable natural resources, mining and quarrying minerals statistics, m&q price index, Hotelling theory

[1] Donatella Vignani, Istat, Italian National Institute of Statistics, Environmental and Territorial Statistics Directorate; email: vignani@istat.it

Sabrina Auci, Department of Political Science and International Relations, University of Palermo; email: sabrina.auci@unipa.it

1 Introduction

Since 1980, the amount of materials extracted has constantly increased. The OECD countries have shown a lesser increase of material extraction than the rest of the world but the 35% of the total material resources extraction takes place in Europe (OECD, 2013). Because the growth rate of the extraction is mainly driven by construction sector, the socio-economic structure puts pressure on natural resources and the environment. Since the '60s, even in Italy a considerable amount of non-renewable mineral resources has been extracted.
The aim of this paper is to disentangle the drivers of raw non-energy materials extraction in Italy. The relevance of a supply curve between mineral quantity extracted by mining and quarrying (m&q) and the domestic m&q producer price index is analyzed by controlling for the openness effect, the composition effect and the territorial effect. The scale effect is irrelevant because of the presence of material extraction decoupling issue for the European countries.
Two are the reasons of preferring the material extraction to the material use. First, raw materials supply derives from a choice on dynamic exploitation of m&q following Hotelling (1931)'s theory. Second, the rate of raw material extraction has important consequences on natural resource stocks, on environmental pressure and on international trade and market prices.
As natural non-renewable resources are concerned, the theory of Hotelling (1931) is prevailing. Any withdrawal of an exhaustible natural resource implies an irreversible reduction of the available stock. Considering the dynamic efficiency in the exploitation of a natural exhaustible resource, the optimal quantity to be extracted will be provided by the market price of the resource. With declining resource availability, the resource price increases over time as the rate of extraction declines.
The remainder of the paper is organized as follows. Section 2 describes the methodology followed and the data collected. Section 3 presents the empirical model and the main results. Section 4 concludes.

2 Data Description and Methodology

Understanding the main drivers of no-energy material supply is a necessary precondition to a sustainable resource management. Very few studies have focused on the drivers of material extraction and to the aim of this analysis, only two papers are relevant. First, Nyambuu and Semmler (2014) shows that whenever the resource reserves are large, the path of the extraction rate and price will monotonically decrease and rise respectively. The study of Menegaki and Kaliampakos (2010) instead, focusing on 26 European countries, find the relevance of total construction and residential building sector as drivers of aggregates production.
Our analysis sheds light on the supply of m&q mineral resources extracted in Italy. As in Auci et al. (2013), our analysis focuses on the no-energy materials extracted in the 21 Italian regions for the period 2013-2016. A new official statistic has been

employed for the dependent variable: the m&q mineral resources extracted intensity indicator. This dataset on no-energy producing minerals at regional level has been collected by Istat (Vignani et al., 2019) through the Survey Anthropic Pressure and Natural Risk since 2015 included in the National Statistical Program, as a statistical information relevant for the country. These new official statistics have been produced for the period spanning from 2013 to 2016 on a yearly base. One of the advantages of this new collection of data consists in better identifying the physical dimension and the environmental pressure of the m&q phenomenon in Italy. The other variables are collected using Environmental Satellite Accounts of National Accounting for material flows statistics, while several surveys – Survey on International Trade, Survey on Production prices, Survey on Industrial Production PRODCOM – and National Accounts are employed for economic statistics. All data are collected yearly.

In 2016, in Italy m&q raw minerals extraction with the exclusion of mineral waters was 167.8 million tons of which 154.2 from quarries and 13.7 from mines. The Extraction Intensity Indicator (IE), which is the ratio between quantities of m&q resources extracted and regional administrative areas, is calcolated on yearly base. In 2016, on average the extraction intensity for Italy is equal to 556 tons per Kmq. The IE indicator calculated at territorial level show that the municipalities with m&q sites in production are 1,224 of which the majority are middle and high extraction intensity, principally located in the North (Vignani et al., 2019).

Using an instrumental variable random effects (IVRE) model, we estimate the impact of m&q producer price index on minerals resources extraction intensity, considering the size of the construction, manufacturing and m&q sector. All these variables are measured in terms of value added (VA). Thus, controlling for time-invariant differences between regions, the random effects model is applied. Because the m&q producer price index is endogenous, the estimation is based on the two-stage least-squares random effects estimator as follows:

$$m\&q\ ext\ inten_{it} = \alpha_i + \alpha_1 m\&q\ prod\ price\ index_{it} + \alpha_2 VA\ Const\ sector_{it} + \alpha_3 VA\ Manuf\ sector_{it} + \alpha_4 VA\ m\&q\ sector_{it} + \mu_i + v_{it}$$

and the endogenous model is:

$$m\&q\ producer\ price\ index_{it} = \beta_1 Openness_{it} + \beta_2 m\&q\ exchange_{it} + u_{it}$$

Where the degree of openness to EU countries in monetary terms and the degree of national exchange of mineral materials with the rest of the world in physical terms are the instruments to control for endogeneity.

3 Empirical Analysis and Results

In Table 1, our main results are reported. Three different models are compared for robustness analysis. In Model IV1 is considered only the m&q producer price index, while in Model IV2 and Model IV3 are added the VA of m&q sector together with respectively the VA of construction sector and manufacturing sector.

Table 1: IV estimations – random effects model

Dep. Var.: m&q Extraction intensity	***Model IV1***	***p-value***	***Model IV2***	***p-value***	***Model IV3***	***p-value***
m&q prod. price index	0.438***	(0.006)	0.303*	(0.099)	0.514***	(0.004)
VA m&q sector			-0.044	(0.481)	-0.028	(0.638)
VA constr. sector			0.413***	(0.003)		
VA manuf. sector					0.316***	(0.001)
Constant	4.068***	(0.000)	1.757	(0.132)	1.167	(0.329)
N. of obs.	84		84		84	
N. of Regions	21		21		21	
F-test	7.948		6.092		7.040	

P-values in parentheses; *** *p<0.01,* ** *p<0.05,* * *p<0.1*
The results of the second equation are available upon request.

Our findings suggest that m&q producer price index as well as manufacturing and construction sectors are the main drivers of m&q extraction intensity in line with Menegaki and Kaliampakos (2010). Moreover, Hotelling (1931)'s prediction is confirmed: the price index monotonically rises. Finally, the relevance of openness and exchange degree as instrument of m&q producer price index is proven.

4 Conclusions

This research contributes to the literature on the drivers of m&q raw materials extraction in threefold. First, a new dataset on no-energy producing materials from m&q has been collected. Second, the supply of raw materials extracted has been analyzed along with other relevant determinants. Third, the endogeneity problem of m&q producer price index is considered. Results confirm the relevance of m&q producer price index as well as manufacturing and construction sectors as the main drivers. Moreover, a positive relationship between extraction intensity and price index in line with Hotelling (1931)'s prediction for non-renewable resources whose reserves are large and well-known is proven.

References

1. Auci, S., Castellucci, L., Vignani, D.: Imposte e Governance Regionale in Materia di Cave e Miniere in Italia. In Ficari, V., Scanu, G. (eds.) Tourism taxation. Sostenibilità ambientale e turismo fra fiscalità locale e competitività, pp. 214-232. Giappichelli, Torino (2013).
2. Hotelling, H.: The economics of exhaustible resources. J. Polit. Econ. 39(2), 137–175, (1931).
3. Menegaki, M.E., Kaliampakos, D.C.: European aggregates production: Drivers, correlations and trends. Res. Pol., 35, 235–244, (2010)
4. Nyambuu, U., Semmler, W.: Trends in the extraction of non-renewable resources: The case of fossil energy. Econ. Mod., 37, 271–279, (2014)
5. OECD: Material Resources, Productivity and the Environment, Organization for Economic Development and Cooperation, Paris (2013).

Mines and quarries: An analysis of withdrawals determinants in Italy
6. Vignani, D., Budano, F., Busetti, C.: Le attività estrattive da cave e miniere 2015-2016. Statistica Report Istat, Roma (2019)

Evaluating people's behaviour towards risk: a multidimensional problem

Valutazione del comportamento individuale verso il rischio: un problema multidimensionale

Luigi Bollani, Guido Antonio Rossi and Ivan Sciascia

Abstract Behaviour towards risk is very hard to assess. It depends on the various environmental situations an individual is subjected to and on the internal characteristics of the subject who has to decide.
An empirical investigation has been carried out considering for the same subjects their behaviour regarding simulated financial choices, their perception of some human values and lifestyles related to sensation seeking, a soft investigation about their personality traits and some socio-demographic information.
A multivariate analysis was performed to jointly encompass the main features of this complex framework.
Abstract *Il comportamento individuale verso il rischio è di difficile valutazione. Esso infatti dipende dalle diverse condizioni di contesto in cui l'individuo si trova a decidere, oltre che dalle sue caratteristiche personali.*
Si è condotta un'indagine empirica dove, per ciascun soggetto, si è considerato il comportamento riguardo a scelte finanziarie simulate, la percezione di alcuni valori personali e stili di vita collegati alla ricerca di sensazioni, un inquadramento di massima su alcuni tratti di personalità e alcune caratteristiche socio-demografiche.
È stata utilizzata un'analisi multidimensionale per porre in evidenza le principali connessioni tra le caratteristiche di questo complesso quadro di riferimento.

Key words: expected utility, rational behaviour, values (vals), sensation seeking, colour test, empirical survey, multivariate analysis

[1] Luigi Bollani, ESOMAS Department, University of Turin; email: luigi.bollani@unito.it
Guido Antonio Rossi, ESOMAS Department, University of Turin; email: guido.rossi@unito.it
Ivan Sciascia, Department of Life Sciences and Systems Biology, University of Turin; e-mail: ivan.sciascia@unito.it

1 Aim and literature background of experimentation

The purpose of this paper is to empirically analyse individual risk attitude, considering a multidimensional approach. An experimental investigation was carried out taking in account - for the same individuals - their behaviour in financial choices, their perception of some human values and lifestyles related to sensation seeking, a soft investigation about their personality traits and some socio-demographic information.

So far, the assisted administration of a questionnaire has involved a sample of 190 university students, with some basic knowledge on probability and willingness to devote 20-30 minutes to answer.

For financial risk attitude analysis, evaluated considering individual preference among lotteries - intended to be random amounts of money with an associated probability - a model discussed in Bollani and Rossi (2005) is used; it empirically develops some more general papers (i.e. Rossi, 1994a, 1994b). The lotteries are only proposed in comparison, i.e. the considered amounts of money are in fact fictitious and not really won or lost. Furthermore, these sums are not intended to be very relevant to the decision maker.

For the individual importance attributed to some human values, the Rokeach's survey (Rokeach, 1973) was considered as a starting point. It is based on a questionnaire that asks to rank some human values: a first sequence of 18 items considered to be "terminal values" and, separately, a second sequence of 18 items introduced as "instrumental values" are proposed.

Lifestyles related to "sensation seeking" investigation is proposed, taking inspiration from the SSS-V Zuckerman Sensation Seeking Scale (Zuckerman, 1994). The scale is determined by a synthesis of four components: thrill and adventure seeking (*Tas*), Experience Seeking (*Es*), Disinhibition (*Dis*) and boredom susceptibility (*Bs*); each dimension is investigated using ten items.

To explore some individual personality traits, the Lüscher test was taken in account as a reference point. This test is used to measure the psychological state of a person, describing both personality and emotions. In its short version (Lüscher, 1969), the subject under examination is asked to put in preference order (the same operation is requested twice) the following eight colours: dark-blue, blue-green, red-orange, bright yellow, violet, brown, black, grey. The first four colours indicated before - which are then called primary colours - are often chosen as first or second place in the preference sequence. In a long experimentation - some years around 1950 - Lüscher defined the linkage between the most frequent preference sequences and the corresponding psychological characteristics of the person involved.

2 Methodological background used to examine financial risk attitudes

In our experimentation, individual preference among lotteries is examined and the following method is used to evaluate the decision maker's financial risk attitude. The result of this evaluation determines if the decision maker is risk averse, risk

prone or incoherent (i.e. sometimes averse and sometimes prone or not respecting some circularities in choices).

We start to consider the concept of certainty equivalent x_j - for lottery j - that is determined by the formula

$$x_j = f^{-1}(\sum_{i=1}^{n} P(a_{i,j}) f(x(a_{i,j})))$$

where a monetary value referred to alternative $a_{i,j}$ (i.e. the *i-th* alternative for lottery j) is denoted by $x(a_{i,j})$, while its probability is denoted by $P(a_{i,j})$; moreover f represents the decision maker's risk attitude. In fact, as in Rossi (1994a), the function f - which is usually called utility function - can actually be decomposed into a function that indicates the attitude towards risk and another one that indicates the reaction to a (smaller or larger) sum of money. If, as assumed in this paper, we refer to sums that have a small but non null relevance to the decision maker, the second function becomes an identity and may be not considered: in this way, f only represents the decision maker's attitude towards risk.

Moreover, to evaluate the individual risk attitude a polynomial is used (again, as in Rossi 1994a), in order to express choices among lotteries in terms of comparisons between their moments. In the current experimentation, f is assumed to be a third degree polynomial because it allows to transform expressed preferences into comparisons between means, variances and asymmetries.

We recall that f is unique up to a linear affine transformation. This property is used to reduce the degrees of freedom of the polynomial (variable coefficients) from four to two: that of the first order term (mean) and that of the second order term (second moment from zero), being 0 the constant term and 1 the leading coefficient.

Furthermore, the choice of a subject between two lotteries, which shows his/her attitude towards risk, can be viewed as the choice of a part of a plane - having the two coefficients as coordinates - divided into two halves by a straight line. Thus, repeated choices correspond to a progressively specified part of the plane, that may be included in a reference area of risk aversion, defined by $f' > 0$ and $f'' < 0$ (Bollani, Rossi, 2005).

In the same way, we could find similar areas - as some example those indicating proneness or incoherence towards risk, or other - on the plane and intercepting the previous ones, thus revealing different characteristics of choices.

3 Survey structure and information acquired from each analysis perspective

As already said, a new empirical investigation, with a new questionnaire (both different from previous analyses), was designed and administered specifically for this paper. The questionnaire was divided into five parts, each of them corresponding to a different analysis perspective. They are: submitting choices among lotteries; questions on the importance of different human values; questions on some life style attitudes mainly oriented at “sensation seeking”; administration of

colour choices (as a soft investigation on some personality traits); questions to gather socio-demographic information.

Each part of the questionnaire (except the self-evident one concerning socio-demographic data) is taken into account in the following subsections and - for each - a brief description of the main information obtained is considered.

3.1 Choices among lotteries

In administering financial choices among lotteries, each lottery was prepared considering a finite set of amounts (up to three) and their connected probabilities, in order to simplify comparisons. In some cases, negative amounts have also been included.

In our experimentation, a classification of subjects was carried out - based on a complete comparison of a set of lotteries - using the method explained in section 2.

This classification determines the following groups:

- risk averse (108 subjects, i.e. 56.84% of the sample),
- moderately risk prone (31 subjects, i.e. 16.32% of the sample),
- very risk prone (47 subjects, i.e. 24.74% of the sample).

The complete comparison among lotteries has avoided finding incoherence situations, however a residual group of 4 subjects did not give complete answers and was not classified. No subject resulted to be risk neutral.

3.2 Importance given to human values

About the importance given to different human values, a preliminary discussion group was formed to adapt the questionnaire used in Rokeach's survey. The group considered the necessity to reduce in number the original 36 values, because the human values evaluation represented only one of the themes embedded in our questionnaire. Moreover, the relevance for group members - because of their reflections about most influent human values - of new different items was considered too and some of them were added.

At last, a set of 15 values was chosen and a Likert scale in seven points was decided to be used. So, a question like "How much important do you feel each of the following aspects for your life? (answer must use a scale from 1 - not important at all - to 7 - absolutely important)" was asked for the following human values: health, beauty, male and female roles, work, possibility of spending, free time, relations with the partner, relationships with the family of origin, relations with friends, economic security, culture, politics, economics, science, physical activity.

A principal component analysis (PCA) was performed followed by a hierarchical cluster analysis (HCA), using - on the most important dimensions of the PCA - the square Euclidean distance as a similarity measure and the Ward method to group sample units by affinity of reply.

Three clusters were determined:

- cluster 1, consisting of people who pay particular attention to their private life context (71 subjects, i.e. 37.37% of the sample),
- cluster 2, consisting of people more oriented to social and economic security (66 subjects, i.e. 34.74% of the sample),
- cluster 3, consisting of people more oriented to social and cultural development (53 subjects, i.e. 27.89% of the sample).

3.3 Importance given to sensation seeking

The attitude towards "sensation seeking" among the interviewees was evaluated starting from the aforementioned studies of Zuckerman. As before, a preliminary discussion group was involved to determine a series of alternatives, introduced into our questionnaire using the semantic differential technique (Snider and Osgood, 1969).

At last, the following ten alternatives were introduced in the questionnaire: work with limited mobility vs work with high mobility; work stable and ordered vs work creative and with unexpected events; spending time with familiar people vs spending time with new people; going in organized trips vs going in trips without organization; accepting to be hypnotized vs not accepting to be hypnotized; accepting a parachute launch vs not accepting a parachute launch; listening familiar musical groups vs listening new musical groups; preferring reliable and predictable friends vs preferring exciting and unpredictable friends; preferring a comfortable hotel vs preferring a tent under the stars; feeling art as clarity, symmetry and harmony of colours vs feeling art as chaos, irregular shapes and dissonance of colours.

Given that each alternative gave rise to a difference from the neutral (central) point in a seven-point scale (resulting in a quantitative variable from -3 to +3, starting from a low "sensation seeking" situation towards a high one), again a PCA followed by an HCA was performed.

Three groups of interviewees were identified, as follows:

- rather low degree of "sensation seeking" (62 subjects, i.e. 32.63% of the sample),
- medium degree of "sensation seeking" (44 subjects, i.e. 23.16% of the sample),
- rather high degree of "sensation seeking" (84 subjects, i.e. 44.21% of the sample).

3.4 Colour choices

Finally, in order to examine some personality traits, the Lüscher Colour Test (in its short version) was administered in part (skipping the second run and not showing

colour charts). In particular, as already said, each respondent was asked for a ranking of eight colours: i.e. the ones we now shortly call blue, green, red, yellow, purple, brown, grey, black.

Only the first two positions have been considered, which are very often chosen among the primary colours: blue, green, red and yellow.

Following the first dimension introduced by Lüscher (the so-called "constellation"), a first contraposition is considered: the choice of blue and yellow (in the first two positions) is more typical of a receptive person, while red and green denote directive attitudes.

A second dimension ("continuity") is based on the contraposition between blue/green, which are associated to a constant trait, and yellow/red which are associated to a variable one.

A third dimension ("communication") considers the contraposition between blue/red, which are associated to an integrative attitude, and green/yellow, which are associated to a separative attitude

Considering the only situations described above the following classification of respondents found in our experiment is shown:

- receptive (8 subjects, i.e. 4.21% of the sample),
- directive (11 subjects, i.e. 5.79% of the sample),
- constant (32 subjects, i.e. 16.84% of the sample),
- variable (11 subjects, i.e. 5.79% of the sample),
- integrative (28 subjects, i.e. 14.74% of the sample),
- separative (this category was skipped, because of the only 2 subjects belonging to it)

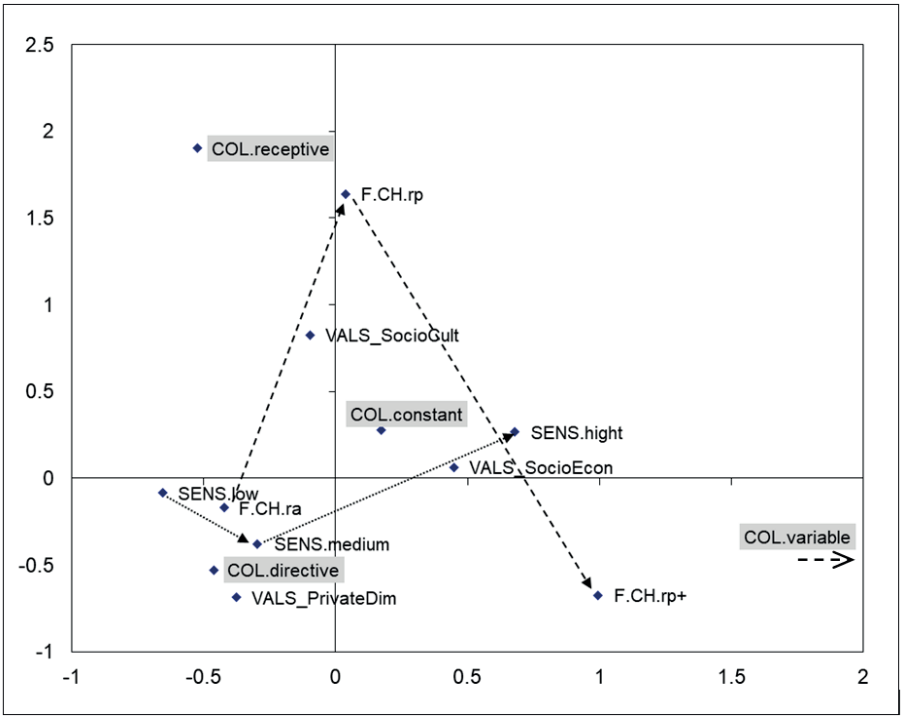
It is to be noted that 98 subjects (51.58% of the sample) resulted not classifiable in the above categories.

4 Combining different analysis perspective

The following output - in figure 1- of a multiple correspondence analysis (MCA) shows some relationships among the items of the variables observed through the different perspectives of investigation.

MCA (graphically representing the underlying well-known Burt scheme) was performed using "FactoMineR" R package (see Escofier, Pagès, 2008).

Figure 1: Relationships among different perspectives of investigation



Following the first dimension (*36.1%*, using Greenacre's revaluation in "ca" *R* package (see Nenadic, Greenacre, 2007 and Greenacre, 2017) from left to right, it is possible to identify a common direction for all the survey perspectives from risk aversion to risk proneness.

Considering financial choices, a path that starts from risk aversion (F.CH.ra), moving towards a moderate risk proneness (F.CH.rp) and a high-risk proneness (F.CH.rp +) is shown.

In the same direction the attitudes towards "search for sensation" grow (following the path SENS.low, SENS.medium, SENS.hight).

Considering human values, those people who pay more attention to a private dimension (VALS_PrivateDim) are shown on the left (i.e. on the risk aversion side); going towards the right, we find represented the socio-cultural interests (VALS_SocioCult) and beyond the socio-economic sensitivity (VALS_SocioEcon).

Considering chromatic choices, there is a minimal difference between the traits of receptive and directive personalities with respect to the attitude towards risk (given that they have almost the same value on the first dimension), while a difference appears between constant and variable personality traits: the latter being decidedly on the side of risk proneness (outside the graph).

It appears that there is some concordance between the various characteristics here examined: risk attitude in financial choices, importance given to human values, to sensation seeking, preference for colours.

These results encourage consideration of approaches to examine a subject's risk attitude, jointly described by financial choices and other choices revealing personal characteristics, and furthermore provide indications for the usefulness of some new and diverse analysis perspectives.

References

1. Bollani, L., Rossi G.A.: Inquiring into decision makers' qualities. In Rossi G.A. (ed.), Changing Models, pp. 25-50. Levrotto & Bella, Torino, 2005.
2. Escofier, B.; Pagès, J.: Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation; Dunod: Paris, France, 2008.
3. Greenacre, M.: Correspondence Analysis in Practice; Chapman and Hall/CRC: New York, NY, USA, 2017.
4. Lüscher M.: The Lüscher Color Test. Scott L.A., Trans. & Ed. New York: Random House. (1969).
5. Nenadic, O.; Greenacre, M.: Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. J. Stat. Softw. 2007, 20.
6. Rokeach M.: The nature of human values. New York, NY, US: Free Press. (1973).
7. Rossi, G.A.: Rational behaviour: a comparison between the theory stemming from the de Finetti's work and some other leading theories. Theory and Decision, 36, pp. 257-275 (1994a).
8. Rossi, G.A.: About explaining in decision theory. Rivista internazionale di scienze economiche e commerciali, 41, pp. 997-1012 (1994b).
9. Snider, J.G., Osgood, C.E.: Semantic Differential Technique: A Sourcebook. Aldine, Chicago (1969).
10. Zuckerman M.: Behavioral Expression and Biosocial Bases of Sensation Seeking, Cambridge University Press, New York (1994).

Inferential versus descriptive statistical approach in the analysis of Delphi performance: A case study.

Analisi delle performance del Delphi. Approccio inferenziale verso descrittivo: studio di un caso

Bolzan M., Auciello M., Pesarin F

Abstract This study aims at evaluating, by permutation methods, the performances of Delphi approach in the research to predict the future of the family in NorthEast of Italy in ten years. The usual descriptives indicators are: stability, consensus and convergence speed. In the work we intend to test – by permutation methods -three equivalent distinct statistical hypotheses: equality, convergence and combination.
Abstract *Questo studio mira a valutare, con metodi di permutazione le performance dell'approccio Delphi nella ricerca per predire il futuro della famiglia nel Nord-Est dell'Italia tra dieci anni. Gli indicatori descrittivi usuali sono: stabilità, consenso e velocità di convergenza. Nel lavoro intendiamo verificare - con metodi di permutazione - tre ipotesi statistiche distinte equivalenti: uguaglianza, convergenza e combinazione.*

Key words: Delphi Approach, performance Indicators; Peermutation methods ...

1 Introduction

The Delphi method is considered by many scholars to be the father of methods that are useful in participatory social research and for the construction of future scenarios on themes that, by nature, do not lend themselves to be analysed by traditional

Bolzan Mario
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: mario.bolzan@unipd.it

Auciello Massimiliano
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: massimilianoauciello@hotmail.it

Pesarin Fortunato
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: fortunato.pesarin@unipd.it

quantitative approaches. It is set up as a multi-interview survey, carried out through a number of rounds with experts or, more generally, with privileged witnesses, who provide a series of opinions on the subject of the research [3, 2, 4, 1]. The iterativity of the Delphi procedure allows a certain degree of consensus to be reached amongst experts and facilitates a comparison and mutual exchange of knowledge whilst allowing the individual respondents to re-evaluate their positions and beliefs up to their adequate and acceptable convergence. The performance indicators of Delphi applications are measures of stability, consensus and convergence speed. With stability, we denote the situation in which the results of two successive Delphi interviews are not consistently different; in fact, this property can be used as a criterion to stop the procedure. It is measured by regression coefficients calculated on the values of the first and third quartiles of the last two rounds. The closer these two coefficients S_1 and S_3 are to zero, the more the evaluations of the panel of experts can be considered stable. The consensus measures the convergence of opinions, so if the process leads to a final interquartile range which is small enough (e.g. 20% of the domain), consensus is considered to have been reached. The final level of consensus expresses a static aspect that evaluates only the last round regardless of the stability of the answers. It is measured by the width of the last interquartile range, denoted with IQ. The convergence speed or the velocity of convergence (V of C), instead, evaluates the dynamic aspect of the process and is measured by the coefficient of variation (CV), calculated over all the rounds. The more these coefficients decrease towards zero, the greater the V of C is. The research question to which we intend to offer some answers of various levels of completeness is as follows: Is the information produced by the descriptive performance indicators of the Delphi process sufficient, or can the integration of inferential statistical instrumentation offer additional useful and pertinent information? The descriptive empirical approach has the considerable advantage of offering indications based on the contingent dimension of the phenomenon under study and on the basis of shared parameters; the inferential one provides information on the risk of false negatives in which conclusions of general value are drawn. In the case of Delphi surveys, the so-called *minimal sufficient statistics is the whole data set*. Then there is no one-dimensional statistic able to summarize the entire set of information contained in the data. To gather this information as best as possible, a plurality of statistics would be necessary. Indeed, if the data are n, the entire set of information is necessarily gathered from no less than n indicators; whose number, therefore, increases as n. It follows that with a smaller number of statistics one can only aspire to an incomplete (insufficient) and approximate representation with consequent loss of information. It should also be kept in mind that the group of experts involved in the survey is not constructed according to the criteria of representativeness produced by a probabilistic selection of the sample. Therefore, the data (the opinions expressed on the individual items) cannot be analysed with classical statistical methods (parametric tests); the application of non-parametric methods is instead required. The work concerns a survey using the Delphi approach conducted in 2016-2017 on the topic *tomorrow in the family* in Northeast Italy. It involved a group of 32 experts selected amongst professionals and scholars in the areas of more specific interest in the study of the family. The

objective is to derive predictions based on the convergence of their opinions on the evolution of phenomena expressed by items (see Table 1) which reflect - specifically in this article - the conditions of parent' lives and which imagine a society and family placed in a future that is sufficiently advanced (10 years) [1].

2 Related statistical methodology

The methodological problem connected with Delphi data, due to its complexity, will be subdivided into some sub-problems in accordance with the hypotheses that are of interest to be analyzed with data observed at two or more time occasions.

The first response model for the analysis concerns a given variable Y observed at two different time occasions, $t = 1,2$, on subject i. In such a context we assume that responses behave on:

$$Y_{1i} = \mu + \eta_i + Z_{1i} \quad \text{and} \quad Y_{2i} = \mu + \eta_i + \Delta_i + \sigma_i Z_{2i},\ i = 1,\ldots,n, \quad (1)$$

where: a) μ is a population constant; b) η_i represents the effect due to the set of covariates specific to subject i (e.g. competence, experience, skillness, and so forth), either observed or not; c) Z_{1i} and Z_{2i} are the so-called error components (*natural deviates*) specific of the adopted instrument, errors that are assumed to be independent between subjects but possibly not time independent within a subject; d) σ_i is the dispersion coefficient specific to subject i: it is > 1 if there is a divergence from occasion *t=2* with respect to that at *t=1*; it is < 1 if, vice versa, there is convergence; e) Δ_i are the Delphi effect descibing how subject i changes its response between two time observations and, although with different meanings, it stands both for *evolution* and *relevance*: it is positive if second response is stochastically larger than the first, otherwise it is negative.

Of course, all such coefficients (parameters) are unknown to the analyst and $[(\eta_i, \Delta_i, \sigma_i), i = 1,\ldots,n]$, three from each subject and each observed response, are to be analyzed. So, there are much more unknown parameters (about 3780) than there are observed subjects ($n = 32$). This implies the necessity for the statistical analysis of taking recourse to nonparametric approaches, because the parametric ones are absolutely impossible.

Considering response differences between two occasions, the model becomes: $Y_{2i} - Y_{1i} = \Delta_i + \sigma_i Z_{2i} - Z_{1i},\ i = 1,\ldots,n$. It is worth noting that this model results independent of constant μ and individual effects η_i; it depends on effects Δ_i and σ_i. To test for such coefficients, nonparametric permutation solutions for their *evolution* , i.e. for their *stability*, the hypotheses under analysis are $H_{0U} : (\Delta_i = 0,\ i = 1,\ldots,n)$ against $H_{1U} : (\exists \Delta_i \neq 0)$, and for their *convergence* (concentration) the hypotheses are $H_{0C} : (\sigma_i = 1,\ i = 1,\ldots,n)$ against $H_{1C} : (\exists \sigma_i \neq 1)$.

However, it is of particular importance to see, separately for each response variable or group of variables, wether the evolution effects, instead of merely different from zero, are positive (i.e. $\Delta > 0$) or negative (i.e. $\Delta < 0$), and those of convergence, instead different from one, are larger ($\sigma > 1$) or smaller ($\sigma < 1$) than stabil-

ity, while identifying for both the direction. This requires that the related alternatives must be written as $H_{1U} : [(\exists \Delta_i < 0) \bigcup (\exists \Delta_i > 0)]$ and $H_{1C} : [(\exists \sigma_i < 1) \bigcup (\exists \sigma_i > 1)]$, respectively. Thus, the statistical problem must imply two separate tests for each variable, one for each of two aspects into which the alternatives are broken-down.

For the permutation tests with paired data we refer to the book by Pesarin and Salmaso (2010, pg. $13 \div 23$). That is: $T_L^* = \sum_{1 \le i \le n} [Y_{Li}(t_2) - Y_{Li}(t_1)] \cdot S_i^*$ where S_i^* is a random permutation of equally likely signs $(-1, +1)$ with $L = U$ or C.

So, the response structure of such a test for evolution is: $T_U^* = \sum_i (\Delta_i + \sigma_i Z_{Bi} - Z_{Ai}) S_i^* = \sum_i \Delta_i S_i^* + \sum_i (\sigma_i Z_{Bi} - Z_{Ai}) S_i^*$. From this structure we see that: a) T_U^* essentially depends on coefficient Δ: if $\Delta > 0$ the observed value of the test is stochastically larger than that under H_0, vice versa if $\Delta < 0$; b) the error component $\sum_i (\sigma_i Z_{2i} - Z_{1i}) S_i^*$ is distributed around zero. So, when $\sigma_i = 1$, $i = 1, \ldots, n$, and error components Z_{2i} and Z_{1i} are symmetrically distributed around zero, such a test is exact, i.e. it exactly controls first kind error rate. Otherwise the test is exact only asymptotically. A simulation study with error components strongly asymmetric has shown that, with sample sizes of $n = 30$ the test is practically exact.

Since the exact determination of the permutation distribution of test T^* it is necessary to consider all the $4.295 \cdot 10^9$ possible permutations of signs, to estimate its p-value it is usual to consider R random permutations. Such p-value estimates are $\hat{\lambda}_U^{>} = [\#(T_U^* > T_U^{oss}) + \frac{1}{2}\#(T_U^* = T_U^{oss})]/R$ and $\hat{\lambda}_U^{<} = [\#(T_U^* < T_U^{oss}) + \frac{1}{2}\#(T_U^* = T_U^{oss})]/R$, respectively, where T_U^{oss} is the observed value.

Of course, the global p-value is then $\hat{\lambda}_U = \min(\hat{\lambda}_U^{<}, \hat{\lambda}_U^{>})$ to be compared with $\alpha/2$ if one wants that the global first kind error rate is α. And if it results that $\hat{\lambda}_U^{<} \le \alpha/2$, then one concludes that data behavior stochastically conforms according to the alternative $H_{1U}^{<} : (\exists \Delta_i < 0)$ at α level, and so having identified both the presence of non-null effects and their direction.

Regarding the test on convergence, to put due emphasis on response variations around a suitable central point, i.e. $|Y_{1i} - \tilde{Y}_1|$ and $|Y_{2i} - \tilde{Y}_2|$, $i = 1, \ldots, n$, respectively, it is worth noting that the related response models are: $|Y_{1i} - \tilde{Y}_1| = |\eta_i - \tilde{\eta} + Z_{1i} - \tilde{Z}_1|$ and $|Y_{2i} - \tilde{Y}_2| = |\eta_i - \tilde{\eta} + \Delta_i - \tilde{\Delta} + \sigma_i (Z_{2i} - \tilde{Z}_2)|$. Thus, the test statistic to take into consideration is: $T_C^* = \sum_i \left[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1| \right] \cdot S_i^*$; the observed value of which is: $T_C^{oss} = \sum_i \left[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1| \right]$. A specific simulation study, carried out to find the most suitable central point providing the best approximation for the null distribution, this results that is the sampling median: $\tilde{Y}_j = Med(Y_{ji}, i = 1, \ldots, n)$, $j = 1, 2$.

The structure of the response model for the difference of absolute values of two deviates can be written as:

$$[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1|] = \varphi[(\Delta_i - \tilde{\Delta}),\ \sigma_i (Z_{2i} - \tilde{Z}_2) - (Z_{1i} - \tilde{Z}_1)],$$

where the function φ, whose specific structure is difficult to define precisely, indicates that when the null hypothesis $H_{0C} : (\sigma_i = 1,\ i = 1, \ldots, n)$ is true, such a function depends on the difference of two pure errors $(Z_{2i} - \tilde{Z}_2) - (Z_{1i} - \tilde{Z}_1)$, on the quantities $(\Delta_i - \tilde{\Delta})$, and on the possible interactions of all such components. However, under the null hypothesis such differences are stationary even with non-constant dispersion. Under the alternative, i.e. in case of convergence, function φ

also depends on coefficients σ_i so it will be suitable to put into evidence the possible convergence as better as the σ_i are far from unity. It is worth noting, however, that the quantities $(\Delta_i - \tilde{\Delta})$ and the interactions may depend on which hypothesis between H_{0C} and H_{1C}, is true. From the one hand, this shows that two test statistics T_U^* and T_C^* are dependent in a way that is too difficult to study and so their joint analysis require the nonparametric combination of dependent permutation tests. From the other hand, it shows that the test T_C^* will be not exact but with a rate of approximation converging to zero as sample size n diverges. We also shown that the dependence of tests T_C^* and T_U^* is asymptotically irrelevant. Simulation trials with sample sizes around $n = 30$, with asymmetric variables while using the same sets of random signs S_i^* for both tests, have shown that their correlation coefficient is practically zero. It is also worth noting that, the joint analysis of two aspects U and C, two test are to be computed on the same random permutations of signs S_i^*.

Similarly to test T_U, when for test T_C it is of interest to also detect the direction of deviates, as with $H_{1C}^{<} : (\exists \sigma_i < 1)$ and $H_{1C}^{>} : (\exists \sigma_i > 1)$, the procedure is the same with obvious substitution of symbols.

The same simulation study has shown that the test T_C^*, being essentially approximate, is somewhat *liberal*, as its rejection probability under H_{0C}, instead of $\alpha = 0.10$ it was of $\alpha = 0.145$, since it suffer from the presence of $(\Delta_i - \tilde{\Delta})$. This requires to empirically adjust its p-value distribution as $[\hat{\lambda}_C]^\gamma$, in place of $\hat{\lambda}_C$ one would have if its null distribution under H_{0C} were exactly uniform. With the same conditions of the real problem under examination, the value of such coefficient is $\gamma \approx 1.2$.

The second model for response variable Y regards the case where it, for subject i, is observed at three time occasions: $t = 1,2,3$. In such a context, responses are assumed to behave according to: $Y_{ti} = \mu + \eta_i + \Delta_{ti} + \sigma_{ti} Z_{ti}$, $i = 1,\ldots,n$, $t = 1,2,3$, where, in particular, $\Delta_{1i} = 0$ and $\sigma_{1i} = 1$, $\forall i$, and where the various coefficients, with obvious modifications of symbols, have the same meaning of the former case.

In the context of observations repeated three times is of particular interest to test for the hypothesis of monotonic convergence if any, since is properly this aspect that plays the fundamental role of Delphi method. That is, with obvious meaning of the symbols, testing for $H_0 : |Y_1 - \tilde{Y}_1| \stackrel{d}{=} |Y_2 - \tilde{Y}_2| \stackrel{d}{=} |Y_3 - \tilde{Y}_3|$, against

$$H_1 : [|Y_1 - \tilde{Y}_1| \stackrel{d}{\leq} |Y_2 - \tilde{Y}_2| \stackrel{d}{\leq} |Y_3 - \tilde{Y}_3|] \bigcup [|Y_1 - \tilde{Y}_1| \stackrel{d}{\geq} |Y_2 - \tilde{Y}_2| \stackrel{d}{\geq} |Y_3 - \tilde{Y}_3|], (2)$$

with at least one strict inequality in either branches and where $\tilde{Y}_t = Med(Y_{ti}, i = 1,\ldots,n)$, $t = 1,2,3$.

Such a kind of testing requires a sort of "*multi-aspect*" method while considering all partial tests for paired observations: $T_{C,12}^*$, $T_{C,13}^*$, and $T_{C,23}^*$ where:

$$T_{C,hj}^* = \sum_i (|Y_i(t_j^*) - \tilde{Y}(t_j^*)| - |Y_i(t_h^*) - \tilde{Y}(t_h^*)|),\ 1 \leq h < j \leq 3.$$

Since all such partial tests are homogeneous (the same sample sizes, the same null distribution, and all significant for large values), for the global inference we use the so-called nonparametric "direct combination": $T_C^* = T_{C,12}^* + T_{C,13}^* + T_{C,23}^*$.

In analogy with the case of two occasions, it is worth observing that such a combination gives a solution to the so-called "directional analysis of variance" problem [5], that is it permits to decide between two components of H_1:

$$H_1^{<} : [|Y_1 - \tilde{Y}_1| \overset{d}{\leq} |Y_2 - \tilde{Y}_2| \overset{d}{\leq} |Y_3 - \tilde{Y}_3|] \text{ and } H_1^{>} : [|Y_1 - \tilde{Y}_1| \overset{d}{\geq} |Y_2 - \tilde{Y}_2| \overset{d}{\geq} |Y_3 - \tilde{Y}_3|],$$

related to the monotonically increasing and, respectively, decreasing stochastic ordering of concentration. In particular, the combined test T_C^* becomes:

$$T_C^* = 2\sum_i [|Y_i(t_1^*) - \tilde{Y}(t_1^*)|] - 2\sum_i [|Y_i(t_3^*) - \tilde{Y}(t_3^*)|],$$

where is apparently only involved the data at times $t = 1$ and $t = 3$. It is, however, to underline that: 1) random permutations $\mathbf{t}^* = (t_1^*, t_2^*, t_3^*)$ of $\mathbf{t} = (1,2,3)$, to preserve the underlying within subjects dependence on observations, are common to three partial tests; 2) test T_C^* would be exact if in place of median estimates $\tilde{Y}_t$, $t = 1,2,3$, true medians Me_1, Me_2 and Me_3 were known; 3) the approximation rate, evaluated by a specific simulation study, is practically negligible as its convergence to zero is fast.

The nonparametric combination of $K \geq 2$ dependent tests is a useful method to make inference when a set of observed variables, for explanatory or interpretative reasons, can form a so-called section of information (for example, it is of interest to jointly see all the variables concerning the section regarding the family, and so forth). The use of such a method is unavoidable when the numbers of V variables and/or of P parameters in the response model are larger than sample size n. We invite readers to see Chapter IV of the book by Pesarin and Salmaso (2010) where the theory and related methodology is wholly discussed.

In this regards, let us suppose that the K partial tests are $(T_{C1}^*, \ldots, T_{CK}^*)$, the p-value of which are $(\lambda_{C1}, \ldots, \lambda_{CK})$. Their nonparametric combination can be done, for instance, by Fisher's combination as:

$$T_F^* = -2\sum_{k=1}^{K} \log(\lambda_{Ck}^*),$$

where $\log(\cdot)$ are natural logarithms, to obtain the p-value of which it is required that the K test statistics are jointly calculated at each data permutation and common to all of them (for instance, in terms of the Delphi data, with $T_{Ck}^*[\mathbf{X}(\mathbf{t}^*)]$, $k = 1, \ldots, K$, where $\mathbf{t}^* = (t_1^*, t_2^*, t_3^*)$ are permutations of $\mathbf{t} = (1,2,3)$, for data observed at three times, and so forth).

3 Analysis of the results

The so-called Delphi effect is configured as the interaction of two distinct but not exclusive contributions: that of the median convergence of expert evaluations, here verified through the non-parametric equality test (U), and that in the distribution of the same assessments verified through the convergence test (C) of the pairwise comparisons of the surveys. The hypothesis H_0 of equality verifies if the distributions

related to the surveys agree in median between them - two by two and amongst all - that is, the tendential idea (expressed by the median of the first survey) is also confirmed in the subsequent interviews. The inevitable and reasonable adjustments of the distribution median, typical of the Delphi method, are not sufficiently large to be significantly relevant, and the interviewees, although from different directions and positions, offer indications during the three rounds that are recognised in the same orientation and opinions expressed by the median value of the distribution. The hypothesis H_0 of convergence states that the dispersion around the median does not decrease significantly, so its rejection refusal is to be read as a confirmation that in the course of the rounds, there is sensitive and gradual convergence and consolidation towards the central value of the distribution. In the case of statistical significance, the experts involved during the surveys move away gradually but consistently from their initial positions and approach the final median. On the basis of model (2), the two contributions on which the hypothesis test is started are not directly separable or evaluable in a strictly separate manner (each one, in fact, conditions the other) and to an increasingly lesser extent as the number of observations increases. This last consideration clashes with a qualifying feature of the Delphi approach in which it is conducted through a very limited number of experts and surveys. Therefore, because of the impossibility of distinguishing the two components of the Delphi effect, the combined Fisher test is applied to measure the joint effect of the two contributions described above. The result of the significance of the latter supports and integrates the summary information offered by the reading of the performance indicators. In Table 1 reports the description by summarising the measures of the performance parameters and the level of significance of the tests.

Table 1 tab:1 Performance parameters stability, convergence speed and consensus for each items and the level of significance of the test on equality (U), convergence (C) and combined with the Fisher (F) test of the pairwise comparisons of the rounds, as well as the multi-aspect one per item

Area 1. Parents (six items)	S(*)	VC	C	1-2 U	1-2 C	**1-2 F**	2-3 U	2-3 C	**2-3 F**	1-3 U	1-3 C	**1-3 F**
1. Parents (father and mother) will devote themselves to training their children.	−	++	++	.519	.068	**.384**	.051	.006	**.004**	.194	.002	**.006**
2. The father will be present in the training and leisure activities of the children (school, sports, associations, etc.).	+	++	++	.287	.103	**.333**	.103	.014	**.027**	.070	.000	**.000**
3. The mother will be able to organise work and family life to be more present in the children's activities of education and free time.	+	−	+	.320	.040	**.166**	.315	.056	**.219**	.217	.004	**.014**
4. For the mother, the organisation of family life will be conditioned by professional rhythms and commitments.	+	+	++	.187	.032	**.090**	.064	.302	**.233**	.053	.005	**.006**
5. The father will try to organise professional commitments according to the organisation of family life.	++	+	++	.273	.033	**.125**	.174	.144	**.251**	.130	.011	**.037**
6. Parents will invest in the role as educators of their children.	++	++	++	.145	.003	**.015**	.090	.110	**.126**	.476	.000	**.002**

* S: Stability; V of C: Speed of convergence; C: Consensus.

Legend common to the three indicators $++$: excellent perfect; $+$: good partially goode; $-$: absence.
Consider that the initial domain for the response scale is 100, so a final interquartile range less than 20 is considered good.

1. Stability: $S_1 = 10$, $S_3 = -10$. There is none; the assessments converge quickly. V of C: $v_1 = 5$, $v_3 = -5$, excellent. Consensus: $IQ = 10$, excellent.
2. Stability: $S_1 = 0$, $S_3 = -10$. Is in the first quartile but not in the third one, which tends to decrease. V of C: $v_1 = 0$, $v_3 = -2.5$, excellent. Consensus: $IQ = 10$, excellent.
3. Stability: $S_1 = 5$, $S_3 = -1.3$. There is in the first quartile, but the third one seems stabler. V of C: $v_1 = -2.5$, $v_3 = -5.6$. It is not good that the coefficients are both negative. Consensus: IQ = 13.7. Good. There is a change in course starting from the second round. The certain and important element, however, is the final consensus that being less than 15 is good.
4. Stability: $S_1 = 10$, $S_3 = 0$. Not good on the first quartile but perfect on the third one.V of C: $v_1 = 5$, $v_3 = 1.3$. The first quartile quickly converges towards the median, whereas the third one stabilises. Consensus: $IQ = 10$, excellent.
5. Stability: $S_1 = 0$, $S_3 = 0$. Perfect. V of C: $v_1 = 0$, $v_3 = -2.5$. Good; the third quartile decreases rapidly. Consensus: $IQ = 10$, excellent.
6. Stability: $S_1 = 0$, $S_3 = -1.3$. Good on the third quartile and perfect on the first one. V of C: $v_1 = 2.5$, $v_3 = -5$ Excellent. Consensus: $IQ = 10$, excellent.

The item, the three parameters provide mostly satisfactory measurements; no item was found to achieve negative performance, except for 3. The consensus is at least good for all items. In the first summary, it is observed that the performance indicators on the evolution of the items offered a generally satisfactory picture, even if apparently contradictory situations were recorded on individual items, as in items 1 and 3, where either stability or the V of C takes on unsatisfactory dimensions or, in any case, is in line with the other parameters.
The analysis of the results of the test application on significance shows that the hypothesis H_0 of equality is almost always accepted even with high levels for all three possible comparisons of the three measurements 1-2, 2-3 and 1-3. This information allows us to confirm that in the median, the three distributions are statistically equivalent, indicating that although the procedure registered different levels of stability amongst the six items - from absent in item 1 to very good in items 5 and 6 - this did not significantly alter the central effect expressed by the medians of the three rounds; this proves that the basic opinion of the group of experts tended to remain the same, albeit with variations. The first contribution of the Delphi effect can be said to be confirmed for these six items. It should be noted, however, that the descriptive analysis of stability is only based on the results of the second and third rounds, whereas with the application of the tests, all three distributions are compared with information that is therefore more exhaustive and differentiated. The analysis of the significance levels of the Ho of the convergence hypothesis (indicated in the columns with C) records less-systematic trends than the previous one. If we want to summarise and conclude, only in comparison 1-3 is the hypothesis Ho is systematically rejected for each of the six items, whereas in the previous comparisons, alternating positions are recorded.
It should be noted that the consensus expresses a measure of the reduction in the range of variation (through the interquartile range) only of the last round, whereas in the inferential analysis, three are jointly examined and compared. From a targeted

reading of the results of significance, it is therefore clear that the three iterations were necessary and also sufficient to detect the monotonic convergence as expected -and desired- precisely by the procedure of successive interviews envisaged in Delphi, otherwise indicated just as Delphi effect. This convergence would not have been achieved, or at least partially only, had we stopped at the first two surveys. This information and conclusion would not have been achieved on the basis of only reading the consensus results. A more careful analysis, however, leads to the consideration of how the latter indicator (third column of Table 1) is always satisfactory in partially confirming the presence of reciprocal and partial integration of the two pieces of information. The significance levels of the combination Ho hypothesis (in Table 1 with F) of the two previous components are rejected and always have a high level of significance only in the comparison between the first and third rounds; in the two other comparisons, the situation is variegated, as the hypothesis Ho is mostly accepted. The high levels of significance of F make it clear that the joint effects of the two distinct contributions of the Delphi effect are consistent and translate to the effective confirmation of information coming from the Delphi procedure applied and described here. Joint effects are still not evident from the first two rounds (results everywhere are predominantly not significant with respect to comparisons between 1-2 and 2-3) to further confirm that we can reiterate the need for the three interviews to converge to expendable results. Finally, the information produced by the inferential analysis offers reflections and indications that the analysis of the indicators may, to some extent, be direct but not defined.

References

1. Bolzan M.(2018), *Domani in Famiglia*. Franco Angeli, Collana Strutture e Culture Sociali, Milano: pp. 1-226. (2017)
2. Dajani J.S., Sincoff M.Z., Talley W.K. (1979), *Stability and agreement criteria for the termination of Delphi studies*, Technol. Forecast. Soc. Chang. 13 pp. 83–90.
3. Glenn J.C. (1972), *Futurizing Teaching vs Futures Course, Social Science Record.* Syracuse university, Vol IX, n.3, Spring.
4. Pacinelli A. (2007), *Metodi per la ricerca sociale finalizzata.* Franco Angeli, Milano.
5. Pesarin F. Salmaso L. (2010),*Permutation test for complex data: Theory, Application and Software*. Wiley, Chichester, UK.

Biplot and unfolding models for evaluation of teacher behaviour in the classroom

Utilizzo dei modelli biplot e unfolding per la valutazione dell'atteggiamento dell'insegnante in classe

Giuseppe Bove, Maria Gaetana Catalano, Paola Perucchini, Alessio Serafini, Giovanni Maria Vecchio

Abstract Biplot and multidimensional unfolding are proposed to make easier interpretation of data obtained by administering questionnaires to evaluate teachers behaviour in classroom. Data visualization in diagrams obtained by biplot allows to detect relationships between teachers and items. Besides, dispersion of ratings inside classrooms are detected in a diagram obtained by the unfolding model.

Abstract *Biplot e unfolding multidimensionale sono proposti per rendere più semplice l'interpretazione di dati desunti dalla somministrazione di questionari di valutazione del comportamento dell'insegnante in classe. La rappresentazione grafica ottenuta con il biplot consente l'analisi della relazione tra insegnati ed item del questionario. Inoltre, la variabilità dei punteggi forniti dagli alunni nelle classi è analizzata nel grafico ottenuto dall'applicazione del modello unfolding.*

Key words: Biplot, Multidimensional unfolding, Data visualization

1 Introduction

Evaluation of teacher behaviour and classroom management can give an important contribution to the creation of a good climate and to the achievement of effective learning.

[1] Giuseppe Bove, Maria Gaetana Catalano, Paola Perucchini, Giovanni Maria Vecchio, Dipartimento di Scienze della Formazione, Università Roma Tre; giuseppe.bove@uniroma3.it, maria.catalano@uniroma3.it; paola.perucchini@uniroma3.it; giovannimaria.vecchio@uniroma3.it.

Alessio Serafini, Università di Perugia; alessio.serafini@uniroma1.it.

Research provided different instruments and methodologies to analyse teachers behaviour, frequently they consist of questionnaires composed of subsets of items, where each subset identify a particular subscale (e.g., Catalano, Perucchini, Vecchio, 2014). When teachers in a school are evaluated the questionnaire is administered to all the pupils (students) in the classroom, so each teacher is evaluated by a different group of pupils (students). In this situations it is interesting to analyse relationships between items (or subscales) and teachers, for instance to detect different rank order of the teachers along an item (or subscale). Besides, it is also important to analyse the level of dispersion of pupils ratings in the classrooms to investigate aspects of rating's reliability. When teachers are also evaluated by other people (e.g., peers, tutors, principals), it is worth to compare the different types of ratings (pupils, peers, tutors, etc.) to detect differences in perception of teachers behaviour.

In the following a proposal to make easier the analysis of the previous aspects is described, capitalizing on diagrams obtained by biplot and unfolding models (e.g., Borg and Groenen 2005, Greenacre 2010, Gower, Gardner and Lubbe 2011). Relationships between teachers and items can be detected in diagrams obtained by biplot models, and dispersion of ratings inside classrooms are represented in a diagram obtained by the unfolding model.

Finally, the main features of the proposal will be illustrated in a study for the assessment of learning teacher behavior in classroom. Data were collected in a research conducted in 2018 at Roma Tre University with students of the degree course in Formazione Primaria during their experience of internship ("tirocinio") at school.

2 Data matrices description and purposes of the analysis

The first type of matrix to analyse is a multivariate data matrix $\mathrm{X} = (x_{ij})$ with each row corresponding to a teacher and each column to an item of the questionnaire. According to the different situations, the entry x_{ij} can be the mean rating obtained by teacher *i* for item *j* in the classroom (the average of pupils ratings), or the rating obtained by teacher *i* for item *j* by a peer or by a tutor (in particular in the case of student teacher). This type of matrix contains the information for comparing teachers and for analysing relationships between teacher and items (or subscales). Besides, as we could observe two or more matrices (pupils, peers, tutors, etc.), an important point will be to compare several evaluations.

The second matrix $\Delta = (\delta_{ij})$ also has each row corresponding to a teacher and each column to an item of the questionnaire, but the entry δ_{ij} is the variation coefficient of the ratings obtained for item *j* in classroom of teacher *i*. Entries of matrix Δ, considered as dissimilarities between teachers and items (this is allowed because ratings are usually positive variables), contain the information for analysing the dispersion of ratings in the classrooms. In the final application, this dispersion aspect will be considered only for ratings obtained by pupils, because each teacher will be evaluated by only one peer, and so for each item only one rating is available.

3 Biplot and Multidimensional Unfolding

Any rectangular matrix may be represented choosing a vector for each row and a vector for each column in such a way that the elements of the matrix are the inner products of the vectors representing the corresponding rows and columns. This representation is based on the *biplot* model (Gabriel, 1971), that in scalar notation is:

$$x_{ij}^{t} = \sum_{s=1}^{r} a_{is}\, b_{js} + \varepsilon_{ij} \qquad (1)$$

where x_{ij}^{t} is the value x_{ij} of a numeric variable opportunely transformed (e.g., mean centred or standardized), a_{is} (component score) and b_{js} (component loadings) are the coordinates respectively of row (teacher) *i* and column (item) *j* on dimension *s* in an *r-dimensional* space and ε_{ij} is a residual term. We can always obtain a perfect representation ($\varepsilon_{ij} = 0$) of the values x_{ij}^{t} by choosing *r*=*rank* (X), but we do not have a graphical representation because *rank* (X)>3 in most applications. So, to obtain and make easier graphical interpretation, rows and columns are usually represented in two dimensions (*r*=2). In order to render unique the bilinear decomposition of the biplot model, coordinates a_{is} and b_{js} can be normalized in different ways. One of the most frequent normalization (principal coordinate) allows to represent each row (teacher) by a point in such a way that differences between the corresponding row vectors of the data matrix (e.g., differences between the rows of obtained mean ratings on the items) are approximated by inter-point distances in the diagram. Each column (item) is represented by a vector (arrow) that identifies a direction (or axis) in the plane. Points representing rows (teachers) can be projected along that direction, and the product of the length of the column vector and the length of the obtained projection approximate the values x_{ij}^{t} (e.g., mean ratings) in that column (item).

In order to simplify interpretation of diagrams, we can limit ourselves to compare point projections along column vector directions, because these projections are proportional (according to the vector length) to the corresponding values x_{ij}^{t} in the columns. So, to summarize, the comparison of row points (teachers) projections along the direction identified by column vector (item) *j* provides an approximation of the rank order of values x_{ij}^{t} in column *j* (that coincides with rank order of values x_{ij} in column *j* of the original data matrix X, because mean centering or standardization does not change rank order).

According to a different choice of the normalization of the coordinates, correlations between items can also be represented.

Model (1) is based on the assumption that entries x_{ij} are values of numeric (interval or ratio) variables and that the relationships between variables are linear. Frequently in the social sciences variables are categorical (nominal or ordinal) and relationships between variables are nonlinear, so, model (1) is not appropriate. To overcome the problem, a nonlinear approach (named NLPCA - NonLinear Principal Component Analysis) developed in the Eighties, that does not make assumptions concerning the measurement level of the variables and the nature (shape) of their

relationships, analyzes the data at a level specified by the researcher (interval, ordinal, or nominal), providing quantifications of the categorical variables accordingly (for a review see Meulman *et al.* 2004, Linting *et al.* 2007, Linting & van der Kooij 2012). In this case the model is reformulated as

$$q_{ij} = \sum_{s=1}^{r} a_{is}\, b_{js} + \varepsilon_{ij} \qquad (2)$$

where q_{ij} is the value of the transformed variable of column (e.g., item) j in row (e.g., teacher) i, a_{is}, b_{js} and ε_{ij} are defined as in model (1). For standard biplot of model (1), component scores a_{is} and component loadings b_{js} in CATPCA are obtained by singular value decomposition (SVD) of the standardized data matrix. Parameters in model (2) are estimated by an iterative process in which a least squares loss function is minimized by cyclically updating one of the three sets of parameters (a_{is}, b_{js}, q_{ij}). For details on the computational algorithm see Linting *et al.* (2007). Since transformations q_{ij} preserve order relationships for ordinal categorical variables, guidelines already provided for interpretation of diagrams obtained by standard biplot hold.

The *unfolding* model, originally proposed by Coombs (1964) for rectangular matrices of preference scores, is given in the following:

$$\delta_{ij} = \sqrt{\sum_{s=1}^{r} \left(a_{is} - b_{js}\right)^2} + \varepsilon_{ij} \qquad (3)$$

where δ_{ij}, a_{is}, b_{js} and ε_{ij} were defined in this and the previous section. We can consider the previous model as the distance version of the biplot model (1), where distances replace inner products. It is worth to notice that the Euclidean distance model usually used in multidimensional scaling (MDS) for square dissimilarity matrices (e.g., Borg & Groenen 2005) is a constrained version of model (3), because for each j it is required $b_{js} = a_{js}$. A diagram for the pattern of relationships is obtained where each row (teacher) is represented as a point with coordinates a_{is} and each column (item) as point with coordinates b_{js}. In the planar representation (r=2), the distance between row (teacher) i and column (item) j approximates the corresponding dissimilarity δ_{ij} (variation coefficient, so, for instance, we can detect in the diagram both the teachers and the items with low (high) homogeneity of ratings in the classrooms). Distances within each of the two sets of the row-points and the column-points are only implicitly defined and do not have corresponding observed entries in the data matrix. Parameters in model (3) are estimated by iterative algorithms that, starting from initial estimates of a_{is}^0, b_{js}^0 (*initial configuration*), iteratively decreases a least squares loss function moving vectors $\boldsymbol{a}_i^0 = (a_{i1}^0, a_{i2}^0, \ldots\ldots, a_{ir}^0)$ and $\boldsymbol{b}_j^0 = (b_{j1}^0, b_{j2}^0, \ldots\ldots, b_{jr}^0)$, until convergence to a minimum. An important point is picking a good initial configuration, in order to avoid the problem of *local minima* (that means the algorithm did not find the best possible choice for coordinate matrices). Available programs for unfolding allow to start with many different configurations, so it is possible to check the stability of the

estimates obtained, repeating their computations with different starts. Stability of the estimates is an important indication for a global minimum.

4 Application

A reduced version of the Teachers' Educational Practices Questionnaire (TEP-Q, Catalano, Perucchini, Vecchio 2014) was administered to evaluate a group of 24 female student teachers of Roma Tre University, during their training (internship) in several primary schools of the Italian region Lazio in school year 2018. The questionnaire consists of 12 items regarding teachers behaviour in the classroom, examples of items are: "He/she was calm and relaxed in the classroom" (Item 1), "He/she helped us to say something better, if we hadn't been so clear" (Item 4), "He/she has made us of group works" (Item 11). Answers were provided on a 4-levels Likert scale (1=almost never, 4=almost always). Ratings were obtained from 418 pupils (204 females, 214 males), aged between 7 and 12 years. For each of the 24 student teachers, a mean rating was obtained for each item by averaging the ratings of the pupils in her classroom. Besides, also a peer provided ratings for the same 12 items by observing the student teacher behaviour during the lesson. So two (24×12) matrices of ratings are available: one for pupils and one for peers.

First, the matrix of pupils ratings was analysed by standard biplot model (1), because we assumed the mean ratings obtained by the pupils in each classroom as values of a numerical variable. This matrix contains the information for analysing relationships between student teachers and items and for comparing student teachers. The procedure named CATPCA available in IBM-SPSS program (release 25) was used to obtained component scores and component loadings. Ratings are standardized for each item in CATPCA before estimation of parameters in model (1). Figure 1 shows the diagram obtained approximating data in $r=2$ dimensions. The variance accounted for (VAF) by this solution is 50.7% of the total variance in the data matrix. As it was explained in section 3, in the diagram each student teacher is represented by a point and each item is represented by a vector. Since ratings are standardized for each item by CATPCA, coordinates of the projections of each student teacher on an item vector provide the position of the student rating respect to the mean rating (that coincide with the origin) on that item (positive direction is identified by the grey point at the end of the vector). Negative coordinates correspond to ratings less than the mean rating of the item, positive coordinates correspond to ratings greater than the mean rating.

Besides, positions of projections along the item vector provide an approximation of the rank order of the ratings obtained by student teachers on that item. The approximation will be the more reliable the higher the value of VAF. The value 50.7% obtained for the VAF of our biplot model allow us to be quite confident in the analysis of the diagram depicted in Figure 1 (that however does not exclude that some rank order is not exactly represented).

To interpret the diagram in Figure 1, we can consider, for instance, positions of student teachers 20 and 14 respect to item 1 and item 11. Projection of student

teacher 20 on the axis generated by item 1 (vector) falls far from the origin on the left side of the axis, and so it represents one of the lowest rating on item 1. Projection of student teacher 14 falls quite far on the right side of the same axis and so represent one of the highest ratings on item 1. So, student teacher 14 in her classroom is judged better than student teacher 20 in her classroom respect to item 1, that means she is judged usually calmer and more relaxed than student 20 (however we should take into account that the two classrooms are usually different). Comparison of projections of student teachers 20 and 14 on the direction generated by item 11 follows the same way and indicates that student 14 has the highest rating and student 20 has a rating just below the mean, that means student 14 more frequently made up group works respect to student 20 (that made up group works as the average of the student teacher group).

On the whole, projection of student teacher 20 on many of the item directions has low coordinate values, that indicates low mean ratings for many items. The opposite happens for student teacher 14, with all projections having high values, that means she is judged always better than the average of the student teacher group and better than many other student teachers. Roughly, the same comment provided for student 14 (ratings higher than the mean rating of the student group) applies to all student teacher points positioned in the portion of the plane bounded by item 11 and item 10 in Figure 1 (e.g., student 10, student 21, etc., in total seven students). Differently, students points positioned outside the portion of the plane bounded by item 11 and item 10 (e.g., student 3, student 4, student 23, etc., in total seventeen students) are commented in a way more similar to teacher student 20, with ratings lower than the mean rating of the student group in at least some of the items.

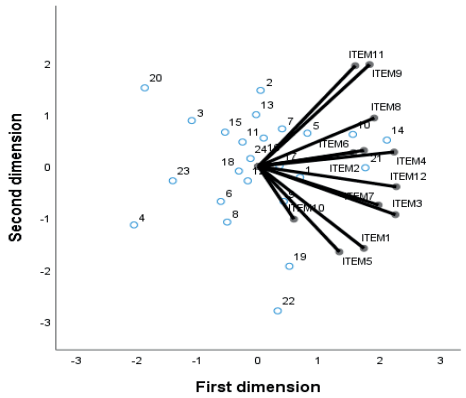


Figure 1: Biplot of teachers and item for pupils ratings

The previous comments concerning the positions of the whole group of teacher students in the plane are useful to compare ratings obtained by pupils in the classrooms with ratings obtained by peers. The diagram obtained by applying model (2) to the matrix of peers ratings is provided in Figure 2. Procedure CATPCA was

used also in this application, with ratings considered defined at an ordinal scale level. VAF was 56.3%.

In summary, the comparison of diagrams in Figures 1 and 2 indicates that the two evaluations (pupils and peers) seem quite different for several student teachers. For instance, teacher student 14 has a position close to the origin in Figure 2, on the border of the portion of the plane delimited by item vector 10 and item vector 11, and as a consequence she has negative values (lesser than the mean ratings) on item 1, item 7 and item 10. So, the peer judged teacher student 14 less well than pupils in her classroom. On the contrary, teacher student 20 received positive ratings on several items by her peer, being judged better than by pupils in the classroom. On the whole, pupils evaluations seem less positive than peer evaluations, because the number of student points positioned outside the portion of the plane delimited by item vector 10 and item vector 11 (eleven student points) in Figure 2 results less than the corresponding number of student points (seventeen students) positioned in the same portion of the plane in Figure 1.

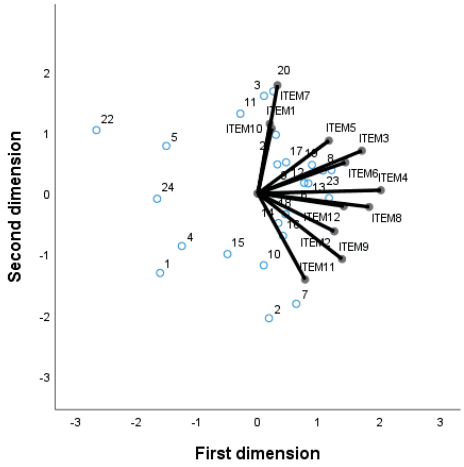


Figure 2: Biplot of teachers and item for peer ratings

Finally, the analysis of dissimilarities δ_{ij} (variation coefficients) was conducted by the unfolding model (3). In Figure 3, the solution for r=2 dimensions is provided. Distances between student teachers and items represent the level of the corresponding variation coefficients for the items in the classroom of the student teacher (the greater the distance the higher the dispersion). Considering again the same two student teachers 20 and 14, we observe that student teacher 20 is far from most items because she has usually high variation coefficients for the ratings obtained in her classroom. On the contrary, student teacher 14 is near the centre of the diagram and close to many items, as a consequence of the homogeneity of

ratings obtained on many items. Items 5, 7 and 10 have high heterogeneity in many cases, so they are positioned far apart from many student teachers.

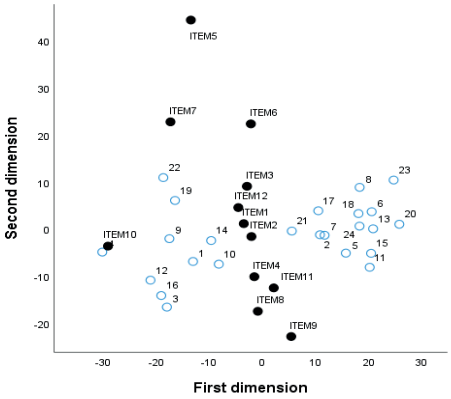


Figure 3: Unfolding of variation coefficients for teachers and items

References

1. Borg I, Groenen, PJF. Modern Multidimensional Scaling. Theory and Applications. (Second Edition), Springer, New York (2005).
2. Catalano, M. G. Perucchini, P., Vecchio ,G. M.: The Quality of Teachers' Educational Practices: internal Validity and Applications of a New Self-evalutation Questionnaire. Procedia-Social and Behavioral Sciences, 141, 459-464 (2014).
3. Coombs C.H. A Theory of Data, Wiley, New York (1964).
4. Gabriel K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58, 3, 453-467 (1971).
5. Gower, J.C., Lubbe, S., Le Roux, N. Understanding Biplots, Chichester, UK: Wiley (2011).
6. Greenacre, M.J. Biplots in Practice, Madrid: BBVA Foundation (2010).
7. IBM Corp. Released. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp. (2017)
8. Linting, M., Meulman, J.J., Groenen, P.J.F., van der Kooij, A.J.: Nonlinear principal components analysis: Introduction and application. Psychological Methods, 12, 3, 336-358 (2007).
9. Linting, M., van der Kooij, A.J.: Nonlinear principal components analysis with CATPCA: a tutorial. Journal of Personality Assessment. 94, 1, 12-25 (2012)
10. Meulman, J.J., van der Kooij, A.J., Heiser, W.J.: Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: Kaplan D. (ed.) The sage handbook of quantitative methodology for the social sciences, pp. 49-70. Sage Publications, London (2004)

Student satisfaction for Teaching and its impact on drop-out rate

La soddisfazione degli studenti per la didattica e il suo legame con il tasso di abbandono

Barbara Cafarelli, Angela Maria D'Uggento and Alessandra Petrucci

Abstract Since 2009 the Department of Economics of Foggia has started the project "Analysis of student satisfaction" with the aim to integrate the internal quality control system also with its students' feedback. In this context, the evaluation of the students' satisfaction about Teaching and its relationship with students' retention are crucial aspects. In order to understand the satisfaction/feeling level of Teaching and the psychological mechanism behind the student evaluation process, MUB models have been adopted. Then, the feelings towards Teaching estimated by MUB models and the relation with the drop-out rate are analyzed by linear models.

Abstract *A partire dal 2009, il Dipartimento di Economia dell'Università di Foggia conduce il progetto "Analisi della soddisfazione degli studenti", con l'obiettivo di integrare il sistema di assicurazione interno della qualità anche con i feedback degli studenti. In questo ambito, assumono grande rilevanza la valutazione della soddisfazione degli studenti per la didattica e la sua influenza sulla decisione di proseguire gli studi nello stesso ateneo. La soddisfazione/feeling degli studenti per la didattica e i fattori che maggiormente lo influenzano sono stimati con i modelli MUB. Successivamente, la relazione tra il tasso di abbandono e la soddisfazione/feeling per la didattica viene analizzata con i modelli lineari.*

Key words: students' satisfaction, drop-out rate, MUB models.

1 Introduction

In the last decade, the Italian university system has undergone a profound reform process that has affected both the quality evaluation and the financing model. An

[1] Barbara Cafarelli, University of Foggia; email: barbara.cafarelli@unifg.it
Angela Maria D'Uggento, University of Bari; email: angelamaria.duggento@uniba.it
Alessandra Petrucci, University of Florence; email: alessandra.petrucci@unifi.it

articulated quality evaluation system concerning administrative, educational and research activities, which is aimed at verifying the correct use of public resources and the performance of universities, has been introduced. The National Agency of Evaluation of the Universities and the Research Institutions (ANVUR) is one of the main protagonists of this process. Strategic planning of Italian universities cannot leave aside the ANVUR criteria in the fields of self-assessment, evaluation, initial and periodic accreditation of universities and degree courses (AVA), evaluation of research quality (VQR) and the third mission, along with the criteria for the PRO3 and FFO allocation, which are the main financial funds provided by the Ministry of Education, University and Research (MIUR). In this context, the need and availability of information to support policy makers and stakeholders' decisions has increased considerably. In particular, data on student careers and quality of services offered by universities have become a very important issue. Consequently, students' perceptions about Teaching and services offered by universities become a useful tool for assessing the efficiency of university policies. For this reason, the Department of Economics at the University of Foggia carries out the "Analysis of student satisfaction" project with the aim of improving the internal quality assurance system. The aim is achieved by monitoring the students' feedback about Teaching and Support services (Registrar's offices, Laboratories, Libraries, Website, etc) [1]. Through the feedbacks provided, it is possible to identify which aspects most affect the students' satisfaction. Moreover, students can choose the university and, then, decide to complete their academic careers or dropping-out if they are not satisfied. Consequently, they can be considered as "customers" [6]. Indeed, there are several factors influencing the choice of university which deals with a complex process involving families and students, their expectations, placement opportunities, financial availability, travelling time and the quality of services offered by the university [4]. In this competitive scenario, universities may be interested in investigating some factors: What factors influence the choice of university? What determines student satisfaction? What is considered to be attractive for a student? How does the quality of Teaching influence the drop-out rate? In this paper the attention is focused on students' satisfaction for Teaching and its possible relation with the drop-out rate by means of the survey carried out by the Department of Economics of Foggia in the last ten academic years. In particular, the MUB models are used to assess how the judgments of the students are influenced by their personal feeling (satisfaction) towards the items under investigation and by the inherent uncertainty associated with the choice of the ordinal values gathered in the questionnaire responses. Then, the possible influence of the students' satisfaction (feeling) for Teaching on the drop-out rate is investigated by using linear models. The drop-out represents an inefficiency for the Italian university system that has long tried to stem it without success. Eurostat data puts Italy in second place, after France, for the number of students who did not complete their academic careers in 2016 [3]. Understanding the drop-out motivations and identifying at-risk students in advance implies the possibility of reducing its impact.

In this paper we only report the results of the satisfaction survey referring to Teaching.

The paper is organized as follows: sections 2 and 3 deal with a description of the survey's characteristics and a presentation of the statistical procedures, the main results are discussed in section 4. Section 5 contains some final remarks.

2 The student survey and the drop out-data

The evaluation of the student satisfaction about Teaching and its support services is performed by means of a questionnaire proposed to the students by the Department of Economics at University in Foggia. In particular, the survey is addressed to the students attending lectures of the two Bachelor's degrees (Economics and Business Administration) and the two Master's degrees (Business Administration and Marketing Management) during the academic years from 2009-10 to 2017-18 for a total of 7,257 interviews. It was decided to address the survey only to attending students as they are those who directly experience the Department services. All the courses are offered in the face-to-face format; students' participation was voluntary, anonymous and the questionnaires always completed.
The services under evaluation are Teaching, Logistics, Registrar's Office and Website along with Overall satisfaction. The judgements are expressed in a Likert scale from 1 (extremely unsatisfied) to 7 (extremely satisfied). In this paper, the main results about Teaching are reported as it has emerged to be one of the most important factors affecting student satisfaction. In particular, ten items of Teaching are evaluated with respect to reliability, responsiveness, empathy, tangibles and assurance dimensions: *Usefulness of teaching materials (MAT); Subjects consistency with course goals (SBJCONS); Teachers' availability (AVAIL); Lessons' clearness (CLRNS); Teachers' professionalism (PROF); Teachers' ability to stimulate students' interest (INTLS); Teachers' celerity to replay to student emails (CELRP); Office hours observed (OFFHOBS); Class schedule observed (CLSCHOBS); Overall satisfaction for Teaching (OVSAT).*
Then, the students' retention is measured by means of its complementary variable, the drop-out rate. Data on the drop-out rate come from *Anagrafe Nazionale degli Studenti,* a huge and complete database containing information about all Italian students' academic careers overseen by MIUR. The drop-out rate is calculated as a ratio between the number of students enrolled in the second-year course and those enrolled in the previous year, following the cohort.

3 The statistical analysis based on MUB models

The students' satisfaction assessment is made by using MUB models as a means to understand how customer preferences are influenced by a subjective personal feeling towards the item under investigation and by the inherent uncertainty associated with the choice of the ordinal values featuring on the questionnaire responses. A MUB

model [2, 7] is a mixture model combining a shifted binomial random variable, used to evaluate the feeling and a discrete uniform random variable, used to express the uncertainty, as follows:

$$P(R=r)=\pi\binom{\mathbf{m-1}}{\mathbf{r-1}}(1-\xi)^{r-1}\xi^{m-r}+(1-\pi)\frac{1}{m}\quad r=1,2,\ldots,m \qquad (1)$$

where $\xi\in[0,1], \pi\in\,]0,1]$, r is the rating and m>3. The ξ and π parameters are related to the latent components of feeling and uncertainty, respectively, and differently characterize the probability distribution of R. The interpretation of the two parameters depends on the scale of judgements. In this study, as m=7 corresponds to the best positive judgment, the quantity (1- $\hat{\xi}$) increases with the respondents' satisfaction toward the item under judgment and can be interpreted as a measure of feeling/liking [8]. The π parameter represents the contribution of the Uniform distribution to the R random variable and it is inversely related to the weight of the uncertainty component (1- $\hat{\pi}$). In particular, the random variable, R, tends to assume the shape of a Uniform distribution, thus suggesting a total uncertainty choice when $\hat{\pi}$ is close to 0; when $\hat{\pi}$ is close to 1, R tends to behave as a shifted Binomial distribution, suggesting a completely thoughtful choice. Inferential issues are fully specified in [7]. Under these premises, in order to make the results interpretation clearer, the estimated feeling and uncertainty are represented in the parameter space and the coordinates are directly related to the latent component of feeling (1- $\hat{\xi}$), in vertical axis, and of uncertainty (1- $\hat{\pi}$), in the horizontal one. The models (1) are estimated using package CUB 3.0, available in R environment [5].

The statistical analysis involves the following steps. A model (1) is estimated for each of the four Bachelor's and Master's degrees, for each of the nine years and for each of the ten aspects of Teaching under evaluation (more specifically, 360 MUB models are estimated). For each estimated model, the goodness of fit is measured by the dissimilarity index [2]. In order to investigate the second research hypothesis, a linear model is estimated to analyse the influence of the feeling for the overall satisfaction for Teaching on the drop-out rate. Following, a stepwise regression with backward elimination according to AIC criterion has been carried out to detect which aspects of Teaching influence the overall students' satisfaction the most. Finally, the possible influence among the drop-out rate and the feelings of the most influencing items with respect to the overall satisfaction for Teaching is investigated by a stepwise regression with backward elimination according to AIC criterion.

4 Main results

The values of the estimated MUB parameters suggest a good level of the students' satisfaction for Teaching in all the courses and in each of the nine years under evaluation: the estimated quantity (1- $\hat{\xi}$) is always higher than 0.5, suggesting a

positive feeling. The quantity $(1-\hat{\pi})$ is always less than 0.5, showing that the interviewees were quite precise in giving marks (Figure 1).

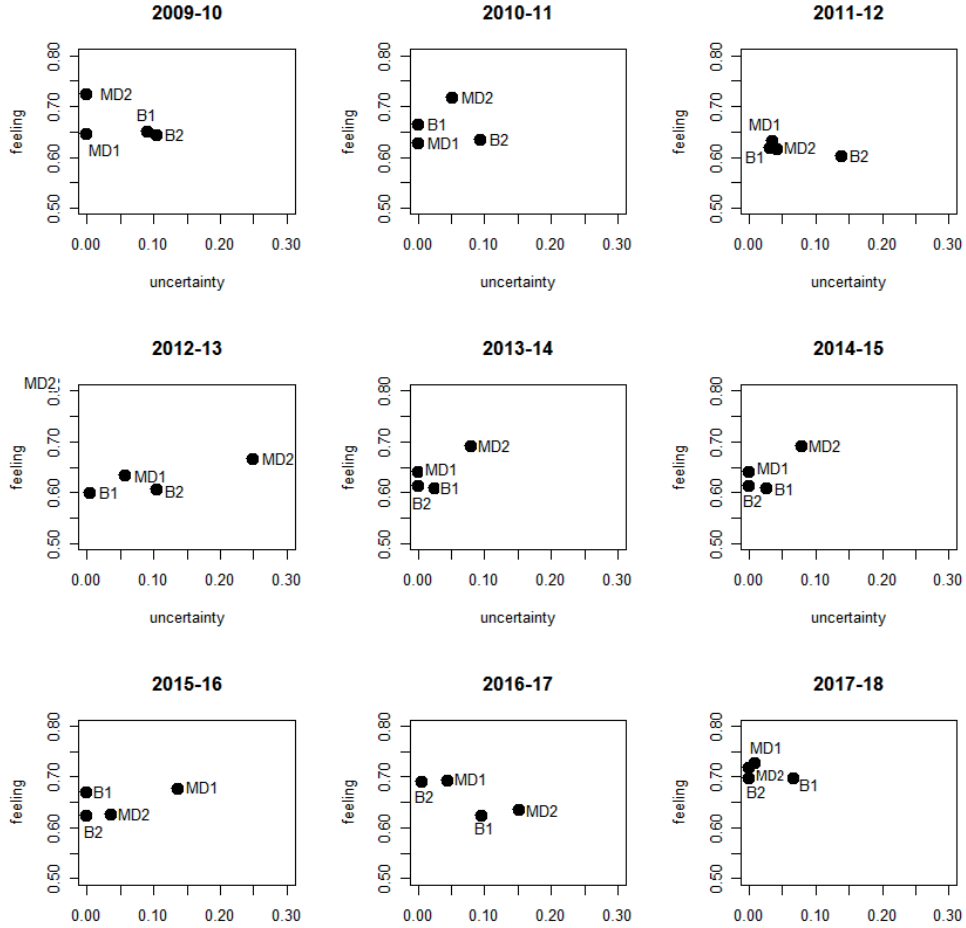


Figure 1: Feeling and uncertainty estimated by MUB models for *Overall satisfaction with Teaching* in Bachelor's degrees Economics (B1) and Business Administration (B2) and the two Master's degrees Business Administration (MD1) and Marketing Management (MD2) in each academic year.

The same results can be observed if one considers the 10 aspects of Teaching represented through their corresponding acronyms. The *Teachers' professionalism* is highly appreciated by students, as highlighted by a high level of feeling along with a low uncertainty (Figure 2).

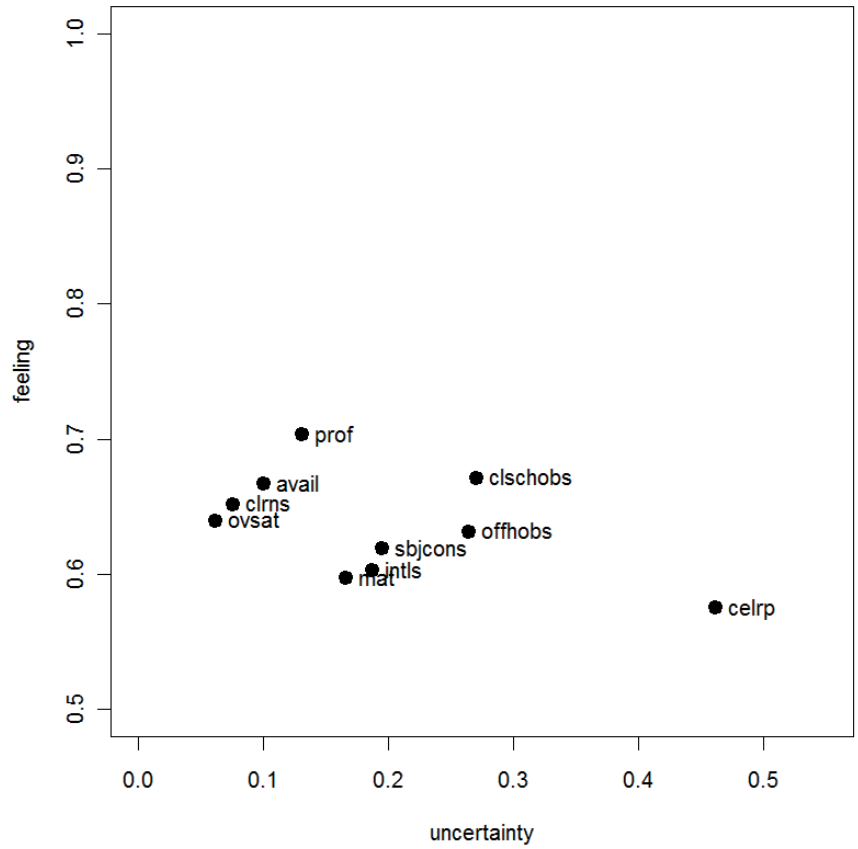


Figure 2: Feeling and uncertainty estimated by MUB models for *Usefulness of teaching materials (MAT); Subjects consistency with course goals (SBJCONS); Teachers' availability (AVAIL); Lessons' clearness (CLRNS); Teachers' professionalism (PROF); Teachers' ability to stimulate students' interest (INTLS); Teachers' celerity to replay to student emails (CELRP); Office hours observed (OFFHOBS); Class schedule observed (CLSCHOBS) and Overall satisfaction with Teaching (OVSAT)* from academic year 2009-10 to 2017-18.

By means of a deeper analysis, the values of the estimated MUB parameters show a good level of satisfaction with Teaching among the respondents except for 5 of the 360 estimated models. In particular, a slight dissatisfaction emerges about the *Consistency of subjects with respect to the degree course goals* for the Master's degree in Marketing Management in the academic years 2012-13 and 2015-16 and the *Teachers celerity to replay to student emails* concerning the students enrolled in the Bachelor's degree in Economics in the academic years from 2012-13 to 2014-15.

The dissimilarity index is always lower than 0.1, indicating a good fit of the estimated MUB models.
The aspects of Teaching which show a significant and positive influence in defining its overall satisfaction are: *Consistency of subjects with respect to the degree course goals, Clearness of lessons and practice exercises, Teachers' celerity to reply to student emails* and *Teachers' respect of class hours* (Table 1).
An inverse relation between *Drop-out rate* and *the estimated feeling for Teaching* is proved to be significant (Table 2). In particular, in order to investigate the relation between the detected items influencing Teaching the most (Table 1) and *Drop-out rate*, *Teachers' celerity to reply to student emails* and *Teachers' respect of class hours* show a significant inverse relation with the *Drop-out rate* (Table 3). The fit of the estimated linear models (Tables 1, 2, 3), assessed by means of a residual analysis, was good.

Table 1: The results of the estimated linear model for student feeling with Teaching versus the most relevant Teaching items detected by step wise regression

Covariate	***Coefficient***	***Std. Error***	***p-value***
Intercept	0.0588	0.0729	>0.05
Subject consistency with course goals	0.1545	0.0482	<0.05
Clearness of lessons and practice exercises	0.3113	0.1101	<0.05
Teachers' celerity to reply to emails	0.2550	0.0741	<0.05
Teachers' respect of class hours	0.2109	0.0878	<0.05

Table 2: The results of the estimated linear model for *Drop-out rate* versus *Student feeling for Teaching*

Covariate	***Coefficient***	***Std.Error***	***p-value***
Intercept	1.4138	0.2661	<0.05
Student feeling for Teaching	-1.8899	0.4080	<0.05

Table 3: The results of the estimated linear model for *Drop-out rate* versus *Teaching items*

Covariate	***Coefficient***	***Std.Error***	***p-value***
Intercept	1.3648	0.2595	<0.05
Teachers' celerity to reply to student emails	-0.8528	0.2557	<0.05
Teachers' respect of class hours	-1.0001	0.4048	<0.05

5 Final remarks

Among the several factors influencing students' satisfaction, this study focuses the attention on Teaching and its determinants. The findings deal with the Department of Economics at University of Foggia. It emerged that the satisfaction's determinants are strictly linked to some latent dimensions as responsiveness, assurance and reliability, more than to tangible ones, like teaching materials.

Another important result concerns the detection of an inverse relation between satisfaction and drop-out, as students are more willing to carry on their studies when they are satisfied. Deepening the data analysis, it was found that establishing an empathetic relationship with the teacher is a very strong determinant in students' retention. In a challenging and uncertain environment such as university, having a reference point in the teacher helps the student to complete his academic career successfully. Further developments would deal with a comparison among the results of student satisfaction surveys carried out in other universities in order to highlight possible common frameworks and to assess if there are significant differences due to their size, in terms of number of students enrolled and departments.

References

1. Cafarelli B., Crocetta C. An evaluation of the student satisfaction based on CUB Models. Eds: Alleva & Giommi. Topics in Theoretical and Applied Statistics. Springer, 73-83. (2016).
2. D'Elia, A., Piccolo, D. A mixture model for preference data analysis. Computational Statistics & Data Analysis. 49, 917–934 (2005). ISSN: 0167-9473.
3. Eurostat, Education and training statistical database. Source: European Labour Force Survey ad hoc module on young people on the labour market (2016). Eurostat News, Product code: DDN-20180404-1, (April 2018). Available at appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfso_16ymgnednc&lang=en.
4. Horstschräer J. University rankings in action? The importance of rankings and an excellence competition for university choice of high-ability students. Economics of Education Review, 31(6), 1162– 1176. (2012).
5. Iannario M. Piccolo D. Simone R. CUB: A Class of Mixture Models for Ordinal Data. R package version 0.1. (2016) Available at: cran.r-project.org/web/packages/CUB/CUB.pdf
6. Petruzzellis L., D'Uggento A. M., Romanazzi S., Student Satisfaction and Quality of Service in Italian Universities. Managing Service Quality. Vol. 16 (4), 349-364 (2006).
7. Piccolo D. (2006) Observed information matrix for MUB models. *Quaderni di Statistica* 8: 33-78.
8. Piccolo D. and D'Elia A (2008) A new approach for modelling consumers' preferences. *Food Quality and Preference* 19(3): 247-259.

Space-time nonparametric analysis for the Italian real estate market

Analisi non parametrica spazio-temporale per il mercato immobiliare italiano

Cappello C. and De Iaco S. and Palma M. and Pellegrino D. and Posa D.

Abstract Recently, the evolution of the real estate market is considered a key indicator of the well-being level of the economic system. In particular, positive performance of the housing sector could drive the economic growth as well as improve social and environmental conditions of a country. Thus, the study of the indices measuring the performance of this sector could be a useful tool for planning macroeconomic policies for a territory. In this paper, the behavior of a real estate index in the Italian provinces, from 2000 to 2018, will be studied through a nonparametric space-time geostatistical analysis. In particular, the probability that the amount of the transactions in the residential sector exceeds some fixed thresholds will be estimated for the period 2019-2021.

Abstract *Recentemente, l'andamento del mercato immobiliare è considerato un indicatore del livello di benessere economico. In particolare, una performance positiva del mercato immobiliare potrebbe guidare la crescita economica, nonché migliorare le condizioni sociali ed ambientali di un paese. Pertanto, lo studio degli indici che misurano le performance di questo settore potrebbe essere utile per pianificare idonei interventi politici che favoriscano la crescita economica di un territorio. Nel presente lavoro è proposto uno studio dell'andamento di un indice del mercato immobiliare nelle province italiane, nel periodo 2000-2018, mediante un'analisi geostatistica non parametrica. In particolare, sarà stimata la probabilità che le transazioni nel settore residenziale superino prefissati valori di soglia, per il triennio 2019-2021.*

Cappello C.
University of Salento, Complesso Ecotekne, e-mail: claudia.cappello@unisalento.it

De Iaco S.
University of Salento, Complesso Ecotekne, e-mail: sandra.deiaco@unisalento.it

Palma M.
University of Salento, Complesso Ecotekne, e-mail: monica.palma@unisalento.it

Pellegrino D.
University of Salento, Complesso Ecotekne, e-mail: daniela.pellegrino@unisalento.it

Posa D.
University of Salento, Complesso Ecotekne, e-mail: donato.posa@unisalento.it

Key words: space-time geostatistics, indicator kriging, real estate transactions

1 Introduction

The study of the performance of the real estate market could be useful in order to assess the socio-economic well-being of a country. As highlighted in the literature [9, 1, 3], this sector is strongly related to economic cycles and reflects their behaviors. In particular, housing market booms are associated with an increase in consumption and firm investment, that boost the economic growth. Moreover, it is well known a strong correlation between the housing market and some macroeconomic variables, such as Gross Domestic Product (GDP), unemployment rate and inflation rate. For instance, due to the global financial crisis started in 2008, after an expansion period (2000-2006), the Italian real estate market, as for the other European countries, was characterized by a recession (2007-2013), finally a slight recovery was recorded in the last 5 years. This counter-trend was certainly supported by the gradual economic resumption, as well as by the spread of forms of investment with more favourable conditions, e.g., lower interest rates [11].

In this context, the analysis of the space-time evolution of the real estate market indexes, such as the number of normalized transactions (NNT), is convenient for policy makers to identify new territorial development and/or investment strategies, for the next years. Moreover, the use of geostatistical techniques to analyze the spatial-temporal behavior of such variables can give valuable hints in predicting socio-economic scenarios as thoroughly described in [14]. Although in this last paper the convenience of using Geostatistics to model jointly the spatial and temporal variability exhibited by a real estate index was discussed, in the literature a nonparametric geostatistical approach has not been widely applied on economic variables.

Hence, the novelty of this paper regards the application of nonparametric space-time geostatistical methods to predict the probability that the Italian residential NNT (per 1000 inhabitants), in the period 2000-2018, exceeds some relevant thresholds. This approach has significant implications to detect important signs regarding the probability of a trend reversal of the cycles in the real estate market and highlights important dynamics of this economic sector. In particular, making inference on the spatio-temporal evolution of the probability law of the variable of interest, instead of the variable itself, is often preferable in this context, especially when the adoption of different programs of incentives for the economic growth depends on the probability of exceedance of a planned target, which can be autonomously fixed by national government bodies or dictated by extra national ones.

2 The Italian NNT in the period 2000-2018

The Italian yearly NNT, per 1000 inhabitants, concerning 86 provinces (except those provinces belonging to Sicily and Sardinia, as well as Gorizia, Trento and Bolzano provinces), for the period 2000-2018, has been analyzed. In particular, the data re-

garding the number of transactions are available for the residential, office, retail and industrial classes and are provided by the Italian Revenue Agency. However only the residential asset has been retained for the analysis since it involves the 90% of total transactions and stocks.
As briefly explained in Sect. 1, the Italian residential NNT is strongly related to the national economic cycles. In particular, after a positive span, started in the 1990s until mid-2000s, when a considerable growth of the real estate transactions is recorded, in the period 2007-2014, a worrying decline occurred. Nevertheless, it is important to underline that in the last 5 years of the period under study, the Italian NNT registered a positive increase.

Areas	Mean	1*st* Quartile	Median	3*rd* Quartile	Std. Deviation
Italy	11.28	8.09	10.80	14.14	2.96
Northern Italy	12.83	10.16	11.81	16.11	3.51
Central Italy	11.31	8.81	10.67	14.39	3.14
Southern Italy	8.42	6.64	8.56	9.89	1.91

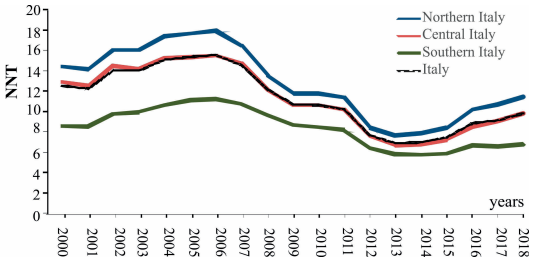


Fig. 1 Descriptive statistics and temporal evolution of the NNT average for Italy, Northern, Central and Southern part, in the period 2000-2018.

This positive national trend is even confirmed by considering a more dis-aggregated level, namely by exploring the NNT trend of the three macro-geographical areas (Fig. 1): Northern Italy (Friuli Venezia Giulia, Lombardia, Emilia Romagna, Piemonte, Valle D'Aosta, Liguria and Veneto regions), Central Italy (Lazio, Toscana, Marche and Umbria regions) and Southern Italy (Puglia, Abruzzo, Campania, Molise, Calabria and Basilicata regions).

Some differences among the three Italian sub-areas are evident: the lowest mean value of NNT (per 1000 inhabitants) can be found in the Southern part of Italy (8.42), while the highest one is observed in the Northern part of the nation (12.83). Moreover, the standard deviations are higher for the Northern and Central Italy (3.51 and 3.14, respectively), with respect to the Southern part of the country (1.91). However, a deeper analysis of the variability which characterizes the data, points out that the dispersion around the mean values is almost equal for the three Italian macro-geographical areas, indeed the coefficients of variation are 0.27, 0.28 and 0.23 for the North, Center and South, respectively. It is important to underline that the differences among the macro-geographical areas were significant from 2000 to 2006, during the real estate recovery, whereas in the following 7 years the amounts of the NNT were almost aligned around the historical minimum for the three areas.

These results confirm the need to study jointly the spatial and temporal profiles which intrinsically characterize the NNT evolution. Furthermore, the space-time indicator approach will allow estimating the probability that the variable under study records values greater than some critical thresholds, supporting the policy makers

to plan and carry out appropriate strategies for controlling and developing the real estate fluctuations which affect the economic growth.

3 Nonparametric space-time analysis

The NNT measurements at different time points and spatial locations can be considered as a finite realization of a second-order stationary spatio-temporal random field ($STRF$), $\{Z(\mathbf{u}), \mathbf{u} = (\mathbf{s},t) \in D \times T\}$, where $D \subseteq \mathbb{R}^d$ (usually $d \geq 3$) and $T \subseteq \mathbb{R}$. It is worth noting that the sample spatial points are the centroids of the Italian provincial polygons, which describe the shapes of these territories.

In nonparametric context, given a $STRF$ Z and a fixed threshold $z \in \mathbb{R}$, the spatio-temporal indicator random field ($STIRF$), denoted with $\{I(\mathbf{u},z), \mathbf{u} = (\mathbf{s},t) \in D \times T\}$, is such that:

$$I(\mathbf{u},z) = \begin{cases} 1 & \text{if } Z(\mathbf{u}) \geq z, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and under the second-order stationarity, the spatio-temporal indicator variogram depends on the threshold z and the lag vector $\mathbf{h}$, i.e. $2\gamma_I(\mathbf{h};z) = E\left[I(\mathbf{u}+\mathbf{h};z) - I(\mathbf{u};z)\right]^2$, where $\mathbf{h} = (\mathbf{h}_s, h_t)$. with $(\mathbf{s}, \mathbf{s}+\mathbf{h}_s) \in D^2$ and $(t, t+h_t) \in T$.

The nonparametric analysis of the Italian NNT has been conducted using four thresholds: the $1st$, $2nd$ and $3rd$ quartiles, as well as the 2018 average value (9.81) of the variable under study, i.e.:

$$I_1(\mathbf{u};8.09) = \begin{cases} 1 & \text{if NTN} \geq 8.09 \\ 0 & \text{otherwise,} \end{cases} \quad I_2(\mathbf{u};10.80) = \begin{cases} 1 & \text{if NTN} \geq 10.80 \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(\mathbf{u};14.14) = \begin{cases} 1 & \text{if NTN} \geq 14.14 \\ 0 & \text{otherwise,} \end{cases} \quad I_4(\mathbf{u};9.81) = \begin{cases} 1 & \text{if NTN} \geq 9.81 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

For each indicator threshold, the empirical spatio-temporal indicator variogram can be modelled by using one of the spatio-temporal models proposed in the literature [2, 6, 7, 10, 13]. In this paper, the space-time dependence has been modelled by using the generalized product-sum model, which is one of the most flexible in the fitting step [6].

After computing the sample spatio-temporal variogram for each indicator threshold (Fig. 2), the following product-sum models have been fitted (Fig. 3):

$$\gamma_{I_i}(\mathbf{h}_s, h_t; z) = c_{s_i} Exp(\|\mathbf{h}_s\|; a_{s_i}) + c_{t_i} Sph(h_t; a_{t_i}) - k_i \left[c_{s_i} Exp(\|\mathbf{h}_s\|; a_{s_i}) \cdot c_{t_i} Sph(h_t; a_{t_i})\right]$$

where $i = 1, \ldots, 4$, $z = (8.09, 10.80, 14.14, 9.81)$, $c_s = (0.08, 0.10, 0.08, 0.09)$, $c_t = (0.15, 0.17, 0.15, 0.16)$, $a_s = (120, 106, 86, 65)$, $a_t = (5.0, 5.5, 6.0, 5.5)$ and $k = (5.88, 3.82, 3.38, 3.43)$. Note that $Exp(\cdot;a)$ and $Sph(\cdot;a)$ denote the well-known exponential and spherical variogram models, with practical range a [8], while k_i, $i = 1, ..., 4$, are the parameters of the spatial-temporal interaction, and are computed such that the admissibility condition is satisfied [6].

Moreover, the fitted global sill, identified through the graphical inspection of each

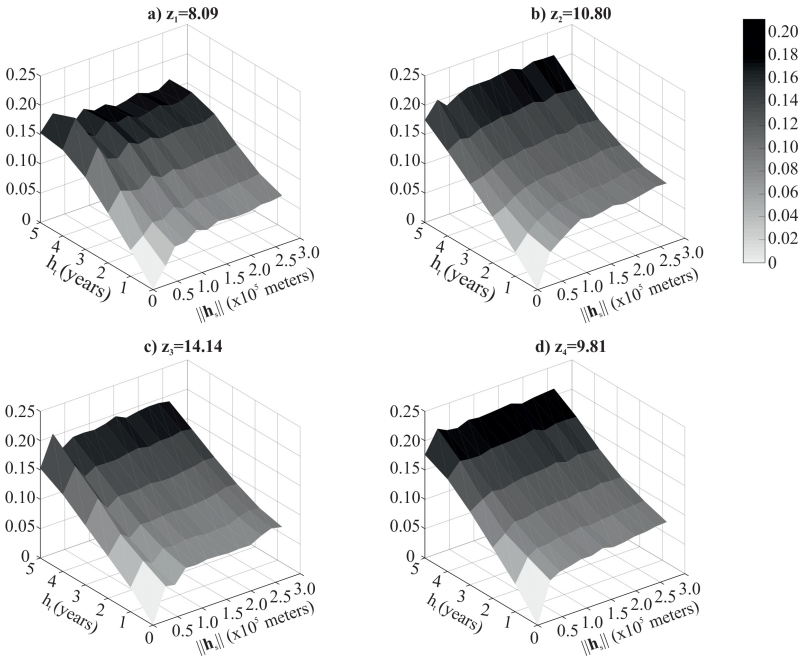


Fig. 2 Sample spatio-temporal indicator variogram surfaces for a) I_1, b) I_2, c) I_3, d) I_4.

spatio-temporal indicator surface, is 0.16 in γ_{I_1}, 0.21 in γ_{I_2}, 0.19 in γ_{I_3} and 0.20 in γ_{I_4}.

The reliability of the spatio-temporal models has been evaluated through

a) two fitting indexes, that is the *Mean Error* (ME) and the *Root Mean Square Error* (RMSE), based on the fitting errors between the empirical variogram surface and the corresponding model,
b) the cross-validation technique, by which the value observed at each sample location is estimated by using all the other sample values and the fitted spatio–temporal variogram model.

In particular, for all thresholds, the values of the ME and RMSE are within the ranges $[0;0.03]$ and $[0;0.28]$, respectively, confirming the accuracy of the fitted spatio-temporal models.

Moreover, the linear correlation coefficient, between the observed values and the estimates obtained by the cross-validation technique, is always greater than 0.85, pointing out the goodness of the fitted indicator variogram models.

4 Spatio-temporal nonparametric predictions

The probability that the spatio-temporal Italian NNT (per 1000 inhabitants) is not smaller than the threshold z at an unsampled point $\mathbf{u}$ is predicted by using the linear

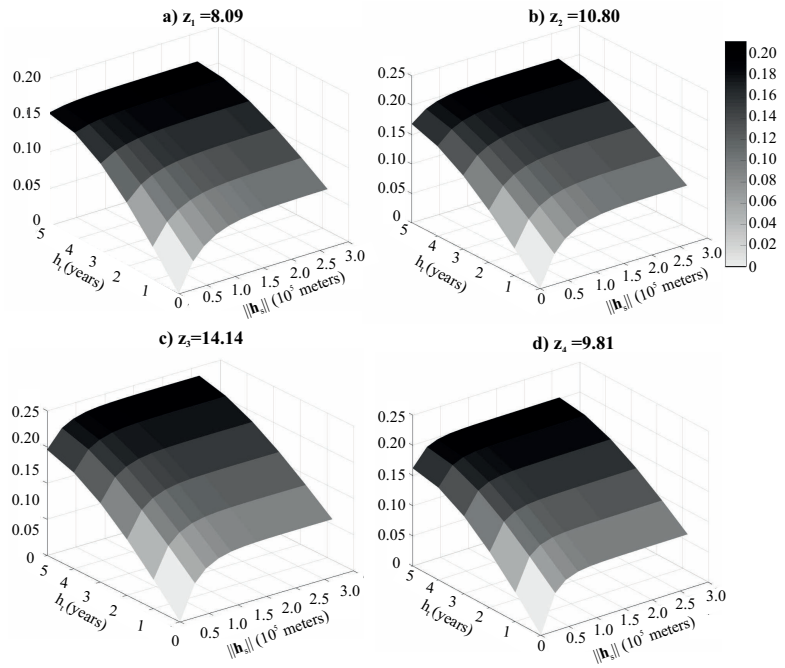


Fig. 3 Spatio-temporal indicator variogram models for a) I_1, b) I_2, c) I_3, d) I_4.

combination of neighbouring indicator variables: $\widehat{I}(\mathbf{u};z) = \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}_\alpha;z) I(\mathbf{u}_\alpha;z)$, where $I(\mathbf{u}_\alpha;z)$, $\alpha = 1,\dots,n$, are the indicator random variables at the sampled points $\mathbf{u}_\alpha \in D \times T$ and $\lambda_\alpha(\mathbf{u}_\alpha;z)$ are the kriging weights, determined by solving the indicator kriging system [12].

The product-sum models fitted for each threshold (Sect. 3) have been applied in order to produce spatio-temporal indicator kriging predictions over the area of interest for the period 2019-2021, through a modified *GsLib* routine [4]. Then, the probability maps of the Italian NNT (per 1000 inhabitants) exceeding the fixed thresholds have been obtained (Fig. 4).

It is interesting to point out that the geostatistical forecasts highlight a constant growth of the probability of exceeding the cut off values over time.

In particular, Figs. 4-a), b) and c) show the probability maps for the thresholds associated to the 1*st*, 2*nd* and 3*rd* quartile, respectively. For each year, the maps are useful for analyzing the spatial variations of the probability that the residential NNT exceeds the above mentioned thresholds, hence these tools are helpful to detect the provinces in which a stronger growth of the real estate market will occur.

For the predicted period (2019-2021), a recovery of the NNT has been estimated throughout the Northern and Central provinces, since the probability of exceeding the 1*st* quartile is greater than 0.6 (Fig. 4-a)). In fact, this result is recorded for all provinces belonging to the Northern part and for the 80.9% of those located in the Central part of Italy (12 provinces out of 21). On the other hand, for these sub-

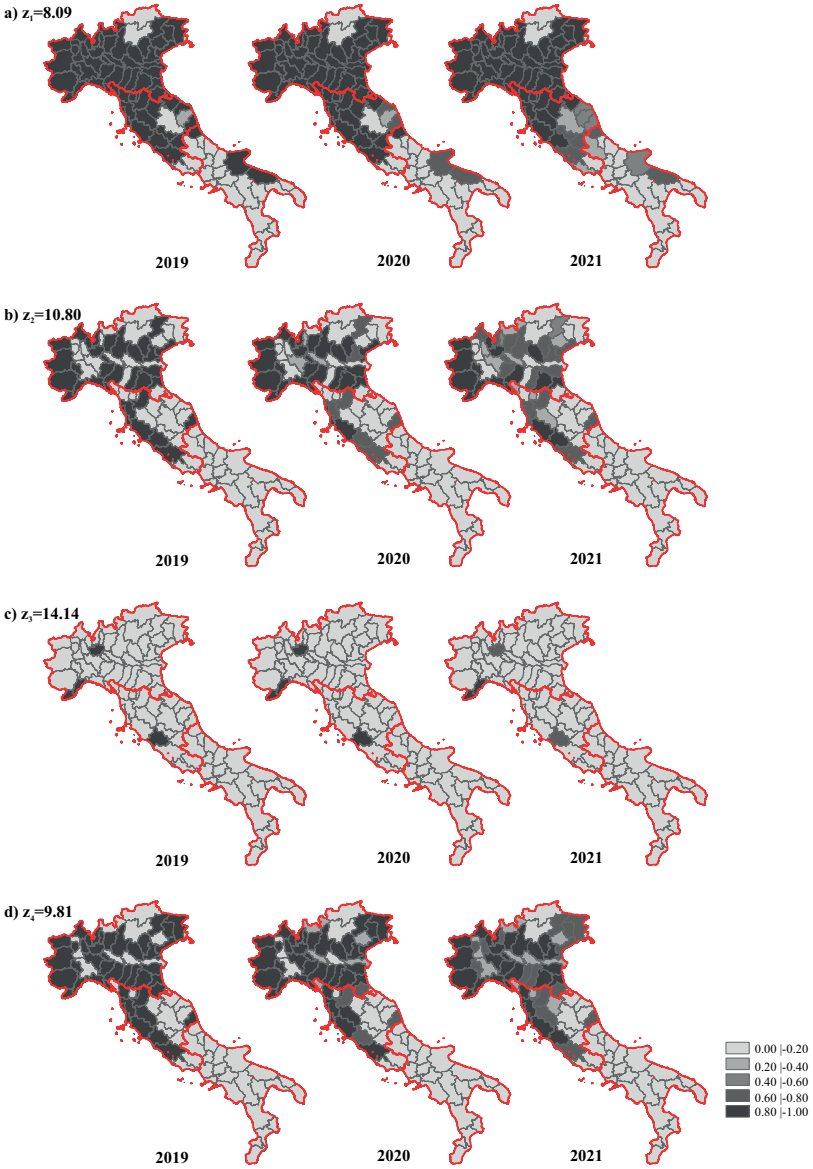


Fig. 4 Probability maps of the Italian NNT, per 1000 inhabitants, for a) $I_1 = 8.09$, b) $I_2 = 10.80$, c) $I_3 = 14.14$, d) $I_4 = 9.81$, for the period 2019-2021.

areas a less extended improvement in the residential real estate market is evident if the 2*nd* quartile is considered (Fig. 4-b)): for the 61.9% (26 out of 42) and 33.3% (7 out of 21) of the provinces in the Northern and Central parts of Italy, respectively, the predicted level of probability that the NNT exceeds this cut off is greater than 0.6. Note that only for 3 provinces out of 23 belonging to the Southern part (13.0%), high probabilities that the NNT exceeds the 2*nd* quartile are estimated. Moreover, a resumption of this economic sector, at an even stronger pace, occurs only in few provinces of the Northern and Central sub-areas (7.1% and 4.8%, respectively), since there is a high probability that the variable under study assumes values greater than the 3*rd* quartile (Fig. 4-c)). Finally, in Fig. 4-d) the probability maps that the analyzed phenomenon exceeds the 2018 NNT national average are given. From these figures it is evident that only in the Northern and Central Italy there are signs of recovery of the analyzed sector, since in the 80.9% and 47.6% of the provinces, respectively, high probability of exceeding the threshold value has been estimated. Hence, this result highlights a positive growth of the real estate market in these areas with respect to the NNT mean level registered in 2018.

References

1. Baltagi, B.H., Fingleton, B., Pirotte, A.: Spatial lag models with bested random effects: An instrumental variable procedure with an application to English house prices. J. Urban Econ. **80**, 76–86 (2014)
2. Cressie, N., Huang, H.: Classes of nonseparable, spatial-temporal stationary covariance functions. J. Am. Stat. Assoc. **94**(448), 1330–1340 (1999)
3. Curto, R., Fregonara, E., Semeraro, P.: Listing behaviour in the Italian real estate market. Int. J. Hous. Mark. Anal. **8**, 97–117 (2015)
4. De Iaco, S., Posa, D., Predicting spatio-temporal random fields: some computational aspects. Comput. & Geosci. (2011), doi: 10.1016/j.cageo.2011.11.014
5. De Iaco, S., Posa, D.: Positive and negative non-separability for space–time covariance models. J. Stat. Plan. Inference **143**, 378–391 (2013)
6. De Iaco, S., Myers, D.E., Posa, D.: Space-time analysis using a general product-sum model. Statist. and Probab. Lett. **52**(1), 21–28 (2001)
7. Dimitrakopoulos, R., Luo, X.: Spatiotemporal modeling: covariance and ordinary kriging systems. In: Dimitrakopoulos, R. (eds.) Geostatistics for the next century, pp. 88–93. Kluwer Academic Publishers, Dordrecht (1994)
8. Deutsch, C.V., Journel, A.G.: GSLib: Geostatistical Software Library and User's Guide, Oxford University Press, New York (1998)
9. Ferreira, F., Gyourko, J., Tracy, J.: Housing bust and household mobility. J. Urb. Econ. **68**, 34–45 (2010)
10. Gneiting, T.: Nonseparable, stationary covariance functions for space-time data. J. Am. Stat. Assoc. **97**(458), 590–600 (2002)
11. Italian Revenue Agency: Rapporto immobiliare 2018 - Settore residenziale (2018). https://www.agenziaentrate.gov.it
12. Journel, A.G.: Nonparametric estimation of spatial distributions. Math. Geol. **15**(3), 445–468 (1983)
13. Ma, C.: Linear combinations for space-time covariance functions and variograms. IEE Trans. Signal Process. **53**(3), 489–501 (2005)
14. Palma, M. Cappello, C., De Iaco, S., Pellegrino, D.: The residential real estate market in Italy: a spatio-temporal analysis. Qual. Quant. **53**(5), 2451–2472 (2019)

A Multidimensional Approach for Classifying Italian University Students by Mobility

Un approccio multidimensionale per la classificazione degli studenti universitari italiani secondo la mobilità

Sara Casacci

Abstract The Italian university system is characterized by both long-range and medium-range flows of students. Based on a novel dataset from integrated administrative sources, this work tries to profile students in tertiary level education by modelling the distance between hometown and destination in relation to (1) socio-demographic characteristics of students and their families and (2) territorial aspects. Results confirms a large incidence of movers from southern regions to universities located in the Centre and North. Besides, factors affecting mobility range are identified. Such kinds of information can be useful for collective services dimensioning and in the housing needs analysis.

Abstract *Il sistema universitario italiano è caratterizzato da flussi di studenti sia a lungo che a medio raggio. Sfruttando dati da fonti amministrative integrate, questo lavoro si propone di classificare gli studenti italiani. La distanza tra residenza e luogo destinazione viene modellata in relazione a (1) caratteristiche socio-demografiche degli studenti e delle loro famiglie e (2) aspetti territoriali. I risultati confermano la presenza di flussi consistenti dalle regioni meridionali alle università situate nel Centro e Nord. Vengono inoltre identificati i fattori che influiscono sul raggio degli spostamenti. Questo tipo di informazioni si rivela utile per il dimensionamento dei servizi collettivi e per l'analisi dei fabbisogni abitativi.*

Key words: Student Mobility, Classification, University

1 Background

The debate on students' mobility in Italy is based on evidence supplied by several studies. A large imbalance of mobility flows between northern and southern regions

[1] Sara Casacci, Italian National Institute of Statistics; email: Casacci@istat.it

is a well-established feature of the Italian academic system, with a high incidence of movers from southern regions to universities located in the Centre and North (Dal Bianco et al., 2009, Vivio, 2016). The average distance between the hometown and the chosen university has significantly increased in the last decade among students from the South of Italy. Among new entrants in first level courses, the number of movers from the South and Islands to the Centre-North increased by 10 per cent in the period from 2007 to 2015 (De Angelis et al., 2017).

The probability of moving is correlated with individual characteristics (such as gender, age and schooling background), with the local supply of academic courses and with the job prospects offered by the hosting university (De Angelis et al., 2016). Family income and cultural background also play a role in the decision to move insofar as they reduce the set of opportunities for students from poorer or less educated families (Lupi and Ordine, 2009). The family educational background has often been recognized as an important factor in determining the investment in human capital and the schooling decision (Checchi, 2003). On the other hand, a study by Pigini and Staffolani (2016) found that higher quality institutions attract more talented students independently from their social status and family background, thus encouraging student selection, which is in turn profitable to the local labour market.

There is also some evidence that mobility is positively associated with the quality of life in cities and educational supply (Dal Bianco, 2007). The increase in student flows both between and within geographical macro-areas probably signals a reallocation towards more attractive universities and prospering local labour market (Dotti et al., 2013). The empirical evidence discussed in some studies (Bratti and Verzillo, 2015; Ciriaci, 2014) suggests that universities' research and teaching quality are explanatory variables in the migration choice of the youngest and most skilled part of the Italian labour force. Other economic characteristics of the areas of destination, such as the availability of fast transport services, may as well influence student mobility in Italy by mitigating the negative effect of distance and increasing university accessibility (Cattaneo et al., 2015).

Based on a novel dataset from integrated administrative sources, this paper aims to classify the new entrants in first level tertiary education, by modelling the distance between hometown and destination in relation to students' socio-demographic characteristics, family income and educational level, and universities' features. This work presents the results of a supplementary analysis conducted within the Italian National Institute of Statistics (Istat) Project "Studenti e bacini universitari" (Vivio, 2016).

2 Data and Method

The work is based on an extensive use of information contained in administrative sources (acquired by Istat for various purposes) properly treated and integrated. Data are derived from the integration of several administrative archives, listed in Table 1. The main source is the University Students Register, held by the Ministry of

Education, University and Research, which collects information about enrolments, students' schooling background, chosen university, degree course and municipality of enrolment. The integration of the other administrative sources permits an informational enhancement and the identification of students' parents. In particular, the available sources make it feasible to identify the parents of the students in the following cases:

- students are dependent family members (regardless of the co-residence with their parents);
- students and their parents are part of the same household (they are co-resident).

Record linkage was carried out by a dedicated structure in Istat, the 'Integrated Micro-data System' (SIM – Sistema Integrato di Microdati)[1].

Table 1: List of administrative sources

Source	***Information***
Student Registers	Enrolments; students' schooling background, chosen university
Social Security Sources (workers)	Students' occupational status
Tax Returns Register	Identification of parents; parents' income
Municipal Population Registers	Students' municipality of residence; identification of parents
Population Census	Identification of parents; parents' educational level

The final dataset consists of 206,627 students enrolled the first year of tertiary education in the academic year 2014-2015, whose municipality of usual residence differs from the university's location[2]. Students of online universities were not considered.

On the integrated dataset, a Classification tree was performed using Chi-square Automatic Interaction Detector (CHAID) (Kass, 1980), specifying the distance[3] (in kilometres) between hometown and university location as the dependent variable. The classification method adopted is of a hierarchical type. CHAID "builds" non-binary trees (i.e., trees where more than two branches can attach to a single root or node), based on an algorithm that is particularly well suited for the analysis of larger datasets. CHAID creates all possible cross tabulations for each categorical predictor until the best outcome is achieved and no further splitting can be performed. In the progressive partition of students, the characteristics considered are in succession the 'branches' and 'leaves' of the 'classification tree', whose ''root' represents the set of students. Like other decision trees, CHAID's advantages are that its output is highly visual and easy to interpret.

The independent variables included in the model are:

1. Sex (male, female);

[1] SIM is the main repository for administrative data in Istat. It is an integrated system since (a) it identifies each object in the administrative sources with a unique and stable (over time) ID number ; (b) it defines, for each object, the logical and physical relationships among different administrative data sources (Ambroselli and Garofalo, 2015).

[2] In the academic year 2014-2015 the number of new enrollments was about 260,000 (Miur, 2015).

[3] Distances are derived from the matrix of the distances between the Italian municipalities (Istat, 2017).

2. Occupational status (employed, not employed);
3. Secondary school grade (high: greater than 84/100, medium and low: less than or equal to 84/100);
4. Macro-area of residence (Centre & North 'CN', Sud & Island 'SI');
5. Urbanization of municipality of residence (high density, intermediate density, low density-rural);
6. The Municipality of residence type (not mountain; totally mountain; partially mountain);
7. Type of university (private, public);
8. Field of study (STEM, non-STEM);
9. Parents' income (high: greater than 55,000€, medium: between 20,000€ and 55,000€, low: less than 20,000€);
10. Parents' educational level (up to lower secondary education 'LSE', upper secondary education 'USE', tertiary education 'TE').

3 Results

Data analysis confirms that universities located in the CN receive mobility flows from the SI. In fact among new entrants in first level courses, the percentage of movers from the SI to the CN is 26.2 per cent (Table 2).

Table 2: University entrants by area of residence and enrollment (raw percentages)

Area of residence	*Area of enrollment*		
	North	***Centre***	***South & Islands***
North	97.0	2.6	0.5
Centre	10.5	85.6	4.0
South & Islands	13.2	13.0	73.8

Movements across geographical macro-areas are only a part of the phenomenon. In particular for the CN, although long-range outflows are modest, the number of students moving to a different municipality in the same macro-area is relevant.

A multidimensional approach (Classification tree) was carried out in order to identify groups of students taking into account different elements. The final model includes six independent variables: area of residence, urbanization and type of municipality of residence, secondary school grade, private or public university, parents' income, parents' educational level. Sex, occupational status and field of study are not significant, so they are not included in the final model. Terminal nodes are 13 (Table 3). They represent the best classification forecasts for the model.

The group (node 20) with the lower average distance (48 kilometers) is composed by students from North and Centre, living in a not mountain municipality whose parents achieved a lower secondary education. The group (node 20) with the greatest average distance (515 kilometers) is composed by students from South and

Islands, enrolled in a private university, with an high school grade greater than eighty-four.

Table 3: Terminal nodes of classification tree

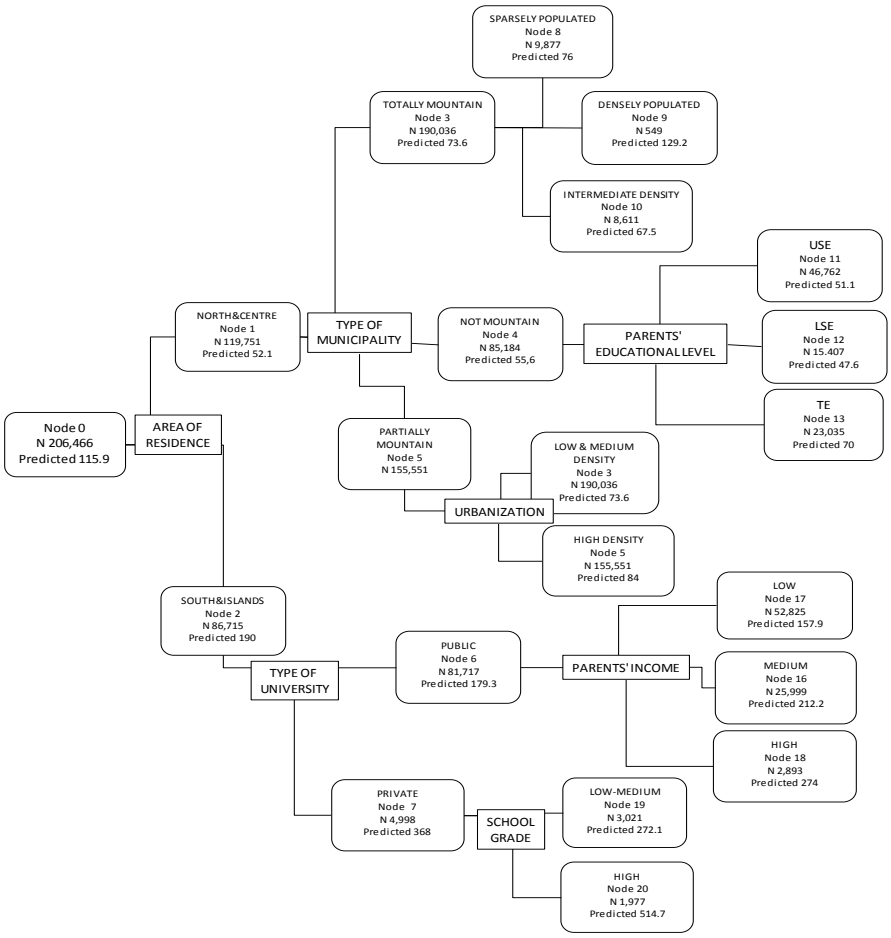
Node	*Number of students*	*Percent (%)*	*Avarage distance (in kilometers)*
8	9,877	4.8	76
9	548	0.3	129
10	8,611	4.2	67
11	46,722	22.6	51
12	15,407	7.5	48
13	23,035	11.2	70
14	11,834	5.7	68
15	3,717	1.8	134
16	25,999	12.6	212
17	52,825	25.6	158
18	2,893	1.4	274
19	3,021	1.5	272
20	1,977	1.0	515

The tree diagram is reported in Figure 1. The first variable to define the first two 'branches', and therefore, to divide students into two large groups is the macro-area of residence: students from CN travel - on average - 52 kilometres, students from SI 190 kilometres. In fact, as documented in the previous literature, long-range mobility is largely due to the movement of students from southern regions in universities located in the Centre-North of the country. The tree diagram for students from CN shows that their mobility is dependent on urbanization and type of municipality of residence and from the parents' educational level. In particular, students whose parents achieved a tertiary education have a wider range of mobility than students with parents with a lower education.

The first classification variables for students from SI is the type of the chosen university: private universities are more attractive than public ones. Also, results of previous studies indicated that private universities attract students and increase their willingness to travel longer distances (Turk, 2017). For students moving to a public university, family income is a strongly discriminating factor. In particular, it discriminates between situations in which a modest or medium income (up to 55,000€) is a barrier to long-range mobility and those in which a high income allows it. As documented by previous research, the cost of living away from home is one of the most significant barriers to access to the Italian university system (Catalano and Fiegna, 2003).

The group of students moving to a private university consists of 4,998 individuals with a predicted distance between home and university location of 368 kilometres (node 7). Secondary school grade is a strongly discriminating variable for this group: the predicted distance between hometown and university for new enrolled in a private university from the SI with school grade greater than 84/100 is 515 kilometres.

Figure 1: Tree diagram



4 Conclusions

As known, the Italian university system is characterized by both long-range and medium-range flows of students, stimulated or refrained by a series of push and pull factors. The present work, tries to profile students in tertiary level education, by using a non-parametric technique for data classification. In particular the groups are the result of a multidimensional approach, which is able to take into account different aspects. Type of municipality of residence, school grade, family income and parents' educational level, in addition to the area of residence and type of

university, are factors affecting mobility range; whereas sex, occupational status, and field of study are not significant.

Some crucial considerations emerges from the analysis. The first is that students from wealthier families and those whose parents achieved higher education levels are more likely to increase the distance between hometown and university location. In addiction to confirming the Italian limited social mobility, this finding imply that, if brain circulation and human capital accumulation are to be encouraged, public support of mobility to study in the best-performing universities, especially for less wealthy students, should be promoted.

The second consideration is that secondary school grade is a strongly discriminating variable, in particular for students of private universities. A large body of research affirms that the decision whether to move to study influences subsequent (post-graduation) migration behaviour (Bacci et al., 2008; Faggian et al., 2007). Therefore, the loss of human capital due to the outflow of skilled students contributes to compromise the growth in the areas of origin (mostly the southern regions), a process already well documented in literature (Fratesi and Percoco, 2009).

Further analysis can improve the classification model by adding variable related to the urban context and to university quality and describe the territorial trajectories of the flows. Beside, an alternative model could be built using travel times as dependent variable.

References

1. Ambroselli S., Garofalo G.: Reversing the flow: from an integrated system of administrative microdata to an infrastructure for the users, NTTS 2015. Available at:https://ec.europa.eu/eurostat/cros/system/files/Ambroselli_etal_NTTS2015abstract_SIM_ARCHIMEDE_0.pdf (Accessed 18 March 2018) (2015)
2. Bacci, S., Chiandotto, B., Di Francia, A., Ghiselli, S.: Graduates job mobility: a longitudinal analysis, Statistica 3–4, pp. 255–279 (2008)
3. Bratti, M., Verzillo, S.: The Gravity of Quality: Research Quality and Universities' Attractivity in Italy, Mimeo (2015)
4. Catalano, G., Fiegna, G.: La Valutazione del Costo degli Studi Universitari, il Mulino (2003)
5. Cattaneo, M., Malighetti, P., Paleari, S., Redondi, R.: Evolution of long distance students' mobility: the role of transport infrastructures in Italy, ERSA conference papers, European Regional Science Association (2015)
6. Checchi, D.: The Italian Educational System: Family background and social stratification, Working Paper 2003-01, University of Milan (2003)
7. Ciriaci, D.: Does university quality influence the interregional mobility of students and graduates? The case of Italy, Regional Studies, 48 (10), pp. 1592-1608 (2014)
8. Dal Bianco, A.: Determinants of student migration in Italy, XXVIII Conferenza Italiana di Scienze Regionali, Bolzano, 28-28 Settembre (2007)
9. Dal Bianco, A., Poggi, E., Spairani, A.: La mobilità degli studenti in Italia, IRER Working Paper, 12 (2009)
10. De Angelis I., Mariani, V., Modena, F., Montanaro, P.: Academic enrolments, careers and student mobility in Italy, Questioni di Economia e Finanza (Occasional Papers), No. 354, Banca d'Italia (2016)
11. De Angelis, I., Mariani, V., Torrini, R.: New evidence on interregional mobility of students in tertiary education: the case of Italy, Questioni di Economia e Finanza, Banca d'Italia (2017)

12. Dotti, N. F., Fratesi, U. Lenzi, C., Percoco, M.: Local labour markets and the interregional mobility of italian university students, Spatial Economic Analysis, 8 (4), pp. 443-468 (2013)
13. Faggian, A., McCann, P., Sheppard, S.: Human capital, higher education and graduate migration: an analysis of Scottish and Welsh students, Urban Studies 44 (13), pp. 2511–2528 (2007)
14. Fratesi, U., Percoco, M.: Selective migration and regional growth: evidence from Italy, Bocconi Working Paper (2009)
15. Istat: Matrici di contiguità, distanza e pendolarismo https://www.istat.it/it/archivio/157423 (2017)
16. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data, Applied Statistics, Vol. 29, No. 2, pp. 119-127 (1980)
17. Lupi, C., Ordine, P.: Family income and students' mobility, Giornale degli Economisti, 68(1), pp. 1-23 (2009)
18. Miur: Focus: gli immatricolati nell'anno accademico 2014/2015 (2015)
19. Pigini, C., Staffolani S.: Beyond participation: do the cost and quality of higher education shape the enrollment composition? The case of Italy, Higher Education, 71 (1), pp. 119-142 (2016)
20. Turk, U.: Socio-economic determinants of student mobility and inequality of access to higher education in Italy, Working Paper Series, Department of Economics, University of Verona (2017)
21. Vivio, R. (ed.): Studenti e Bacini Universitari, Letture statistiche Istat (2016)

Robust analysis of the labor market

Analisi robusta del mercato del lavoro

Aldo Corbellini, Marco Magnani and Gianluca Morelli

Abstract The work presents a robust approach to labor share analysis. The estimate of labor share presents various complexities related to the nature of the data sets to be analyzed. Typically, labor share is evaluated by using discriminant analysis and linear or generalized linear models, that do not take into account the presence of possible outliers. Moreover, the variables to be considered are often characterized by a high dimensional structure. The proposed approach has the objective of improving the estimation of the model using robust multilevel regression techniques and data transformation.

Key words: labor share, robust multivariate regression, data transformation

1 Introduction

The analysis of the labor share is a field of analysis which involves both the macro and the micro level. The relevance of this issue indeed is mostly related to the empirical analysis of the level and evolution of the aggregate labor share. A large share of the theoretical literature however, has studied the dynamics and the determinants of the labor share at the micro level. This contradiction has been solved converging to a paradigm where the macro level is concealed with the micro level by assuming that a representative firm is operating in the economy. This approach characterizes most of the literature. In particular, since the seminal analysis of Bentolila and Saint Paul (2003) [6] where, the theoretical determinants of the labor share are summarized in the definition of the SK schedule, several studies have tried to provide an explanation for the persistent declining trend of the labor share identifying different causes for it. Most of these causes have to do with the behaviour of the

Aldo Corbellini, Marco Magnani and Gianluca Morelli
University of Parma, Department of Economics and Management, Italy, e-mail: aldo.corbellini@unipr.it, marco.magnani@unipr.it, gianluca.morelli@unipr.it

representative firm, and thus concern the micro level. They include the elasticity of substitution between labor and capital (Bentolila and Saint-Paul, 2003; Lawless and Whelan, 2011 [11]; Antras, 2004 [2]) capital deepening Piketty and Zucman, 2014 [13]), workforce composition (Arpaia et al., 2009 [3]; Elsby et al., 2013 [8]) capital composition Lawless and Whelan, 2011). Firm relative bargaining power is further affected by a large set of factors ranging from product market competition (Autor et al., 2017 [5]) and to the effects of international trade, offshoring and globalization (Guscina, 2006 [9]); These determinants of the labor share operate differently in different firms and interact within it either strengthening or weakening their effects. It follows that, adopting a description of the economy where only a representative firm operates does not allow to investigate these mutual interactions nor the channels through which they affect the macro level. These shortcomings are specially relevant when, as shown by the results of several studies, a substantial fraction of the decline in national labor shares can be ascribed to changes in the sectoral composition of the economy (Moral and Genre, 2007 [12]). These authors suggest indeed that declines in the industry-level labor share may be driven by shifts toward firms with below-average labor shares rather than by within-firm changes in labor shares. The present paper analyzes the determinants of the labor share at the micro level using a large set of firm-level data and aims at investigating many issues that in the empirical analyses at the macro level are discarded. In particular, we will discuss the role of the elasticity of substitution between productive factors and its interactions with the main structural factors of the firm as its size and its sector of activity.

2 The case of Italy

The declining trend in the Italian labor share reversed at the beginning of 2000's and, excluding the housing sector, reached historically high levels during the Great Recession (Torrini, 2016) [18]. The aggregate labor share for the whole Italian economy reaches a maximum in the mid-1970s. After this historical high the Italian labor share starts a declining trend which leads to a new point of minimum at the beginning of the 2000s. This evidence is coherent with what was observed in many advanced economies. A trend reversal then occurs which is not generally observed in other countries. This recovery starting well before the Great Recession, is the peculiarity of the Italian case where a medium-run upward trend of the labor share is observed which goes beyond the cyclical fluctuations induced by the regression.

This trend, started in the second half of the 1990s in manufacturing and in most others business sectors, is not evident in the data due to the effects of the transformations which occurred in this period in the regulated sectors (RS): energy, transport, communication and finance. The slight increase in the labor share in non-regulated sectors in facts has been more than offset by the sharp decline in the labor share of RS, hiding its effects at the aggregate level. Since 2001 an upward trend has prevailed in the labor share of manufacturing and business sectors other than RS. The

divergence with respect to RS, which was already evident at the end of the 1990s, thus has further strengthened.

Torrini (2016) proposes an explanation for this recent upward trend in the Italian labor share, which focuses on the effects of a reduction in markups over marginal costs, possibly related to loss of competitiveness in the Italian economy. This reduction in markups is in part the outcome of the cyclical impact of the Great Recession, and is likely to reverse when a recovery of the economy will take place. But it is also the result of the inability of the Italian production system to cope with a more competitive and deeply integrated global market which requires more capacity of innovation. Italian firms in fact reacted to increased international competition by pursuing an innovation strategy based on product quality upgrading. This quality enhancement though has not been rewarded by the market with a proportional price increase, hence causing the dissipation of innovation efforts which only guaranteed firm survival. As a consequence firm markups have dropped, and a shift in the relation between factor shares and factor prices occurred which induced a reduction in profit margins, a slowdown in productivity and a rise in the labor share which took place mostly in the form of an increase in employment. A drop in profit margins is consistent with a declining user cost of capital which leads to an increase in the capital-output ratio.

3 The Model

The analysis of the firm-level determinants of the labor share relies on the analysis by Bentolila and Saint-Paul (2003) which is summarized by the *SK* schedule. This is a one-for-one technological relationship between the labor share, S, and the capital-output ratio, k, derived by assuming competitive markets, constant return to scale production function and labor-augmenting technical progress.

The empirical specification proposed by Bentolila and Saint-Paul (2003) and derived from this theoretical background, adopts a general multiplicative form for the production function:

$$S_{i,j,r} = \phi\left(k_{i,j,r}, A_{i,j,r}\right)\gamma\left(X_{i,j,r}\right) \tag{1}$$

where the subscript denotes respectively firm i, industry j, and region r. The term $\phi\left(k_{i,j,r}, A_{i,j,r}, \alpha_j\right)$ summarizes the characteristics of the production function and depends on the capital-output ratio, $k_{i,j,r}$, and on a measure of capital-augmenting technical progress $A_{i,j,r}$. The term $\gamma(X_{i,j,r})$ captures possible discrepancies between the marginal product of labor and the wage rate and depends on the vector $X_{i,j,r}$ which includes three variables: a measure of markup, $\mu_{i,j,r}$, a measure of labor adjustment cost, $c_{i,j,r}$, and the unemployment rate of region r, u_r, as a measure for worker relative bargaining.

Following Bentolila and Saint-Paul (2003), it is assumed that the functions $\phi\left(k_{i,j,r}, A_{i,j,r}, \alpha_j\right)$ and $\gamma(X_{i,j,r})$ are multiplicative so that:

$$\phi\left(k_{i,j,r},A_{i,j,r}\right)=\left(A_{i,j,r}\right)^{\beta_0}\left(k_{i,j,r}\right)^{\beta_1} \tag{2}$$

and

$$\gamma(X_{i,j,r})=\exp\left(\beta_2\cdot\mu_{i,j,r}+\beta_3\cdot c_{i,j,r}+\beta_4\cdot u_r\right). \tag{3}$$

Substituting the previous equations into Equation 1 and taking natural logarithms gives the basic estimated equation:

$$\ln S_{i,j,r}=\beta_{0j}\cdot\ln A_{i,j,r}+\beta_1\cdot\ln k_{i,j,r}+\beta_2\cdot\mu_{i,j,r}+\beta_3\cdot c_{i,j,r}+\beta_4\cdot u_r+v_{i,j,r} \tag{4}$$

where $v_{i,j,r}$ is the error term.

An augmented version of Equation 4, which can be referred as a multivariate linear regression model including a set of control variables $Z_{i,j,r}$, is estimated using cross-section data techniques.

4 The forward search for linear models

In the regression model $y=X\beta+\varepsilon$, y is the $n\times 1$ vector of responses, X is an $n\times p$ full-rank matrix of known constants, given in equation 4 with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ε_i are i.i.d. $N(0,\sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e=y-\hat{y}=y-X\hat{\beta}=(I-H)y$, where $H=X(X^TX)^{-1}X^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2=e^Te/(n-p)=\sum_{i=1}^n e_i^2/(n-p)$. The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$. To start we take $m_0=0$, since the prior information specifies the values of β and σ^2.

To test for outliers the deletion residuals are calculated for the $n-m$ observations not in $S^*(m)$. These residuals are

$$r_i(m)=\frac{y_i-x_i^{\mathrm{T}}\hat{\beta}_1(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}}=\frac{e_i(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_i(m)\}}}, \tag{5}$$

where the leverage $h_i(m)=x_i^{\mathrm{T}}\{X_0^{\mathrm{T}}X_0+X(m)^{\mathrm{T}}X(m)/c(m,n)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min}$ where

$$i_{\min}=\arg\min_{i\notin S^*(m)}|r_i(m)|.$$

To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{\mathrm{imin}}(m)=\frac{e_{\mathrm{imin}}(m)}{\sqrt{\hat{\sigma}^2(m)\{1+h_{\mathrm{imin}}(m)\}}}, \tag{6}$$

as a test statistic. If the absolute value of (6) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$.

In order to detect outliers and departures from the fitted regression model, FS uses least squares to fit the model to subsets of m observations. The initial subset of m_0 observations is chosen robustly, for example by least trimmed squares. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest residuals. For each m $(m_0 \leq m \leq n-1)$, we test for the presence of outliers, using the residual $r_{\text{imin}}(m)$ defined in (6).

In order to test for outliers we need a reference distribution for $r_i(m)$ in (5). If we estimated σ^2 from all n observations, the statistics would have a t distribution on $n-p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated normal distribution containing the central m/n portion of the full distribution be

$$\sigma_T^2(m) = 1 - \frac{2n}{m}\Phi^{-1}\left(\frac{n+m}{2n}\right)\phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\}, \tag{7}$$

where $\phi(.)$ and $\Phi(.)$ are respectively the standard normal density and c.d.f. See Riani, M., Atkinson, A. C., and Cerioli, A. (2009) [14], for a derivation from the general method of Tallis, G. M. (1963) [16]. We take as our approximately unbiased estimate of variance $s_T^2 = s^2(m)/\sigma_T^2 = s^2(m)/c(m,n)$. In the robustness literature, the important quantity $c(m,n)$ is called a consistency factor, Johansen and Nielsen, (2016) [10].

5 The dataset and variables

Our main sample of firms is composed of more than thirty thousand firms in a timespan of ten years going up to year 2017, representative of the manufacturing sector and extracted from the Buerau Van Dijk's AIDA data base that contains comprehensive information on capital companies in Italy. A rich set of information is collected by this survey, including firm-specific characteristics, investment and (international) trade activities. The model variables are as follows: Y: Labor share proxied by the ratio of labor cost to value-added; This indicator is an alternative version of the ratio of wage to total company assets. Using the added value instead of total assets, this variable can assume negative or positive values. X1: The ratio of tangible fixed assets to added value. The book-value of gross investments of this year has been adjusted to account for inflation using a measure of vintage. Then the deflated value of investments in the next years has been added using sectoral deflators for gross fixed capital calculated by the ISIC/ATECO assuming year 2007 as the base reference. X2: The ratio of intangible assets to total assets: this ratio measures the percentage of investments on intellectual capital, research and development and other intangi-

ble assets over the company total assets. X3: The ratio of industrial equipment to total asset: this variable measures the theoretical productive potential of the firm and is one of the primary drivers of company value. X4: Return On Sales, (ROS), that measures firm operating profitability proxied by the ratio of operating margins to sales. We expect a negative effect on the default risk, as the higher a firm's profitability the higher the flow of internal resources available to cover debt exposure should be; X5: measures the firm's interest burden, proxied by the ratio of firm total asset to net capital; high interest burden may worsen the financial risk associated with external finance. X6: Sales, X7: Age of the firm.

6 Transformations

The Box-Cox transformation produces a smooth relationship between $y(\lambda)$ and the original y determined by the value of λ. A non-parametric alternative is to use smoothing to estimate this relationship. Both methods can transform explanatory variables and response. The assumed model is a generalized additive model, that is one with transformations of both response and explanatory variables but without interactions. Both rely on repeated application of univariate smoothers. In ACE (alternating conditional expectations) Breiman and Friedman (1985) [7] maximize a measure of correlation between all variables; in regression the response variable is not treated as being different from the explanatory variables. Tibshirani (1987) [17] describes a related method in which the transformation for the response is intended to yield additivity and variance stabilization (AVAS). The asymptotic variance stabilizing transformation is applied to the response. Hastie and Tibshirani (1990, Chapter 7) provide a description of both ACE and AVAS with an emphasis on response transformation and the mathematical relationship to the Box-Cox transformation. The output of ACE and AVAS are a set of transformed responses The original programs for both ACE and AVAS are written in Fortran 77, without comments and with many non-informative variable names. This Fortran code also provides the basis of the R package Acepack. We have rewritten the programs in Matlab and also created an extensive documentation. These new programs have been thoroughly compared with the Fortran programs and validated to give identical numerical results; they are used in our calculations.

7 Results

Table 1 first column shows the results of the application of the multivariate regression model on the original data. The R^2 of the multivariate model is high, but the significance of β coefficients is very poor, except for the intercept. The main reason of this behaviour lies in the presence of extremely high leverage points in the data that are affecting the β estimates in the linear model; this effect is well know in

statistical literature and was first pointed out by Sastry and Nag (1990) [15] which summarized it in a theorem that states that $R^2 \to 1$ as the remoteness of the leverage units increases. Table 1 4th column shows the results of the application of the multivariate regression model after removing these outlying observations by means of the Forward Search, Atkinson and Riani (2000) [4] as described in the previous section. The new pseudo R^2 is very low, near 0.01 even if the significance of the β estimates now improves considerably and most of them are now significant.

	$\hat{\beta}_{ML}$		$\hat{\beta}_{MLFS}$		$\hat{\beta}_{MLFSdt}$	
Variable	Estimate	p-value	Estimate	p-value	Estimate	p-value
intercept	0.44924	6.9299e-05	0.68786	2.9993e-163	0.35525	2.5439e-239
TFA/AddVal	0.055476	0	0.018411	1.0344e-05	-0.051072	7.5262e-179
IA/TA	-0.65996	0.16129	-0.10365	0.29107	-0.66183	3.0051e-57
IE/TA	0.29917	0.78016	0.20128	0.59472	-0.49701	0.001858
ROS	0.00039737	0.95317	-0.02762	6.4095e-79	-0.094043	0
Debt ratio	0.00042643	0.70808	0.001237	0.10774	0.0019641	1.4619e-09
Sales	-1.3991e-08	0.97737	7.3252e-06	0.1368	-5.2079e-05	1.0936e-137
Age	0.0041725	0.29412	0.0020498	0.012091	0.0089754	3.1016e-148

Table 1 $\hat{\beta}$ comparison between non robust regression, robust regression and robust regression after data transformation

The analysis of the distribution of the regression residuals leads us to think that transformation of the response is required. To this purpose we use the non parametric conditional expectation methods (i.e. ACE and AVAS). Applying the transformations on the cleaned dataset we were able to dramatically improve the goodness of fit, $R^2 = 0.33$. The analysis of the results still shows the presence of several regression outliers, therefore we performed again the Forward Search to remove the atypical units. Table 1 6th column shows the results of the regression model applied on the clean transformed data. The new value of the pseudo $R^2 = 0.39$, all the variables are now highly significant and -finally- the signs of the coefficients are in agreement with those suggested by the economic theory. Note that this goal was reached removing a small percentage of units that were biasing the model estimates.

8 Discussion and conclusions

The present paper studies the determinants of labor share dynamics using the approach developed by Bentolila and Saint-Paul (2003), which characterizes a one-for-one relationship between the labor share and the capital output ratio, the *SK* schedule. The sign of the relationship depends on the elasticity of substitution between labor and capital. An elasticity larger than unity implies a negative relationship, an elasticity smaller than unity implies a positive relationship, and unit elasticity implies that the labor share is constant. In the present context, the coefficient multiplying the capital output ratio, measured as the book-value of tangible assets on value added, highlighting that, as in most of the literature using micro-data, the

productive factors capital and labor are largely substitute. Forward Search monitoring and outlier detection gives us the flexibility to find outlying units or even group of outliers that are hard to remove from the raw dataset. The forward search coupled with ACE transformations enabled us to detect outliers at successive stages and to find the correct structure of the data. All steps of robust analysis was performed on the (growing set of robust procedures) of FSDA toolbox for MATLAB, developed with a joint collaboration between the Department of Economics of the University of Parma and the European Joint Research Center which is freely downloadable from the Mathworks website and now also directly form MATLAB from the Add-Ons menu.

References

1. Alvarez, I. (2015). Financialization, non-financial corporations and income inequality: the case of France. *Socio-Economic Review*, 13(3), 449-475.
2. Antras, P. (2004). Is the US aggregate production function Cobb-Douglas? New estimates of the elasticity of substitution. *Contributions in Macroeconomics*, 4(1).
3. Arpaia, A., Pérez, E., & Pichelmann, K. (2009). Understanding labour income share dynamics in Europe. Economic Papers 379, European Commission.
4. Atkinson, A.C., & Riani, M.(2000). *Robust diagnostic regression analysis*. New York, Springer-Verlag.
5. Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2017). The fall of the Labor share and the rise of superstar firms. NBER workin paper No. w23396. National Bureau of Economic Research.
6. Bentolila, S., & Saint-Paul,G. (2003). Explaining movements in the labor share. *The BE Journal of Macroeconomics*, **3**(1), 1–33.
7. Breiman, L. & Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, Vol. 80, pp. 580–597.
8. Elsby, M. W., Hobijn, B., & Nahin, A. (2013). The decline of the US labor share. *Brookings Papers on Economic Activity*, 2013(2), 1-63.
9. Guscina, A. (2006). Effects of globalization on labor's share in national income. IMF Working Papers No. 2006-2294. International Monetary Fund.
10. Johansen, S., & Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, 43(2), 321-348.
11. Lawless, M., & Whelan, K. T. (2011). Understanding the dynamics of labor shares and inflation. *Journal of Macroeconomics*, 33(2),121–136.
12. Moral, E., & Genre,V. (2007). Labor share developments in the Euro area. *Economic Bulletin, Banco de España*, July 2007.
13. Piketty, T., & Zucman, G. (2014). Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010. *The Quarterly Journal of Economics* 129(3), 1255-1310.
14. Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, 71, 447—466.
15. Sastry, D. V. S., & Nag, A. K. (1990). Transfer of Resources from Centre and Growth in State Domestic Product. *Economic and Political Weekly*, 738-742.
16. Tallis, G. M. (1963). Elliptical and radial truncation in normal populations. *The Annals of Mathematical Statistics*, 34(3), 940-944.
17. Tibshirani R. (1987). Estimating optimal transformations for regression, *Journal of the American Statistical Association*, Vol. 83, 394–405.
18. Torrini, R. (2016). Labour, profit and housing rent shares in Italian GDP: long-run trends and recent patterns. *Bank of Italy Occasional Paper*, (318).

Analysis of the financial performance in Italian football championship clubs *via* longitudinal count data

Analisi della performance finanziaria delle squadre di calcio Italiane attraverso longitudinal count data

Anna Crisci and Luigi D'Ambra

Abstract Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. The main goal of any Football Championship club is to achieve sport results. The study of the relationship between sport and economic results attracts the interest of many scholars belonging to different disciplines. Very informative is considered the connection, over short or long periods of time, between the points in the championship and the resource allocation strategies. The aim of this paper is to give an interpretation of this last link using the longitudinal count data.

Abstract *Il calcio è senza dubbio lo sport più potente e popolare in Italia, che collega le comunità e stimola le emozioni. L'obiettivo principale di qualsiasi società di campionato di calcio è quello di ottenere risultati sportivi. Lo studio della relazione tra sport e risultati economici attira l'interesse di molti studiosi appartenenti a diverse discipline. Molto informativo è considerato il collegamento, tra i punti del campionato e le strategie di allocazione delle risorse. Lo scopo di questo articolo è di fornire un'interpretazione di quest'ultimo collegamento usando i dati di conteggio longitudinale.*

Key words: Italian Football championship clubs, Longitudinal count data, random and fixed effects

Anna Crisci

Department of Economics, Management, Institutions, University of Naples Federico II, Naples, Italy, anna.crisci@unina.it

Luigi D'Ambra

Department of Economics, Management, Institutions, University of Naples Federico II, Naples, Italy, dambra@unina.it

1. Introduction

Football is undoubtedly the most powerful and most popular sport in Italy, linking communities and stirring emotions. Professional business operators consider football an important industry with enormous potential in terms of growth and also for the indirect benefits gained by investors and management due to the popularity of football teams. In the football world, major consulting companies provide statistical data relating exclusively to athletic performance and sports results. The recipients of such data can be placed in two main categories. The first concerns professional football players, sports clubs, coaches, sports directors, etc. Such information is sold, in some cases, for payment. The second category is represented by media outlets, which release statistical reports to fans and sports people. The main goal of any Football Championship club is to achieve sport results. Nevertheless, football has also become one of the most profitable industries, with a significant economic impact in infrastructure development, sponsorships, TV rights and transfers of players. Very informative is considered the connection between the points in the championship and the resource allocation strategies. The aim of this paper is to give an interpretation of the link between the points in the championship and the resource allocation strategies using the longitudinal count data. In addition to the introduction, this paper consists of two further sections. In Sect. 2, the overview panel data approach is described. In Sect. 3 the Hausman test is shown. Finally, in Sect.4, empirical results for the Panel Poisson Model are shown.

2. Overview panel data

We often have data where variables have been measured for the same subjects (or countries, or companies, or whatever) at multiple points in time. These are typically referred to as Panel Data or as Cross-Sectional Time Series Data. With panel data you can include variables at different levels of analysis (i.e. students, schools, districts, states) suitable for multilevel or hierarchical modeling. Why do we use panel data? (Hsiao, 1985).

Benefits:

- They allow to identify the effects that are not identified in the cross-section data (Ben-Porath,1973).
- The panel allows to study the dynamics: while the cross-section allows you to estimate what proportion of the population is unemployed in a unit of time, the panel data show how this share varies over time;
- The panel data contain more information, more variability and therefore less collinearity among the variables and produce estimates more efficient, more precise parameters.
- They allow to control the effect of individual heterogeneity: i.e variables constant over time (individual heterogeneity) not observed (for which no data are available) (Baltagi and Levin, 1992).

Limits:
- Difficulty in the sample design and data collection.
- Distortion of the measurement errors.
- Problem of selection, no answers nor dissensions
- Limited dimension of time series.

2.1 Fixed and random effects

The fixed effects (FE) explore the relationship between predictor and outcome variables within an entity (persons, teams, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables. Each entity is different, therefore the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others (Stock and Watson, 2012).

The fixed effect model is:

$$\boldsymbol{y}_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \alpha_i + \boldsymbol{\varepsilon}_{it} \quad (1)$$

where

α_i (i=1…n) is the unknown intercept for each entity (n entity-specific intercepts).
$\boldsymbol{y}_{it}$ is the vector of dependent variables where $i = \text{entity}$ and $t = \text{time}$.
$\boldsymbol{\beta}$ is the vector of parameters to be estimate.
$\boldsymbol{x}_{it}$ represent the vector of covariates.
$\boldsymbol{\varepsilon}_{it}$ is the vector of error terms.

In the random effects the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model. The crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are "stochastic or not". The random effect model is:

$$\boldsymbol{y}_{it} = \boldsymbol{\beta}'\boldsymbol{x}_{it} + \boldsymbol{\nu}_{it} \quad (2)$$

where $\boldsymbol{v}_{it} = \alpha_i + \boldsymbol{\varepsilon}_{it}$ is the error of the random effect model.

2.2 Fixed and random effects Poisson Model

The Poisson fixed effects model has been proposed by Palmgren (1981) and Hausman et *al. (*1984). The standard way of evaluating the parameters of this model is the conditional maximum likelihood of Anderson (1970). The idea of the conditional method is to obtain an estimator FE without having to estimate each individual term α_i. We define λ_{it} as a linear combination of the observable characteristics:

$$\lambda_{it} = \boldsymbol{exp}(\boldsymbol{\beta}' \boldsymbol{x}_{it})$$

In the fixed effects model α_i is treated as a parameter to be estimated for each individual.

The RE estimator for count data assumes that random effects are gamma- distributed. This is due to the fact that we assume that random effects are multiplicative. Similar to linear models, it is assumed that, random effects are not correlated with the explanatory variables. Alternatively, the pooled Poisson model can be considered. However, if there is there is a systematic (constant over time) unobserved cross-sectional heterogeneity then the error term will be equicorrelated (as in the linear RE estimator). Estimation of parameters is performed using maximum likelihood estimators.

The GEE method of Liang & Zeger (1986) can be a solution to account for the dependence between each observation of the same insured as shown in Denuit, Pitrebois & Walhin (2003)

3. Hausman test

The generally accepted way of choosing between fixed and random effects is running a Hausman H-test (Hausman, 1978). Statistically, fixed effects are always a reasonable thing to do with panel data (they always give consistent results) but they may not be the most efficient model to run. Random effects will give you better p.values as they are a more efficient estimator, so you should run random effects if it is statistically justifiable to do so. Under the null hypothesis, the random effects is correctly specified, so both the fixed and random effects model are consistent, while under the alternative hypothesis, the random effects are correlated with the regressors, so the random effects model loses its consistency. Thus, the Hausman test (table1):

Table 1. Hausman test

	Random Effects	Fixed Effects
H_0: Random effect is not correlated with the regressors	Consistent and Efficient	Consistent and Inefficient
H_1: Random effect is correlated with the regressors	Inconsistent	Consistent

4. Empirical results for the Panel Poisson Model

The data used for our case study was obtained from the financial statements filed by the Serie A football teams. The period of study concerned the championship from season 2010/2011 up to 2014/2015.

The focus of the analysis is to verify the impact that some financial indicators have on the points achieved by football teams. We consider the following independent variables:

- **Depreciation Expense of multi-annual player contracts (DEM):** The depreciation expense of multi-annual player contracts are carried out a cost with an amortization plan. In the particular case of football club financial statements, the purchase of a football player is considered to be an immaterial immobilization, since it is the company's "right" to be the exclusive recipient of an athlete's sporting performance for a certain number of years. This investment is a cost shared for a period of time equal to the duration of the contract that the company has signed with the player. In our study we consider the log of depreciation Expense of multi-annual players.
- **Revenue net of player capital gain (RNC):** Like all companies, football clubs have different categories of income: characteristic and accessories: a) Typical revenues: revenue from the stadium, television rights, sponsorships, football rights, participation rights in European competitions; b) Revenue accessories: capital gains. The log of RNC considers both Typical Revenues and Revenue accessories.
- **Net Equity (NE):** Net equity is the difference between assets and liabilities and all the resources that the company has as a form of internal financing. Equity may be affected positively by contributions from shareholders (capital increases, retained earnings, etc.) and the profits generated by the company. Operating losses, the repayment of capital to members, results in a decrease in share- holders' equity. We consider the log of Net Equity.

In addition, we have considered, on the bases a bivariate descriptive analysis, also the square effect of DEM (DEM^2), given the non- linear relationship between Point and DEM. Finally, the interaction between DEM and NE (DEM*NE) also was considered.

In order to explore the panel data, figure 1, shows Points versus Year from 2010 to 2015; a line connects the five observations within each team. These lines represent a change over time.

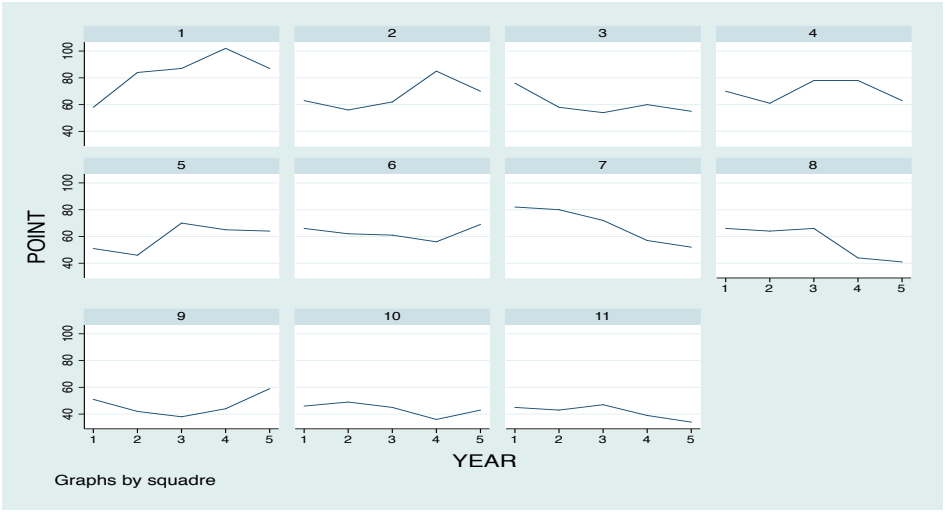


Figure 1: Plot Points versus Year from 2010 to 2015

The fixed effects (FE) Poisson model, in table 2, shows a significant overall model (p.value = 0.0397), with only one statistically significant variable: the RNC.

Table 2. Estimation of Panel Poisson Model with Fixed effects

Point	Coef.	Std.Err.	z	p. value
DEM	-0.426	2.864	-0.15	0.882
NE	-0.737	0.801	-0.92	0.358
RNC	0.288	0.104	2.75	0.006
DEM^2	-0.011	0.099	-0.11	0.914
DEM*NE	0.044	0.046	0.96	0.336

The output of the random effects (RE) Poisson model is shown in table 3:

Table 3. Estimation of Panel Poisson Model with Random effects

Point	Coef.	Std.Err.	z	p. value
DEM	3.211	1.546	2.08	0.038
NE	-0.776	0.383	-2.02	0.043
RNC	0.316	0.060	5.25	0.000
DEM^2	-0.120	0.050	-2.3	0.017
DEM*NE	0.048	0.022	2.14	0.033
Cons.	-22.261	12.861	-1.73	0.083
/ln alpha	-7.3223	2.5236		
alpha	0.0006	0.0016		
Likelihood-ratio test of alpha = 0 chibar2(01)=10.61 p.value = 0.001				

In the random effects model we have all variables statistically significant. The output also includes a likelihood-ratio test of $\alpha = 0$, which compares the panel estimator with the pooled (Poisson) estimator. We find that the random-effects model is significantly different from the pooled model.
Although the random effects model include time-invariant variables and is more efficient when compared to the fixed effects model, it is consistent only when it is correctly specified. So, it is necessary to test which one is better for our data. Under the null hypothesis, the random effects is correctly specified, so both the fixed and random effects model are consistent, while under the alternative hypothesis, the random effects are correlated with the regressors, so the random effects model loses its consistency. From the result,

Table 4. Results Hausman test

	(b) fixed	(B) random	(b-B) Difference	S.E.
DEM	-0.426	3.211	-3.637	2.410
NE	-0.737	-0.776	0.0389	0.703
RNC	0.288	0.316	-0.028	0.085
DEM^2	-0.107	-0.120	0.109	0.085
DEM*NE	0.044	0.048	-0.004	0.039

Chi2(5) = 6.22, p. value = 0.2851

So now we can not to reject the null hypothesis of uncorrelation between regressors and random effects (p.value is well above the critical value of 0.05). This result means that the individual term in the random effects model are not correlated with the regressors.

Discussion

In the present work we have analyzed the link between the championship points and the resource allocation strategies through a longitudinal count data. In particular, we analyze the impact that some economic-financial variables have on the points made by football teams participating in the Series A championship (2010-2015), by using Panel Poisson Model. We have compared the fixed effects with random effects poisson model. The choice between fixed and random effects was performed running Hausman H-test. In particular, the choice of dealing with individual effects as fixed or random enough delicate and some final considerations should be made. The validity of the RE Poisson depends on very strong distributional assumptions. So, we would just stick to the FE regression. In particular, the choice of dealing with individual effects as fixed or random enough delicate. The fixed effects should be used to estimate the specific effects of the sample (i.e, an exhaustive sample countries, a sample of companies in a particular industry in which the selected sample is representative of the characteristics of the industry). By contrast, the random effects should be used for random samples and to make

inference on the population. Then, in our case the choice could be cast on the fixed effects model, as our entity can not really be thought of as random draws from a population. In fact, the inferences that we have drawn are conditioned to the individuals included in the sample as opposed to a random model where the individual characteristics become a component of the population and the inferences are then related to the same population (Crisci *et.al*, 2014)

References

Anderson, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society B **32**, 283-301.

Baltagi, B. H. & Levin D. (1992). Cigarette Taxation: Raising Revenue and Reducing Consumption. *Structural change and Economics Dynamic* **3** (2) 321-335.

Ben-Porath, Y. (1973). Labor-Force Participation Rates and the Supply of Labor. *Journal of Political Economy*, **81**(3), pp. 697-704.

Crisci,A., D'Ambra,A.& Paletta, A.(2014). A Panel data approach to evaluate the Passenger Satisfaction of a Public Transport Service. *Procedia Economics and Finance*, Elsevier, **17** (2014) 231 – 237.

Denuit, M., Pitrebois, S. & Walhin, J.F. (2003). Tarification automobile sur donnees de panel. Bulletin of the Swiss Association of Actuaries, 51-81.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica*, **46**(6), pp. 1251-1271.

Hausman, J.A., Hall, B.H., & Griliches, Z. (1984). Econometric models for count data with application to the Patents-R&D Relationship. Econometrica 52, 909-938.

Hsiao, C. (1985). Benefits and Limitations of Panel Data. *Econometric Reviews*, **4**(1) pp.121- 174.

Liang, K.Y., & Zeger, S.L.. Longitutinal data analysis using generalized linear models. Biometrika **73**, 13-22.

Palmgren, J. (1981). The Fisher information matrix for log-linear models arguing conditionnally in observed explanatory variables. Biometrika **68**, 563-566**.**

Stock, H.J & Watson, M.W. (2012). *Introduction to Econometrics*, 3rd ed., Pearson Addison Wesley. Chapter 10.

A Poset perspective for the evaluation of self-reported health of the elderly in Italy

Un metodo basato sul POSET per valutare lo stato di salute auto-percepita della popolazione anziana in Italia

E. Furfaro, L. Pagani and M. C. Zanarotti

Abstract Measuring health status is becoming a more and more relevant task, especially in relation to ageing societies. In this contribution, we first propose the use of a methodology based on the theory of Partially Ordered Sets that allows to build synthetic indicators out of a set of ordinal variables, respecting the ordinal nature of the variables included. Secondly, using survey data, we calculate two synthetic indicators to evaluate self-rated health status of the elderly population living in Italy.
Abstract *Misurare lo stato di salute è un compito molto importante soprattutto nelle società contemporanee caratterizzate da un aumento della popolazione in età anziana. In questo contributo, dapprima proponiamo l'uso di un approccio basato sulla teoria degli insiemi parzialmente ordinabili che permette di costruire indicatori sintetici a partire da variabili categoriali conservandone il naturale ordinamento. Successivamente, gli indicatori proposti sono stati utilizzati per valutare lo stato di salute della popolazione anziana residente in Italia.*

Key words: self-reported health, ordinal data, regression trees, POSET

1 Introduction

The adage 'health is wealth' is a timeless truth that becomes even more relevant in ageing societies: health is wealth for both individual well-being and for population prosperity, particularly in the face of the new challenges connected with population ageing. The state of health impacts all dimensions of individuals' life, and poor

E. Furfaro
Università Cattolica del Sacro Cuore, Milano, e-mail: emanuela.furfaro@unicatt.it

L. Pagani
Università di Udine e-mail: laura.pagani@uniud.it

M.C. Zanarotti
Università Cattolica del Sacro Cuore, Milano e-mail: chiara.zanarotti@unicatt.it

health conditions dramatically influence it as a whole. Individual health in turn impacts societies, for example by increasing the need for care and assistance. Getting older frequently means getting worse in health conditions so being able to measure and to monitor individual health is of crucial importance in ageing societies.

Measuring health implies a clear definition which may not be trivial. In 1948 the World Health Organization defined health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" ([16]). It was a new and ambitious formulation because it overcame the definition of health as merely absence of disease. More recently this definition has been criticized and some new proposals have been made with the aim to define health taking into account new goals and new needs ([10]). What is clear when looking at the different definitions is that health is a complex phenomenon that involving different factors that affect health of individuals and/or communities. This has led to the deveolpment of many methods to measure health, different approaches, different purposes and levels of measurement ([12, 13]).

Health can be evaluated at individual level through one or more indicators or at aggregate level considering population indicators (life expectancy, mortality rate, incidence of some pathology, etc.). In the former case, measurements can be classified as "subjective", based on self-perceived health; or "objective", based on diagnosis by physicians and/or by other procedures like laboratory or screening examinations. In this contribution we focus on individuals' self-reported health measures, as they have been proven to be efficient tools for health status evluation ([11, 9]). Even in this case there are more possibilities. Here we consider the measurement of subjective health through a set of individual indicators. Specifically we refer to the SF-12 Questionnaire of Health Survey developed during the 90s in United States by the Medical Outcomes Trust of Boston ([15]). The SF-12 is a psychometric questionnaire based on twelve items widely used in internationla studies in the last decades. Items in the questionnaire are then summarized in two synthetic indices (or compiste indicators), one representing physical health (PCS) and the other one mental health (MCS). Please refer to Table 1 for details on the twelve items together with the labels used in this contribution, and the reduced eight-dimensional scale -as presented by [15]- useful for variables interpretation. PCS and MCS are obtained by aggregating the items into two composite indicators by means of a weighting system.

In this contribution we propose the synthesis of the two health indicators with the use of an alternative approach based on the Partially Ordered Set theory (POSET). This approach allows to build synthetic indicators out of a set of categorical variables without the need of any aggregative procedure and respecting the ordinal nature of the variables in the SF-12. After giving some basic background of the POSET theory and providing the details of the proposed indicators, we calculate them for evulating the health conditions of the ageing population living in Italy.

The rest of the contribution is organised as follows: Section 2 is devoted to briefly present the POSET approach; Section 3 summarises the main results on the ageing population who lives in Italy; while concluding remarks are contained in Section 4.

Table 1: The SF-12 measurement model ([15]), labels used in this contribution

Summary measures	Item	Label	Scales
Physical Health	Perceived health	X_1	General Health
	Limited activities	X_2	Physical Functioning
	Difficulties in climbing several fights of stairs	X_3	Physical Functioning
	Accomplished less due to phyisical condition	X_4	Role-Physical
	Limited work due to physical condition	X_5	Role-Physical
	Pain interferes with everyday activities	X_6	Bodily Pain
Mental Health	Accomplished less because of emotional status	X_7	Role Emotional
	Less concentrated because of emotional status	X_8	Role Emotional
	Felt calm	X_9	Mental health
	Felt full of energy	X_{10}	Vitality
	Felt sad	X_{11}	Mental Health
	Emotional status compromised social life	X_{12}	Social Functioning

2 The POSET approach for building composite indicators

The theory of Partially Ordered Sets is a well established mathematical theory that has recently been leveraged to calculate synthetic measures out of a set of ordinal variables. Its use in the calculation of synthetic measures is motivated by the fact that, differently from aggregative procedures, it preserves the ordinal nature of the variables ([5]). It has been succesfully used for producing synthetic measures of wealth, life satisfaction, gender gap and for the evaluation of frailty in the elderly population ([3, 4, 2, 14]). In this section we give some basic definitions useful for the purpose of understanding our work (for more details see, among others, [6]).

A *Partially Ordered Set* (POSET) is a finite set X with a partial order relation, i.e. a binary relation "$\leq$" satisfying the properties of *(i)* reflexivity, *(ii)* antisymmetry and *(iii)* transitivity (for more details see for instance [6]). Two elements a and b of the set X are comparable if $a \leq b$ or $b \leq a$, otherwise we say they are *incomparable*. The elements of X can hence be ordered based on the partial order relation, generating *linear extensions* (see [8] for formalisation). When two elements are incomparable, they generate more linear extensions as there is more than one way to order them.

For the purpose of our study we consider S ordinal variables, each with k_s possible responses (with $s = 1,...,S$). The elements of the set X are combinations of the values of the S ordinal variables and they are called profiles. Two profiles $\mathbf{p}_a = \{p_a^1, p_a^2, ..., p_a^S\}$ and $\mathbf{p}_b = \{p_b^1, p_b^2, ..., p_b^S\}$ are comparable if and only if $p_a^s \leq p_b^s \; \forall \; s, \; s = 1,...,S$, or viceversa. In other words, $\mathbf{p}_a$ and $\mathbf{p}_b$ are comparable if and only if the values observed on $\mathbf{p}_a$ are higher or equal than those observed on $\mathbf{p}_b$ for all variables, or viceversa. Note that in our case, the S variables are those given in Table 1 that we will call elementary variables. The profiles can then be ordered, as mentioned above, generating linear extensions. A possibility for evaluating the

profiles is to set a threshold profile τ so that profiles can be classified above τ or below τ, hence creating two groups. On the different linear extensions, profiles may always be classified in the same group or they may be classified differently on the different linear extensions, representing fuzzy states. Drawing on the threshold definition and on the computation of linear extensions, different synthetics measures have been proposed for the evaluation of profiles.

In this contribution, we evaluate the *Height* of a profile that is a combination of two other measures, namely *wealth* and *severity* ([3]). The severity function provides a measure of the depth of a profile into a group. In order to calculate severity, the first step is to compute, for every linear extension, the severity function, i.e. the distance between a profile and the first element lower than τ. The distance is computed on the rank of the two objects. Let $\Omega(P)$ be the set of all linear extensions on a POSET P, the severity function is given by:

$$svr_l(\mathbf{p}) = \begin{cases} r_l(\mathbf{q}_l) - r_l(\mathbf{p}), & \text{if } \mathbf{p} \leq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $\mathbf{q}_l$ is the first element lower than τ, $r_l(\mathbf{q}_)$ and $r_l(\mathbf{p})$ are the ranks of profile $\mathbf{p}$ and of profile $\mathbf{q}_l$ respectively on a linear extension $l \in \Omega(P)$. The severity value of a profile is then obtained by aggregating the results observed on all linear extensions:

$$svr_l(\mathbf{p}) = \frac{1}{|\Omega(P))|} \sum_{l \in \Omega} svr_l(\mathbf{p}) \tag{2}$$

With reference to health status, severity allows for an evaluation of the intensity of poor health. High values of severity indicate that not only the profile is often classified as a poor health profile, but, when classified as poor health, it positions very far from the threshold indicating the intensity of poor health. Similarly, the wealth function provides a measure of how good is the condition of those classified as good health profiles, indicating on average how far deep into the group of good health profiles a profile is positioned.

The *Height*, that is the profiles evaluation measure we use in this contribution, is then given by the following:

$$H_\tau(\mathbf{p}) = wea(\mathbf{p}) - svr(\mathbf{p}) = \frac{1}{|\Omega(P))|} \sum_{l \in \Omega} wea_l(\mathbf{p}) - \frac{1}{|\Omega(P))|} \sum_{l \in \Omega} svr_l(\mathbf{p}) \tag{3}$$

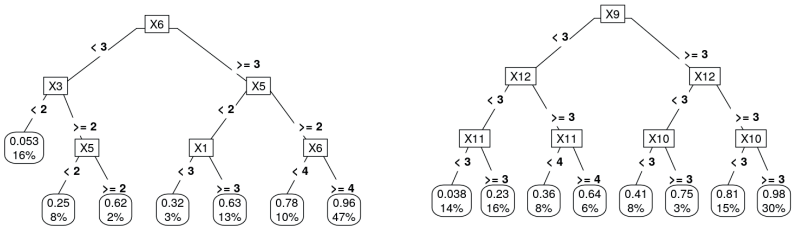
High values of H_τ correspond to profiles in good health and low values correspond those in poor health. Note that high values correspond to profiles that when classified as in good health are very far from the threshold (high values of wealth) and, if classified as poor health, are not too severe.

3 Application: evaluating health conditions of elderly population living in Italy

We use data from the 2013 Multipurpose Survey on Health Conditions carried out by the Italian National Institute of Statistics and we focus on people aged ≥ 65 years. The sample includes 49.811 households, for a total of 119.000 individuals, of which 27.003 are above 65 years old. Following the methodolgoy described in 2 and using the variables in the SF-12 (see Table 1), we built two indicators based on the Equation 3, one for physical health (PCS) and one for mental health (MCS). They were calculated using elementary variables $X_1 - X_6$ and $X_7 - X_{12}$ respectively and setting the threshold on the basis of external information. All the computations were carried out in the R environment, using the R package PARSEC for the computation of posetic measures ([7]).

In order to better understand which elementary variables characterise low and high values of PCS and MCS, we implemented a regression tree for each of the synthetic indicators. PCS and MCS are output variables, the elementary indicators are the regressors. Figure 1 syntheses our results, showing the percentage size of the groups obtained, along with the average value of PCS (left panel) and MCS (right panel), and the values of the elementary indicators at each splitting node. PCS and MCS were normalised to simplify interpretation. An interesting finding regard the role of X_{12} in discriminating between poor mental health profiles and good mental health profiles: in fact X_{12} represents a compromised social life that highlights the importance of social ties for healthy (mental) ageing ([1]).

Secondly, we use quantile regressions to further study the sub-groups in poorer health. Thanks to it, we investigate the role of structural and economic variables on different quantiles. In particular, we are interested in studying gender differences, territorial differences, which is a long-standing issue in Italy, and the role of social relationships which have attracted increasing interest in the context of active and health ageing ([1, 17]). We control for age, citizenship, education and type of income.



(a) Physical health indicator (b) Mental health indicator

Fig. 1: Groups as identified by the regression tree, for PCS (left panel) and MCS (right panel). Variables labels are given in Table 1

Figure 2 and 3 show main results, with the estimated values for the coefficients (confidence intervals in the shaded area) at each quantile regression. More results are available upon request. For both PCS and MCS, the variables considered exhibit smaller coefficients on the very extreme quantiles, while the coefficients seem larger around the 20-30th percentile for physical health (Figure 2), and around the 30-40th percentile for mental health (Figure 3). This suggest that very poor health below those quantiles may be related to some other factors, such as for example having chronic diseases, and it may be independent of structural and contextual characteristics. Women generally report lower levels of mental and physical health, with significant coefficients in all percentiles, intensiying for the mentioned middle-low percentiles. Similarly, North-South territorial differences widen in such percentiles suggesting that the geographical context may play a role in fuzzy states. Being widow is also related to lower levels of both physical and mental health, with coefficients being higher in correspondance of the middle-low quantiles. Regarding living arrangements, we compared those living in couple and those who live with other people (either family members or not) with those living alone. Results highlight no differences in any quantiles with the regard to the former and poorer health

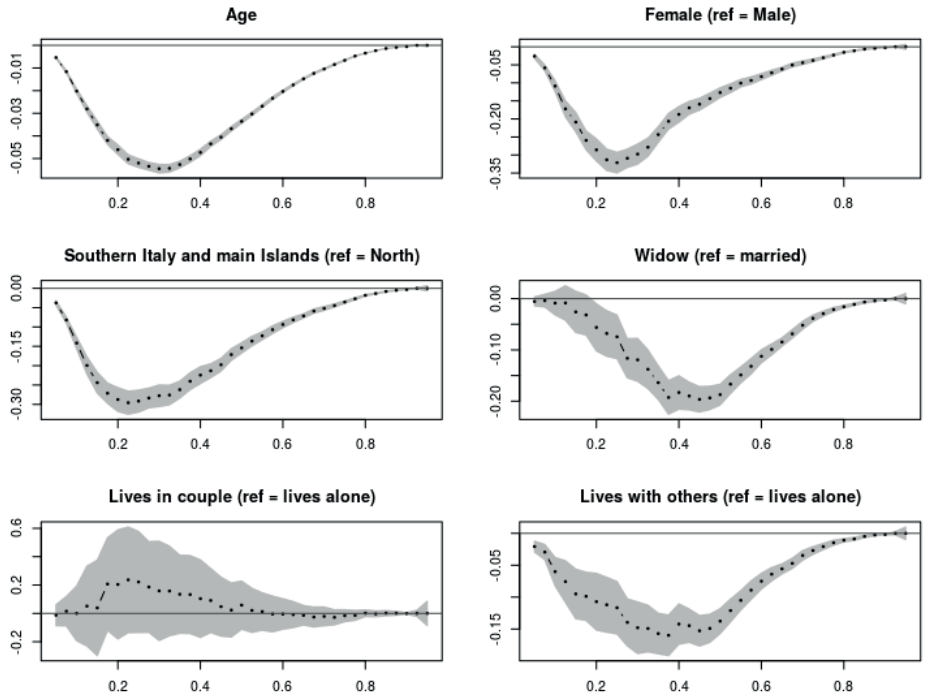


Fig. 2: Physical health indicator: estimated quantile regression coefficients at each quantile. Gray area represents the 95% confidence interval.

status for the latter. This may seem counterintuitive, but it can be interpreted as evidence of the need of people in poor health to stay with others, highlighting the need for social support.

4 Final remarks

This contribution addresses the important task of synthetically measuring health conditions. We propose the use of a methodology based on the POSET theory that allows to create synthetic indicators out of a set of ordinal variables. We build a posetic version of PCS and MCS out of the variables included in the SF-12 questionnaire, widely considered a valuable starting point for the analysis of health conditions. After defining the indicators and identifying which elementary ordinal variables discriminate between good and poor health profiles, we calculated the syntehtic indicators to provide a synthetic measure of the health status of the elderly population living in Italy.

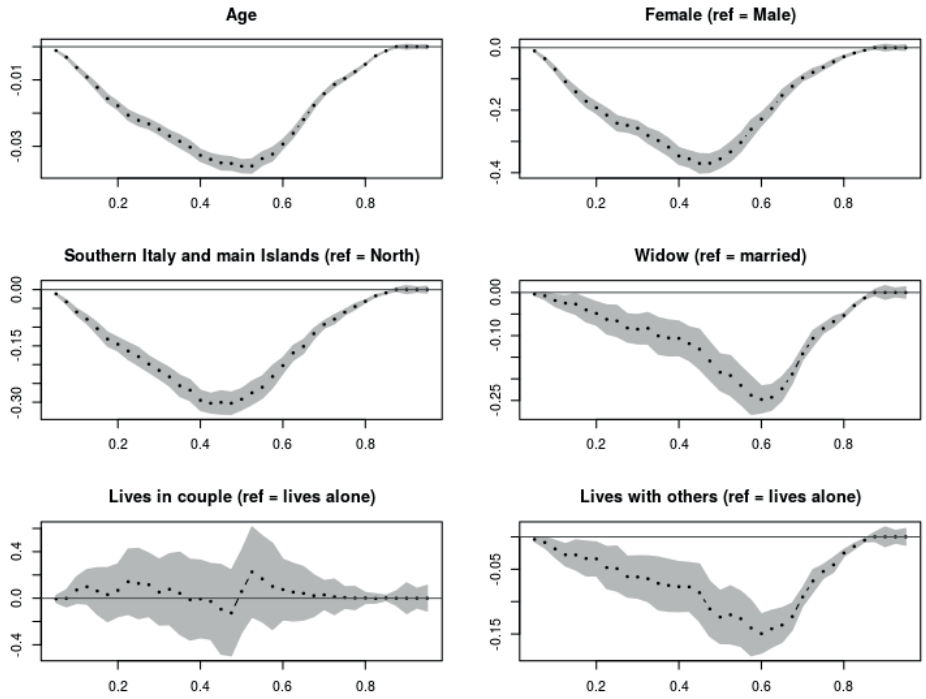


Fig. 3: Mental health indicator: estimated quantile regression coefficients at each quantile. Gray area represents the 95% confidence interval.

By means of quantile regressions, we found that the middle-low quantiles seem to be largely affected by gender and territorial differences, suggesting that it is in fuzzy situations that structural and contextual variables matter. We also looked at the social dimension, highligthing lower levels of health for widowed people and finding that people with lower health tend to be in larger households.

References

1. Adams, K. B., Leibbrandt, S., Moon, H.: A critical review of the literature on social and leisure activity and wellbeing in later life. Ageing & Society, **31(4)**, 683–712 (2011)
2. Arcagni, A., di Belgiojoso, E. B., Fattore, M., Rimoldi, S. M.: Multidimensional Analysis of Deprivation and Fragility Patterns of Migrants in Lombardy, Using Partially Ordered Sets and Self-Organizing Maps. Soc. Indic. Res., **136(3)**, 551–579 (2019)
3. Boccuzzo, G., Caperna, G.: Evaluation of Life Satisfaction in Italy: Proposal of a Synthetic Measure Based on Poset Theory. In: Maggino F. (ed.) Complexity in Society: From Indicators Construction to their Synthesis, pp. 291-321. Springer, Cham (2017)
4. Di Brisco, A. M., Farina, P.: Measuring Gender Gap from a Poset Perspective. Social Indicators Research, 136**(3)**, 1109-1124 (2018)
5. Fattore, M., Maggino, F., Greselin, F.. Socio-economic evaluation with ordinal variables: Integrating counting and poset approaches. Statistica & Applicazioni, 1, pp. 31–42 (2011)
6. Fattore, M., Maggino, F., Colombo, E.: From composite indicators to partial orders: evaluating socio-economic phenomena through ordinal data. In: Maggino, F., Nuvolati, G.(eds.) Quality of life in Italy, pp. 41-68. Springer, Dordrecht (2012)
7. Fattore, M., Arcagni, A.: PARSEC: an R package for poset-based evaluation of multidimensional poverty. In Multi-indicator systems and modelling in partial order, pp. 317-330. Springer, New York, NY (2014)
8. Fattore, M.: Synthesis of indicators: the non-aggregative approach. In: Maggino F. (ed.) Complexity in Society: From Indicators Construction to their Synthesis, pp. 193-212. Springer, Cham (2017)
9. Golini, N., Egidi, V.: The Latent Dimensions of Poor Self-Rated Health: How Chronic Diseases, Functional and Emotional Dimensions Interact Influencing Self-Rated Health in Italian Elderly. Soc. Indic. Res. **128(1)**, 321–339 (2016)
10. Huber, M., Knottnerus, J.A., Green, L. et al.: How should we define health?. BMJ. **343**: d4163 (2011)
11. Idler, E. L., Benyamini, Y.: Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. J. Health. Soc. Behav. **38(1)**, 21—37 (1997)
12. McDowell, I., Spasoff, R., Kristjansson, B.: On the classification of population health measurements. Am. J. Public. Health. **94**, 388—393 (2005)
13. Murray, C. J. L., Salomon, J. A., Mathers, C. D., Lopez, A. D. (eds.) Summary measures of population health : concepts, ethics, measurement and applications . World Health Organization, Geneva (2002), https://apps.who.int/iris/handle/10665/42439
14. Silan, M., Caperna, G., Boccuzzo, G.: Quantifying Frailty in Older People at an Italian Local Health Unit: A Proposal Based on Partially Ordered Sets. Social Indicators Research, pp. 1-26 (2019).
15. Ware, J. E., Kosinski, M., Keller, S.D.: SF-12: How ro Score the SF-12 Physical and Mental Health Summary Scales. Boston, MA: The Health Institute, New England Medical Center, First Edition (1995)
16. WHO: Preamble to the Constitution of WHO as adopted by the International Health Conference. New York, 19 June - 22 July 1946. Official Records of WHO, no. 2, p. 100.
17. Zaidi, A., Howse, K.: The policy discourse of active ageing: Some reflections. Journal of Population Ageing, **10(1)**, 1–10 (2017)

Reduced K-means Principal Component Multinomial Regression with external information to evaluate consumer's preferences

Reduced K-means Principal Component Regression con informazioni esterne per la valutazione delle preferenze dei consumatori

Antonio Lucadamo and Pietro Amenta

Abstract In last years some procedures for dealing with Multicollinearity in Multinomial Regression Model have been developed. The aim of this paper is to introduce a new strategy to improve the efficacy of the Reduced K-means Principal Component Multinomial Regression. The new methodology has been applied to study how the consumers' preferences about different characteristics of coffee can be influenced by some physical-chemical properties.

Abstract *Negli ultimi anni diverse procedure sono state sviluppate per affrontare il problema della Multicollinearità nei modelli di regressione multinomiale. L'obiettivo di questo lavoro è quello di introdurre una strategia per migliorare l'efficacia del Reduced K-means Principal Component Multinomial Regression. La nuova metodologia è stata applicata per studiare come le preferenze dei consumatori relativamente ad alcune proprietà del caffè possono essere influenzate dalle caratteristiche fisico-chimiche.*

Key words: Principal Component Multinomial Regression, Multicollinearity, Reduced K-means.

1 Introduction

Principal Component Multinomial Regression (PCMR) [2, 4] and Reduced K-means Principal Component Multinomial Regression (RKPCMR) [1] are useful methods to face with multicollinearity problem in Multinomial Regression Model. Both the methods are based on the building of new variables, the components, lin-

Lucadamo Antonio
DEMM - University of Sannio, e-mail: antonio.lucadamo@unisannio.it

Amenta Pietro
DEMM - University of Sannio, e-mail: amenta@unisannio.it

ear combinations of the original regressors. They are orthogonals and so they can be used as new explicative variables to overcome the problem of high correlation among regressors. In this paper we introduce a new version of RKPCMR, proposing the introduction of external information in the building of the components and showing how the new method works better for prediction purposes.

2 Reduced K-Means Principal Component Multinomial Regression

Reduced K-means [3] considers the unweighted least squares estimation of the model:

$$\mathbf{X} = \mathbf{UMA}^T + \mathbf{E}$$

where:

- $\mathbf{X}$ is the $(n \times p)$ data matrix with n observations and p continuous variables;
- $\mathbf{U}$ is the membership matrix: $u_{nk} = 1$ if the i^{th} observation belongs to group k, 0 otherwise;
- $\mathbf{M}$ is the $(k \times q)$ centroid matrix in a reduced space, with m_{kq} centroid value of the q^{th} component obtained on the k^{th} cluster;
- $\mathbf{A}$ is $(p \times q)$ component weights matrix for variables;
- $\mathbf{E}$ is a $(n \times p)$ matrix of error terms.

The model is equivalent to the maximization of the following objective function:

$$G_{RKM}(\mathbf{A}, \mathbf{U}) = ||\mathbf{H_U XA}||^2 = max$$

with the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}_q$. $\mathbf{H_U}$ is the $(n \times n)$ projection matrix on the space spanned by the columns of $\mathbf{U}$: $\mathbf{H_U} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$. The components can be required uncorrelated considering the following constraint $\mathbf{A'X'XA} = \mathbf{D}$ with $\mathbf{D}$ diagonal.

Following the approach used in PCMR, the RKPCMR considers the scores obtained on the reduced space as new values of explicative variables in a Multinomial Logit Model. The prediction of a categorical response variable can be improved if we considers these new components, built keeping in consideration a possible clustering structure. The probability to choose the alternative k among the s can be expressed in terms of PCs as:

$$P_k = \frac{e^{\gamma_{0k} + Z\gamma_k}}{\sum_{j=1:s} e^{\gamma_{0j} + Z\gamma_j}}$$

where γ_{0k} and γ_k are the intercept and the vector of estimate coefficients and $\mathbf{Z} = \mathbf{XA}$ are the new scores with $\mathbf{A}$ representing the loadings matrix.

3 Principal Component Analysis with external information

When there are additional information about subjects and variables, they can be used to aid subjective interpretations of Principal Component Analysis using the Takane and Shibayama's approach [6]. "External Analysis" of the Principal Component Analysis with external information on both subjects and variables [6] considers the additive data decomposition of a single quantitative matrix $\mathbf{X}$ according to the row and column external information matrices $\mathbf{H}$ and $\mathbf{Z}$, respectively

$$\begin{aligned}\mathbf{X} &= (\mathbf{P_H} + \mathbf{P_H^\perp})\mathbf{X}(\mathbf{P_Z} + \mathbf{P_Z^\perp}) \\ &= \mathbf{P_H X P_Z} + \mathbf{P_H^\perp X P_Z} + \mathbf{P_H X P_Z^\perp} + \mathbf{P_H^\perp X P_Z^\perp}. \end{aligned} \quad (1)$$

where $\mathbf{P_H} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ and $\mathbf{P_H^\perp} = \mathbf{I} - \mathbf{P_H}$ are the idempotent orthogonal projection operators onto $Im(\mathbf{H})$ (the subspace of $\Re^n$ spanned by the column vectors of $\mathbf{H}$) and onto the ortho-complement subspace $Im(\mathbf{H})^\perp$, respectively, such that $\Re^n = Im(\mathbf{H}) \oplus Im(\mathbf{H})^\perp$, $\mathbf{P_H} + \mathbf{P_H^\perp} = \mathbf{I}$ and $\mathbf{P_H P_H^\perp} = \mathbf{0}$ [5] with $\mathbf{0}$ zero matrix of order $n \times n$. Analogously, $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ is the idempotent orthogonal projection operator onto the subspace of $\Re^p$ spanned by $\mathbf{Z}$: $Im(\mathbf{Z})$. Every pair of terms of the right hand of the decomposition (1) are columnwise and rowwise trace-orthogonal, where two matrices $\mathbf{A}$ and $\mathbf{B}$ of same order are said to be trace-orthogonal when the sum of elements on the main diagonal (trace) of each matrix $(\mathbf{A}^T\mathbf{B})$ and $(\mathbf{A}\mathbf{B}^T)$ is always zero. For example, the second and the third term in (1) respect the condition of trace-orthogonality because $tr(\mathbf{P_Z X}^T\mathbf{P_H^\perp P_H X P_Z^\perp}) = tr(\mathbf{P_H^\perp X P_Z P_Z^\perp X}^T\mathbf{P_H}) = 0$. The property of trace-orthogonality of decomposition (1) implies then the sum of squares of elements in $\mathbf{X}$ is decomposed into the sum of sums of squares corresponding to the four terms in the right hand of (1). PCA of matrix $\mathbf{X}$ extracts the most important dimensions of the components to be analyzed. Additional information can therefore be obtained by applying PCA to each term (or combined) of the right hand of (1) separately ("Internal Analysis"). Each term has a precise statistical meaning in it in explaining the total inertia of $\mathbf{X}$:

- $\mathbf{P_H X P_Z}$ reflects the effect of row and column information;
- $\mathbf{P_H^\perp X P_Z}$ highlights only the effect of column information because the row ones have been deleted;
- $\mathbf{P_H X P_Z^\perp}$ points out only the effect of row information because the colum ones have been removed;
- $\mathbf{P_H^\perp X P_Z^\perp}$ reflects the part of the total inertia of $\mathbf{X}$ where the effects of the external information have been completely disregarded.

For our purpose we consider the case in which the information are only on the subjects. Matrix $\mathbf{X}$ can be then decomposed as follows:

$$\mathbf{X} = (\mathbf{P_H} + \mathbf{P_H^\perp})\mathbf{X} = \mathbf{P_H X} + \mathbf{P_H^\perp X}. \quad (2)$$

The first term on the right hand of the decomposition represents then the row constraint effect, the second one pertains to what can not be explained by row information.

4 RKPCMR with external information to evaluate consumer's preferences

In this paper we want to evaluate if the preferences of some consumers of coffee can be influenced by physical-chemical properties. A group of judges was asked to indicate which characteristic (quality of aroma, intensity of aroma, bitterness, quality of flavour) of some coffees they preferred. The coffee can be classified in 3 groups according to their typologies: waffle coffee, capsula coffee and ground coffee. The variables (and the label) that can affect the choice are synthesized in table 1.

Table 1 Explicative variables

Variable	Label
Caffeine	CAF
Viscosity	VIS
Conductivity	CDT
Ph	PHH
Acidity	CID
Extraction tax	TEE
Extraction per sec	EXS
Optic density at 430 n.m	DO4
Optic density at 510 n.m.	DO5
Capacity of retention	CPR

To reach our aim, the first idea is to apply a classic multinomial logit model, but, if we consider the correlation matrix among the explicative variables, it is evident that there are many high correlations among them. This can lead to some problems in estimation procedure due to possible multicollinearity problem.

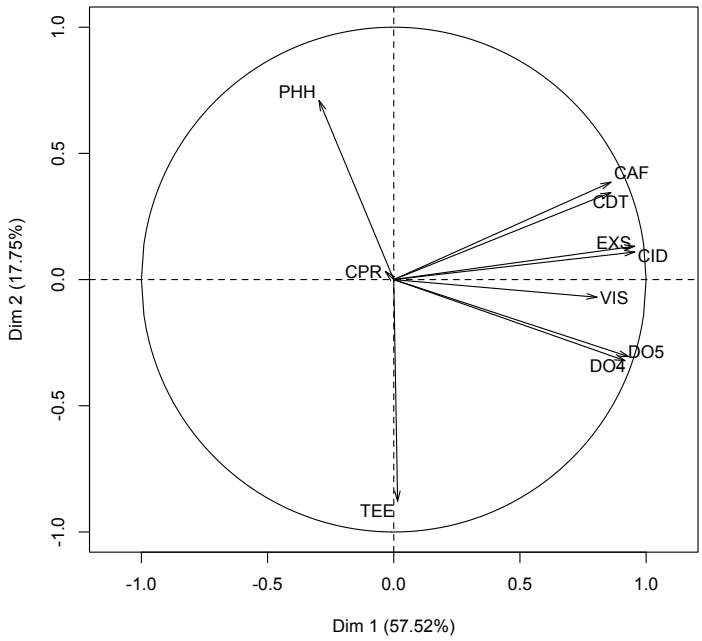
The situation is evident if we look at the plot obtained considering the first two axes of a Principal Component Analysis. It is easy to notice that, if we look at the first axis, the most part of the variables is on the right side. The factor indeed explain more than the 50% of the variability in the data.

In order to verify the multicollinearity we consider the Condition Index (CI), given by the square root of the ratio between the maximum and the minimum eigenvalue of the correlation matrix.

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Table 2 Correlation matrix

	EXS	TEE	PHH	CID	DO4	DO5	CDT	CAF	VIS	CPR
EXS	1.00	0.02	0.03	0.97	0.89	0.86	0.90	0.86	0.68	-0.04
TEE	0.02	1.00	-0.49	-0.02	0.29	0.28	-0.18	-0.28	-0.28	-0.02
PHH	0.03	-0.49	1.00	0.03	-0.21	-0.21	0.09	0.18	-0.07	-0.09
CID	0.97	-0.02	0.03	1.00	0.89	0.86	0.92	0.88	0.65	-0.11
DO4	0.89	0.29	-0.21	0.89	1.00	0.99	0.79	0.70	0.68	-0.10
DO5	0.86	0.28	-0.21	0.86	0.99	1.00	0.76	0.66	0.69	-0.13
CDT	0.90	-0.18	0.09	0.92	0.79	0.76	1.00	0.89	0.60	0.04
CAF	0.86	-0.28	0.18	0.88	0.70	0.66	0.89	1.00	0.64	-0.16
VIS	0.68	-0.28	-0.07	0.65	0.68	0.69	0.60	0.64	1.00	0.08
CPR	-0.04	-0.02	-0.09	-0.11	-0.10	-0.13	0.04	-0.16	0.08	1.00



When the value of this index is higher than 30, it indicates the presence of multicollinearity in the data. In our analysis it is equal to 103.525, so the multicollinearity can be a serious problem and so it affects the parameter estimation.

It has been shown in previous papers [1, 2, 4] that possible solutions can be the PCMR and RKPCMR. In this paper the idea, according to the Takane and Shibayama's approach [6], is to perform PCMR and RKPCMR using as explicative variables the scores obtained by PCA on the $\mathbf{X}$ matrix as well as on the two terms of the decomposition (2).

We introduce as external information the typologies of coffee as previously described. To evaluate how the different methods work, we split the dataset in two sub-samples: 70% of observations for the estimation sample and 30% for the validation one. We estimate the parameters on the first sample and then we use the second one to calculate the rate of well classified.

For the classical multinomial logit model the rate of correct classification on the estimation sample and on the test sample is 82.34% and 42.31%, respectively, whereas the percentages for the other models are in table 3 and 4.

Table 3 Correct classification rate (CCR) for different methods on estimation sample

Method	CCR	Method	CCR
PCMR	71.43%	RKPCMR	74.37%
PCMR-EI	56.28%	RKPCMR-EI	53.74%
PCMR-removing EI	77.54%	RKPCMR removing EI	78.53%

Table 4 Correct classification rate (CCR) for different methods on validation sample

Method	CCR	Method	CCR
PCMR	53.85%	RKPCMR	53.85%
PCMR-EI	46.15%	RKPCMR-EI	42.31%
PCMR-removing EI	61.54%	RKPCMR removing EI	69.23%

It is easy to notice that, for the estimation sample, the multinomial logit has an high rate of well classified, but this is due probably to an overfitting to the data. On this sample, the percentages for the other methods are lower, even if, except for PCMR-EI and RKPCMR-EI, the correct classification rate is higher than 70%. The use of Principal Components as new variables, leads instead to better results on the validation sample. Indeed, the rate increases from 42% for the classical Multinomial Logit Model, to about 54% both for PCMR and for RKPCMR.

Furthermore, if we consider the analysis performed removing the external information, the percentage increases considerably, reaching the value of 61.54% for the PCMR and of 69.23% for RKPCMR. Additional deeper structural knowledge are then given by the use of external information on the statistical units.

5 Conclusions and perspectives

In this paper we show how using external information in the building of components can improve the performances of two known methods for dealing with multicollinearity in MNL models. The application shows that the proposed approach performs quite well on real data. However, a deeper analysis is necessary to evaluate the goodness of the method: for example a cross-validation analysis has to be

performed on the real data set. Furthermore an extensive simulation study, considering different scenario about the number of observations, predictors and the level of correlation among predictors must be evaluated.

References

1. Amenta, P., A. Lucadamo, and A. P. Leone (2017). Reduced k-means principal component multinomial regression for studying the relationships between spectrometry and soil texture. In *Cladag2017 Book of Short Papers*. Universitas Studiorum, Mantova.
2. Camminatiello, I. and A. Lucadamo (2010). Estimating multinomial logit model with multicollinear data. *Asian Journal of Mathematics and Statistics 3*(2), 93–101.
3. De Soete, G. and J. D. Carroll (1994). k-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis*, pp. 212–219. Springer, Heidelberg.
4. Lucadamo, A. and A. Leone (2015). Principal component multinomial regression and spectrometry to predict soil texture. *Journal of chemometrics 29*(9), 514–520.
5. Takane, Y., Hwang, H., 2002. Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research* 37, 163–195.
6. Takane, Y. and T. Shibayama (1991). Principal component analysis with external information on both subject and variables *Psychometrika 56* (1), 97–120.

A two-part finite mixture quantile regression model for semi-continuous longitudinal data

Modello di regressione quantilica mistura a due parti per dati longitudinali

Maruotti Antonello, Merlo Luca and Petrella Lea

Abstract This paper develops a two-part finite mixture quantile regression model for semi-continuous longitudinal data. The components of the finite mixture are associated with homogeneous individuals in the population sharing common values of the model parameters. The proposed methodology allows heterogeneity sources that influence the first level decision process, that is, the model for the binary response variable, to influence also the distribution of the positive outcomes. Estimation is carried out through an EM algorithm without parametric assumptions on the random effects distribution. A penalized version of the EM algorithm is also presented to tackle the problem of variable selection. The suggested modelling framework has been discussed using the extensively investigated RAND Health Insurance Experiment dataset in the random intercept case.

Abstract In questo paper sviluppiamo un modello a due parti mistura per dati longitudinali. Le componenti della mistura catturano cluster di individui che condividono caratteristiche simili. La metodologia adottata suppone che l'eterogeneità nella popolazione influenzi entrambe le due parti del modello, cioè la parte binaria e la parte positiva. La stima dei parametri è ottenuta tramite l'algoritmo EM senza formulare assunzioni parametriche sulla distribuzione degli effetti casuali. Viene proposta, inoltre, una estensione dell'algoritmo per la selezione delle variabili. L'analisi empirica si concentra sul dataset RAND Health Insurance Experiment utilizzando un modello ad intercetta casuale.

Maruotti Antonello
Centre for Innovation and Leadership in Health Sciences, University of Southampton
Department of Economic, Political Sciences and Modern Languages, LUMSA, Rome, Italy, e-mail: a.maruotti@lumsa.it

Merlo Luca
Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, e-mail: luca.merlo@uniroma1.it

Petrella Lea
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano 9, e-mail: lea.petrella@uniroma1.it

Key words: Two-part model, Quantile regression, Random effect models, Longitudinal data, Variable selection, Healthcare expenditure

1 Introduction

Two-part models (TPM), also known as Hurdle models, involve a mixture distribution consisting in a mixing of a discrete point mass (with all mass at zero) and a discrete or continuous random variable. The TPM introduces modelling flexibility by allowing the zero and the positive values to be generated by two different processes. In its basic formulation, the TPM assumes independence between these two processes. Here the modelling framework is extended to model dependence between the spike-at-zero and the positive outcome accounting for potential heterogeneity in both processes. When dealing with longitudinal or hierarchical data, zero inflation may also occur when repeated data are analyzed. In addition, because measurements recorded on the same individual are likely correlated, the potential association between dependent observations should be taken into account in order to provide a correct inference. In such cases, random effect models have been proposed to accommodate for within-subject correlation and between subject heterogeneity, the latter accounting for zero inflation; see [5, 8]. Random effect models accommodate such source of random variation by considering unobserved heterogeneity in the model parameters via individual-specific random coefficients. With a parametric distribution for the random coefficients, one may use either a Monte Carlo EM algorithm or a Maximum Likelihood approach via Gaussian quadrature for parameters estimation. As an alternative, a non-parametric approach can be adopted in which the distribution of the random effects is left unspecified and approximated by using a discrete finite mixture distribution in order to prevent inconsistent parameter estimates due to misspecification of the underlying distribution. Within this scheme, the components of the finite mixture represent clusters of individuals that share homogeneous values of model parameters [2]. In the analysis of real data, applied researchers commonly focusing on estimating the conditional mean, might miss some underlying truth on the entire conditional distribution of the response variable of interest. To overcome those problems, in this paper we propose a quantile regression approach that allows estimating the full range of the conditional quantiles of the response variable. In addition, it is very common that a large number of candidate covariates available are included in the initial stage of modelling for the consideration of removing potential modelling biases. In order to gain in parsimony and to conduct a variable selection procedure, we propose a penalized version of the EM algorithm (PEM) adding the Least Absolute Shrinkage and Selecting Operator (LASSO) L_1 penalty term of [9]. Eventually, this paper can be considered as a unifying framework for the works of [3, 7, 1, 2] modelling time-constant random effects in a two-part finite mixture quantile model for longitudinal data. We examine the empirical behaviour of the proposed approach by the analysis of a sample taken from the RAND Health Insurance Experiment (RHIE).

2 Methodology

Let y_{it}, $i = 1,...,N$, $t = 1...,T$ be a semi-continuous variable for unit i at time t and let $\mathbf{b}_i = (\mathbf{b}_{i0}, \mathbf{b}_{i1})$ be a time-constant, individual-specific, random effects vector having distribution $f_{\mathbf{b}}(\cdot)$ with support $\mathscr{B}$ where $\mathbb{E}[\mathbf{b}_i] = 0$ is used for parameter identifiability. Let us assume that the random variable y_{it} has probability density function given by:

$$f(y_{it}) = p_{it}^{d_{it}} \Big[(1 - p_{it})\, g\big(h(y_{it}) | y_{it} > 0\big) \Big]^{1-d_{it}} \tag{1}$$

with

$$d_{it} = I(y_{it} = 0), \quad p_{it} = \Pr(y_{it} = 0) = \Pr(d_{it} = 1)$$

where d_{it} denotes the occurrence variable for unit i at time t, $g(\cdot)$ is the density function for the positive outcome and $h(y_{it})$ denotes the intensity variable, with $h(\cdot)$ being a transformation function of y_{it}. Formally, the model is completed by defining the linear predictors for the binary and the positive parts of the model. The spike-at-zero process is governed by a binary logistic model such that:

$$\text{logit}(p_{it} \mid \mathbf{v}_{it}) = \mathbf{v}_{it}'\boldsymbol{\gamma} + \mathbf{c}_{it}'\mathbf{b}_{i0} \tag{2}$$

where $\mathbf{v}_{it} = (v_{it1}, \ldots, v_{itm})$ is the m dimensional set of explanatory variables, $\boldsymbol{\gamma}$ its corresponding parameter vector and $\mathbf{c}_{it}$ is a subset of covariates of $\mathbf{v}_{it}$. As mentioned in the Introduction, in this paper we are interested in modelling the truncated at zero part using the quantile regression approach. In particular, the positive part of the dependent variable can be written by exploiting the equivariance property of quantiles to monotone transformations: if $h(\cdot)$ is monotonic increasing on $\mathtt{R}_+$, e.g. the logarithm, the quantiles of the transformed variable are the transformed quantiles of the original one: namely, for a given quantile $\tau \in (0,1)$, $Q_\tau(\cdot|\mathbf{x}_{it}) = h\big(Q_\tau(\cdot|\mathbf{x}_{it})\big)$ where $Q_\tau(\cdot)$ is the quantile function. That is, conditionally on $\mathbf{b}_{i1}$, we assume that, after log-transforming the outcome variable, that is $\tilde{y}_{it} = \log(y_{it})$, the conditional density in (1) is:

$$g(\tilde{y}_{it}|y_{it} > 0, \mathbf{x}_{it}, \mathbf{b}_{i1}) = g_{it} = \frac{\tau(1-\tau)}{\sigma_\tau} \exp\left\{ -\rho_\tau\left(\frac{\tilde{y}_{it} - \mu_{it}}{\sigma_\tau}\right)\right\}, \tag{3}$$

where $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{its})$ represents a covariates s dimensional vector which may differs from $\mathbf{v}_{it}$, σ_τ is the scale parameter and $\rho_\tau(\cdot)$ denotes the quantile asymmetric loss function of [4]. Eq (3) represents an Asymmetric Laplace Distribution (ALD) discussed in [10] whose location parameter μ_{it} is defined by the linear model:

$$\mu_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_\tau + \mathbf{z}_{it}'\mathbf{b}_{i1} \tag{4}$$

where $\mathbf{z}_{it}$ is a subset of covariates of $\mathbf{x}_{it}$. It is easy to see that (4) defines μ_{it} to be the τ-th conditional quantile function of the working variable $\tilde{y}_{it}$ given $y_{it} > 0$ and $\mathbf{x}_{it}$.

As it is clear from (3) and (4), all parameters β_τ, σ_τ and $\mathbf{b}_{i1}$ depend on the quantile level τ. Naturally, if the simple random intercept model is adopted, we have $\mathbf{z}_{it} \equiv 1$.

Parametric assumptions on the distribution of the random coefficients, $f_{\mathbf{b}}(\cdot)$, can be too restrictive and misspecification of the mixing distribution can lead to biased parameter estimates. In addition, an important disadvantage of this approach lies in the required computational effort. For these reasons, we may rely on Nonparametric Maximum Likelihood (NPML) estimation theory of [6]: if $f_{\mathbf{b}}(\cdot)$ is left unspecified, we approximate it by using a discrete distribution on $G < N$ locations $\mathbf{b}_k = (\mathbf{b}_{0k}, \mathbf{b}_{1k})$, with associated probabilities defined by $\pi_k = Pr(\mathbf{b}_i = \mathbf{b}_k)$, $i = 1, ..., N$ and $k = 1, ..., G$. That is, $\mathbf{b}_i \sim \sum_{k=1}^{G} \pi_k \delta_{\mathbf{b}_k}$ where δ_θ is a one-point distribution putting a unit mass at θ. Because responses are assumed to be independent conditional on the random vector $\mathbf{b}_k$, in this case, the likelihood of the model has the form:

$$L(\Phi) = \prod_{i=1}^{N} \left\{ \sum_{k=1}^{G} \prod_{t=1}^{T} p_{it}^{d_{it}} \left[(1 - p_{it})\, g_{itk} \right]^{1-d_{it}} \pi_k \right\}, \tag{5}$$

where, depending on the k-th component of the mixture, the spike-at-zero process is governed by the binary logistic model, $\text{logit}(p_{it} \mid \mathbf{v}_{it}) = \mathbf{v}'_{it}\gamma + \mathbf{c}'_{it}\mathbf{b}_{0k}$, and where g_{itk} is the ALD density in (3) with location parameter given by $\mu_{it} = \mathbf{x}'_{it}\beta_\tau + \mathbf{z}'_{it}\mathbf{b}_{1k}$, for $k = 1, ..., G$.

Locations $\mathbf{b}_k$ and corresponding masses π_k represent unknown parameters, as well as the unknown number of components G, which should be estimated along with other model parameters via selection model techniques. Here, the optimal number of components is based on penalized likelihood criteria such as the AIC and the BIC. As it can be easily noticed, (5) represents the likelihood function of a finite mixture of a quantile TPM with parameters vector $\Phi = \{\gamma, \beta_\tau, \mathbf{b}_1, ..., \mathbf{b}_G, \sigma_\tau, \pi_1, ..., \pi_G\}$.

2.1 Estimation

Given the finite mixture representation in (5), each unit i can be conceptualized as drawn from one of G distinct groups: we denote with w_{ik} a discrete latent variable indicating component membership, i.e. the indicator variable that is equal to 1 if the i-th unit belongs to the k-th component of the finite mixture, and 0 otherwise. We obtain maximum likelihood estimates using the EM algorithm by treating the hidden random variable as missing and representing the complete data set as $(y_{it}, v_{it}, x_{it}, w_{ik})$ for $i = 1, ..., N$, $t = 1, ..., T$ and $k = 1, ..., G$. The log-likelihood for the complete data has the following form:

$$\ell_c(\Phi) = \sum_{i=1}^{N} \sum_{k=1}^{G} w_{ik} \left\{ \log \left(\prod_{t=1}^{T} p_{it}^{d_{it}} \left[(1 - p_{it})\, g_{itk} \right]^{1-d_{it}} \right) + \log(\pi_k) \right\}. \tag{6}$$

In the E-step of the algorithm, the presence of the unobserved group-indicator is handled by taking the conditional expectation of w_{ik} given the observed data and

the parameter estimates at the r-th iteration $\hat{\Phi}^{(r)}$. We replace w_{ik} by its conditional expectation $\hat{w}_{ik}^{(r)}$. The quantity $\hat{w}_{ik}^{(r)}$ is the posterior probability that unit i comes from the k-th component of the mixture model. Subsequently, conditionally on the posterior probabilities $\hat{w}_{ik}^{(r)}$, the M-step solutions are updated using an iteratively weighted least squares (IWLS) algorithm via an appropriate weighted quantile regression for cross sectional. Standard error estimates for model parameters are obtained through non-parametric bootstrap.

With a large number of predictors, one often would like to determine a smaller subset of covariates that exhibit the strongest effects. We implement the variable selection method by maximizing the penalized complete data log-likelihood. Compared to the EM, the PEM algorithm leaves the E-step unchanged and modifies the M-step introducing a penalty function to achieve shrinkage. The log-likelihood for the complete data has the following form:

$$\ell_{pen}(\Phi|\lambda) = \ell_c(\Phi) - \lambda J(\beta_\tau), \tag{7}$$

where $J(\beta_\tau) = ||\beta_\tau||_1$ is the convex LASSO penalty function of [9], $\ell_c(\Phi)$ has been defined in (6) and λ is a tuning parameter that regulates the strength of the penalization assigned to the coefficients in the model which is selected via cross-validation.

3 Application

The above-presented methodology has been applied to study the driving factors of medical expenditures using the data from the RHIE. The RHIE experiment, conducted from 1974 to 1982, collects data from about 8000 enrollees in 2823 families, from six sites across the United States. It assesses how medical care costs affect a patient's use of health services and quality and it is regarded as the basis of the most reliable estimates of price sensitivity of demand for medical services. We consider one measure of utilization: the total spending on health services defined as the sum of outpatient, inpatient, drugs, supplies and psychotherapy expenses. Table 1 reports the number of free model parameters, the log-likelihood and the penalized likelihood criteria (AIC and BIC) for different number of mixture components G at three quantile levels $\tau = (0.25, 0.50, 0.75)$.

References

[1] Alfò, M. and Maruotti, A. [2010], 'Two-part regression models for longitudinal zero-inflated count data', *Canadian Journal of Statistics* **38**(2), 197–216.

[2] Alfò, M., Salvati, N. and Ranallli, M. G. [2017], 'Finite mixtures of quantile and m-quantile regression models', *Statistics and Computing* **27**(2), 547–570.

[3] Geraci, M. and Bottai, M. [2006], 'Quantile regression for longitudinal data using the asymmetric Laplace distribution', *Biostatistics* **8**(1), 140–154.
[4] Koenker, R. and Bassett, G. [1978], 'Regression quantiles', *Econometrica: Journal of the Econometric Society* **46**(1), 33–50.
[5] Lam, K., Xue, H. and Bun Cheung, Y. [2006], 'Semiparametric analysis of zero-inflated count data', *Biometrics* **62**(4), 996–1003.
[6] Lindsay, B. G. et al. [1983], 'The geometry of mixture likelihoods: a general theory', *The annals of statistics* **11**(1), 86–94.
[7] Maruotti, A., Raponi, V. and Lagona, F. [2016], 'Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure', *Biometrical Journal* **58**(2), 280–302.
[8] Min, Y. and Agresti, A. [2005], 'Random effect models for repeated measures of zero-inflated count data', *Statistical modelling* **5**(1), 1–19.
[9] Tibshirani, R. [1996], 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistal Society. Series B* **58**, 267–288.
[10] Yu, K. and Moyeed, R. A. [2001], 'Bayesian quantile regression', *Statistics & Probability Letters* **54**(4), 437–447.

G	2	3	4	5	6
# par.	34	37	40	43	46
$\tau = 0.25$					
log-lik	-36157.85	-35818.09	-35691.69	-35615.53	-35566.29
AIC	72383.71	71710.18	71463.38	71317.06	**71224.58**
BIC	72610.99	71957.52	71730.77	71604.50	**71532.08**
$\tau = 0.50$					
log-lik	-35703.81	-35374.62	-35254.80	-35174.27	-35147.60
AIC	71475.63	70823.25	70589.60	70434.53	**70387.19**
BIC	71702.91	71070.58	70856.99	70721.98	**70694.69**
$\tau = 0.75$					
log-lik	-37379.27	-36614.51	-36422.58	-36374.00	-36313.25
AIC	74826.54	73303.02	72925.17	72834.00	**72718.50**
BIC	75053.82	73550.36	73192.56	73121.44	**73026.00**

Table 1 Number of parameters, log-likelihood and penalized likelihood criteria for different number of mixture components G at quantile levels $\tau = (0.25, 0.50, 0.75)$. The optimal number of components are displayed in boldface.

New developments in the evaluation of goodness of fit for multidimensional IRT models based on posterior predictive assessment: Results from the INVALSI data

Nuovi sviluppi nello studio della bontà di adattamento per i modelli di IRT multidimensionali basati sulla valutazione predittiva a posteriori: Alcuni risultati sui dati INVALSI

Mariagiulia Matteucci and Stefania Mignani

Abstract The issue of model fit assessment is crucial within the framework of item response theory (IRT) models. To overcome the limitations of classical tools which are affected by the problem of sparse data, Bayesian posterior predictive model checking (PPMC) was recently introduced. The purposes of this study are: a) to examine the feasibility of the PPMC method in practice when investigating multidimensionality in IRT models; b) to propose the Hellinger distance within PPMC to be used as a goodness of fit tool. These methods are applied to the INVALSI Italian test data of grade 5. The results support the existence of a predominant general ability without excluding the presence of specific sub-dimensions.

Abstract *Lo studio della bontà di adattamento è di cruciale importanza nell'ambito dei modelli di item response theory (IRT). Per superare i limiti degli approcci classici che risentono della presenza di dati sparsi, sono stati recentemente proposti metodi bayesiani basati sulla valutazione predittiva a posteriori (PPMC). Gli obiettivi di questo lavoro sono: a) studiare le possibilità di utilizzo del metodo*

[1] Mariagiulia Matteucci, Department of Statistical Sciences, University of Bologna; email: m.matteucci@unibo.it

Stefania Mignani, Department of Statistical Sciences, University of Bologna; email: stefania.mignani@unibo.it

PPMC per investigare la multidimensionalità nei modelli di IRT; b) proporre l'uso della distanza di Hellinger nell'ambito PPMC. Questi metodi sono stati applicati ai dati INVALSI relativi alla prova di italiano di livello 5. I risultati evidenziano la presenza di una abilità generale prevalente senza escludere la presenza di sotto-dimensioni specifiche.

Key words: goodness of fit, IRT models, posterior predictive assessment, Hellinger distance, INVALSI data.

Introduction

In educational and psychological measurement, item response theory (IRT) models (see, e.g., van der Linden and Hambleton, 1997) are commonly used to estimate the characteristics of both the categorical items and the test takers. Several IRT unidimensional and multidimensional models have been proposed to account for different data structures. While unidimensional models assume the presence of a single latent variable underlying the response process, the multidimensional ones allow for multiple abilities. In this setting, the issue of model goodness-of-fit is crucial to investigate both absolute and relative fit.

Classical method for assessing model fit suffer from the presence of sparse contingency tables due to the exponential increase in response patterns as the test length increases. In fact, in a frequentist maximum likelihood-based framework, the issue of sparse data prevents the direct application of standard chi-square tests, such as the Pearson chi-square χ^2 and the likelihood ratio G^2, which are full-information statistics based on the complete response patterns. A considerable amount of literature has been published on the development of limited-information statistics based on low-order margins to overcome the sparse data problem, but new computational issues arise as well (among others, see Bartholomew and Leung 2002).

Due to the increasing model complexity, a considerable amount of literature has been recently focused on Bayesian estimation of IRT models via Markov chain Monte Carlo (MCMC) methods due to its flexibility. Starting from a MCMC output, one possibility for examining model fit is using Bayesian posterior predictive model checks (PPMC; Rubin, 1984). Considerable advantages of the method are that it does not rely on distributional assumptions, and it is relatively easy to implement, given that the entire posterior distribution of all parameters of interest is obtained through MCMC algorithms.

PPMC has been used within IRT to check for the behavior of unidimensional models fitted to potential multidimensional data (see, among others, Sinharay, 2006; Sinharay, Johnson, and Stern, 2006; Levy, Mislevy, and Sinharay, 2009; Levy and Svetina, 2011). In these studies, PPMC has been implemented with graphical analyses and the estimation of the posterior predictive *p*-values (PPP-values) to investigate the degree to which observed data are expected under the model, given a discrepancy measure.

The aim of this study is to investigate the dimensionality of response data coming from the administration of the INVALSI Italian language test to students at the end of the lower secondary school (grade 5) in the scholastic year 2015/2016 (INVALSI, 2016). Classical PPMC methods are used and a new solution based on the Hellinger distance to measure the distance between the realized and the predictive distributions is proposed.

PPMC within IRT Models

PPMC techniques are based on the comparison of observed data with replicated data generated or predicted by the model by using a number of diagnostic measures that are sensitive to model misfit (Sinharay, Johnson, and Stern, 2006). Given the data $\mathbf{y}$, let $p(\mathbf{y}|\boldsymbol{\omega})$ and $p(\boldsymbol{\omega})$ be the likelihood for a model depending on the set of parameters $\boldsymbol{\omega}$ and the prior distribution for the parameters, respectively. In the IRT context, $\boldsymbol{\omega}$ consists of the item parameters, person parameters, and trait correlations.

Once defined a suitable discrepancy measure $D(\cdot)$ able to capture relevant features of the data, a graphical analysis is conducted to show the differences among realized and replicated discrepancy measures. Then, it is possible to estimate the PPP-value by computing the proportion of MCMC replications which satisfy the following

$$\text{PPP-value} = p\big(D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega}|\boldsymbol{y})\big). \qquad (1)$$

The PPP-values provide a measure of the degree to which observed data would be expected under the model: values close to 0 or 1 mean that the realized values fall far in the tails of the distribution of the discrepancy measure based on the posterior predictive distribution, indicating misfit.

Effective diagnostic measures in checking for different dimensionality structures such as unidimensional, multi-unidimensional, additive (Sheng and Wikle, 2009) are based on the association or on covariance/correlation among item pairs. Examples are the Mantel-Haenszel (MH) statistic and the model-based covariance (MBC), see Levy, Mislevy, and Sinharay (2009).

While the PPP-value counts the number of replications for which the predictive discrepancy exceeds the realized one, the researcher may be interested in measuring the size of the difference itself. For this reason, we propose the use of the Hellinger distance, based on the Hellinger integral (Hellinger, 1909), which is symmetric, it does obey the triangle inequality and its range is 0-1. Since the Hellinger distance is used to quantify the distance between two probability measures, it can be used to measure the distance between the realized and the predictive distribution within PPMC as follows

$$\mathrm{H}(P,Q) = \sqrt{1 - \int \sqrt{p\big(D(\boldsymbol{y},\boldsymbol{\omega})\big)p(D(\boldsymbol{y}^{rep},\boldsymbol{\omega}))}d\boldsymbol{y}d\boldsymbol{\omega}}. \quad (2)$$

The direct calculation of (2) is computationally demanding and it is usually done via MCMC. Specifically, it is calculated by using the normal kernel density estimates to represent the probability density functions of the realized and the predictive discrepancy measures, given the MCMC replications. In order to check for model unidimensionality, we propose the use of the Hellinger distance with the MBC discrepancy measure, which is based on both data and model parameters, to take into account a fit measure for each item pair.

In this work, models for binary data are taken into account, where Y_{ij} represents the response variable for respondent i to item j, with i=1,…,n and j=1,…,k, taking the values 1 and 0 to represent the dichotomy. By considering the class of models with two item parameters, the unidimensional model is formulated as the two-parameter normal ogive (2PNO) model (Lord and Novick, 1968) as follows

$$P\big(Y_{ij} = 1|\theta_i, \alpha_j, \delta_j\big) = \Phi\big(\alpha_j\theta_i - \delta_j\big), \quad (3)$$

where the probability of a positive response for item j by individual i is expressed as a function of the item parameters α_j and δ_j and the latent trait score θ_i. By considering
a test designed to assess a set of m domains, i.e., a test with k items is divided into m subtests each containing k_v items, where v=1,..,m, the unidimensional model (3) can be extended to the 2PNO multi-unidimensional model (Sheng and Wikle, 2007), as follows

$$P\big(Y_{vij} = 1|\theta_{vi}, \alpha_{vj}, \delta_j\big) = \Phi\big(\alpha_{vj}\theta_{vi} - \delta_j\big), \quad (4)$$

where the probability of a positive response is expressed for item j belonging to subtest v by individual i, and parameters are specific for the v-th dimension. The latent traits can potentially be correlated. Lastly, when the concurrent presence of a general and m specific latent traits is assumed, the 2PNO additive model (Sheng and Wikle, 2009), can be expressed as follows

$$P\big(Y_{vij} = 1|\theta_{0i}, \theta_{vi}, \alpha_{0j}, \alpha_{vj}, \delta_j\big) = \Phi\big(\alpha_{0j}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_j\big), \quad (5)$$

In addition to the measurement structure of model (4), model (5) assumes the presence of an overall trait θ_0, which is related to all items by means of the general discrimination parameters α_0. As can be clearly seen, a lack in the general trait can be compensated by the specific trait in determining the probability of a positive response, and vice versa, because the nature of the linear predictor is compensatory.

Results

The Italian language INVALSI 2016 test for grade 5 consists of 33 reading and comprehension items and 10 grammar items. Binary response data are used where 1 means a correct response and 0 an incorrect one. A sample of n=5083 examinees is considered. Despite the test is scored under the assumption of unidimensionality, the presence of two different subgroups of items suggests the investigation of a multidimensional structure in the data. The following models were estimated: unidimensional, multi-unidimensional with two specific dimensions (reading comprehension and grammar abilities) and additive assuming a multi-unidimensional structure with the addition of an overall dimension (general Italian language ability) where all the traits may correlate.

Table 1 shows the first results. The correlation between the two specific traits in the multi-unidimensional model is rather strong (about 0.8) while, by introducing an overall trait in the model in the additive approach, the specific traits look more correlated to the general Italian language ability. In order to make a first comparison of the model performances, the deviance information criterion (DIC) was computed. A better fit for the additive model is shown, even if in presence of small differences among the DIC values.

Table 1: Estimated trait correlations (ρ), DIC and proportions of extreme PPP-values for the three different models.

Model	ρ_{01}	ρ_{02}	ρ_{12}	***DIC***	***Extreme PPP-values < 0.05 or > 0.95***	
					MH	***MBC***
Unidimensional	-	-	-	234144.65	0.367	0.346
Multi-unidimensional	-	-	0.814 (0.00)	232703.51	0.308	0.252
Additive	0.731 (0.05)	0.596 (0.01)	0.378 (0.03)	231630.26	0.210	0.197

By considering the MH statistic and the MBC for all the 903 different item pairs, the results show that the proportion of extreme PPP-values decreases as the number of dimensions in the model increases. In fact, around a 35% of extreme PPP-values are observed for the unidimensional model for both discrepancy measures while the proportion reduces to around 20% for the additive model. These findings are further confirmed from the plots in Figure 1, where the item pairs with extreme PPP-values are reported. From the figure, it is clear that the additive model is associated to the smallest number of item pairs showing poor fit.

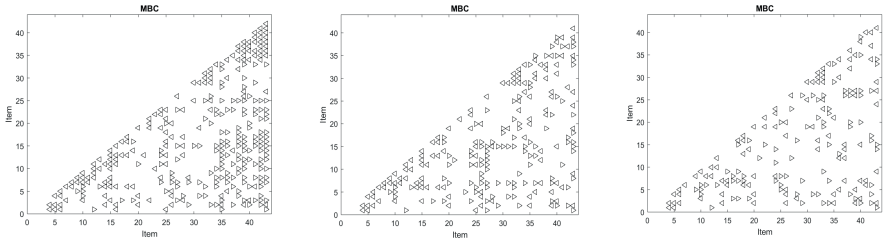


Figure 1: Plots of item pairs with extreme PPP-values for the MBC for the three different models (Note: right triangles indicate PPP-values greater than 0.95; left triangles indicate PPP-values lower than 0.05).

Table 2: Results of the MBC-Hellinger distance for the three different models.

Model	*Mean*	*Sd*	*Median*	*Min*	*Max*
Unidimensional	0.756	0.134	0.729	0.495	1.000
Multi-unidimensional	0.704	0.139	0.678	0.370	1.000
Additive	0.634	0.152	0.610	0.200	1.000

Table 2 reports some statistics computed for the estimated values of the Hellinger distance used with MBC. Average values among item pairs are 0.76, 0.70, and 0.63 for the unidimensional, multi-unidimensional, and additive models, respectively, with similar variability. This means that, once again, the additive model seems to fit the data best among the other two competing models. This is further remarked in Figures 2, 3 and 4 where plots highlighting the item pairs with values of the Hellinger distance higher than 0.5 or 0.8 are represented. From the figures, it is clear that increasing the model complexity reduces the number of item pairs with "high" values for the Hellinger distance.

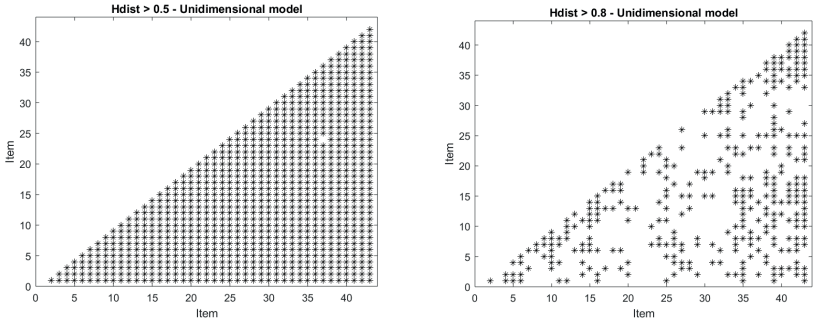


Figure 2: Plots of item pairs with values of the MBC-Hellinger distance larger than 0.5 (on the left) or 0.8 (on the right) for the unidimensional model.

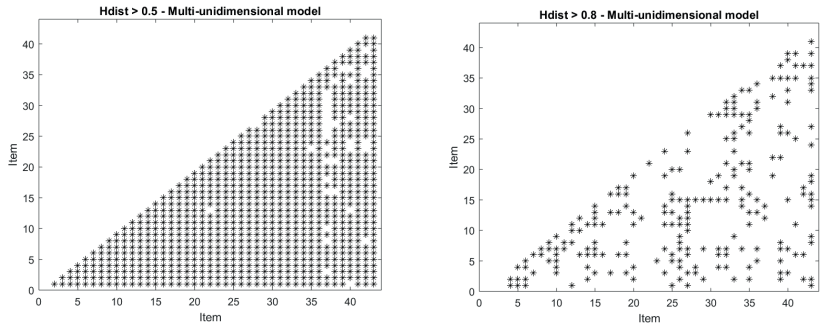


Figure 3: Plots of item pairs with values of the MBC-Hellinger distance larger than 0.5 (on the left) or 0.8 (on the right) for the multi-unidimensional model.

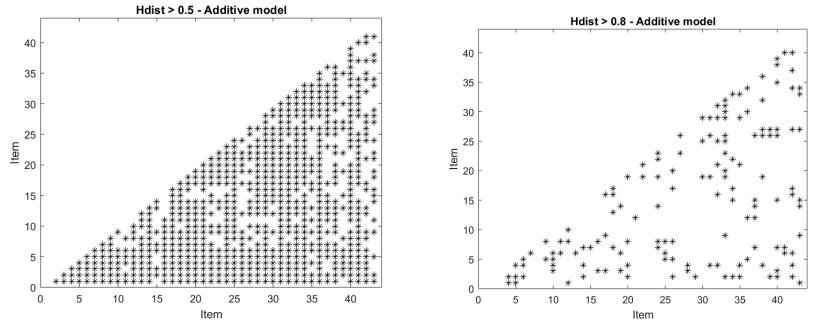


Figure 4: Plots of item pairs with values of the MBC-Hellinger distance larger than 0.5 (on the left) or 0.8 (on the right) for the additive model.

Discussion

The results show that both the amount of extreme PPP-values and the Hellinger distance are useful in assessing model fit and these measures can be used to investigate misfit due to specific items. We believe that the presence of a predominant general latent ability, besides the specific abilities, is well supported by the response data from the INVALSI Italian language test. However, these results should be deepened also by considering the different types of item pairs: both items belonging to the same subtest or items belonging to different subtests.

Future methodological research should consider finding a different discrepancy measure which is able to detect overall misfit and is most sensitive to different sources of misfit. Moreover, the performance of the Hellinger distance for model comparison should be further investigated.

References

1. Bartholomew, D.J., Leung, S. O. A goodness of fit test for sparse 2^p contingency tables. Br. J. Math. Stat. Psychol. **55**, 1-15 (2002).
2. Gelman, A., Meng, X.L., Stern, H.S.: Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. **6**, 733-807 (1996).
3. Hellinger, E.: Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. Journal für die reine und angewandte Mathematik (in German) **136**, 210–271 (1909).
4. INVALSI: Rilevazioni nazionali degli apprendimenti 2015-2016. La rilevazione degli apprendimenti nelle classi II e V primaria, nella classe III (Prova nazionale) della scuola secondaria di primo grado e nella II classe della scuola secondaria di secondo grado. https://invalsi-areaprove.cineca.it/docs/file/08_Rapporto_Prove_INVALSI_2016.pdf (2016).
5. Levy, R., Svetina, D.: A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. Br. J. Math. Stat. Psychol. **64**, 208-232 (2011).
6. Levy, R., Mislevy, R.J., Sinharay, S.: Posterior predictive model checking for multidimensionality in item response theory. Appl. Psychol. Meas. **33**, 519-537 (2009).
7. Lord, F.M., Novick, M. R.. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA (1968).
8. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applies statistician. Ann. Stat. **12**, 1151-1172 (1984).
9. Sheng, Y., Wikle, C.: Comparing multiunidimensional and unidimensional item response theory models. Educ. Psychol. Meas. **67(6)**, 899-919 (2007).
10. Sheng. Y., Wikle. C.: Bayesian IRT models incorporating general and specific abilities. Behaviormetrika **36**, 27-48 (2009).
11. Sinharay, S.: Bayesian item fit analysis for unidimensional item response theory models. Posterior predictive assessment of item response theory models. Br. J. Math. Stat. Psychol. **59**, 429-449 (2006).
12. Sinharay, S., Johnson, M.S., Stern, H.S.: Posterior predictive assessment of item response theory models. Appl. Psychol. Meas. **30**, 298-321 (2006).
13. van der Linden, W. J., Hambleton, R.K. Handbook of Modern Item Response Theory. Springer-Verlag, New York (1997).

Models for measuring well-being. Reflective or Formative?

Modelli per misurare il benessere. Riflessivi o formativi?

Matteo Mazziotta and Adriano Pareto

Abstract Measuring well-being is a very difficult task as it is characterized by a multiplicity of aspects or dimensions. The complex and multidimensional nature of this concept requires the definition of partial goals whose achievement can be observed and measured by individual indicators. Finally, individual indicators can be summarized, on the basis of an specific measurement model, in order to construct a single index of well-being. In this paper, we discuss which measurement models can be used and why.

Abstract *Misurare il benessere è un compito molto difficile, in quanto esso è caratterizzato da una molteplicità di aspetti o dimensioni. La natura complessa e multidimensionale di questo concetto richiede la definizione di obiettivi parziali il cui raggiungimento può essere osservato e misurato per mezzo di indicatori elementari. Gli indicatori elementari possono, poi, essere sintetizzati, sulla base di uno specifico modello di misura, per costruire un singolo indice di benessere. In questo lavoro, si discute quali modelli possono essere usati e perché.*

Key words: Multivariate Analysis, Measurement model, Composite index

1 Introduction

In recent years, interest in the measurement of well-being at the community, national and international levels has grown greatly. In particular, the 2009 Stiglitz-Sen-Fitoussi Commission on "The Measurement of Economic Performance and Social Progress" concluded that it is necessary to move "Beyond GDP" when assessing a country's health, and complement GDP with a broader set of individual indicators

[1] Matteo Mazziotta, Italian National Institute of Statistics; email: mazziott@istat.it
Adriano Pareto, Italian National Institute of Statistics; email: pareto@istat.it

that would reflect the distribution of well-being in all of its social, economic and environmental dimensions.

When a large set of individual indicators is available, it can be useful to construct a composite index for each dimension or an overall composite index of well-being.

One of the main issues in constructing composite indices is determining an appropriate measurement model for aggregating individual indicators. In this work, the two major models are evaluated and compared, emphasising their properties and discussing when and why they can be used for measuring well-being.

2 Measurement models: reflective and formative approach

As it is known, a model[1] of measurement can be conceived through two different conceptual approaches: reflective or formative (Jarvis et al. 2003; Diamantopoulos and Winklhofer 2001; Diamantopoulos et al. 2008; Coltman et al. 2008).

The most popular approach is the reflective model, according to which individual indicators denote effects (or manifestations) of an underlying latent variable. Therefore, causality is from the concept to the indicators and a change in the phenomenon causes variation in all its measures. In this model, the construct exists independently of awareness or interpretation of the researcher, even if it is not directly measurable (Borsboom et al. 2003).

Specifically, the latent variable R represents the common cause shared by all indicators X_i reflecting the construct, with each indicator corresponding to a linear function of the underlying variable plus a measurement error:

$$X_i = \lambda_i R + \varepsilon_i \tag{1}$$

where X_i is the indicator i, λ_i is a coefficient (loading) capturing the effect of R on X_i and ε_i is the measurement error for the indicator i. Measurement errors are assumed to be independent and unrelated to the latent variable.

A fundamental characteristic of reflective models is that individual indicators are interchangeable (the removal of one of the indicators does not change the essential nature of the underlying concept) and correlations between indicators are explained by the measurement model (all indicators must be intercorrelated).

Another important issue concerns the polarity of the individual indicators. The 'polarity' of a individual indicator is the sign of the relation between the indicator and the concept to be measured. For example, in the case of well-being, "Life expectancy" has positive polarity, whereas "Unemployment rate" has negative polarity. In a reflective model, individual indicators with equal polarities must be positively correlated, whereas individual indicators with opposite polarities must be negatively correlated. Otherwise, the model will produce inconsistent results (Mazziotta and Pareto 2019).

[1] For the sake of simplicity, only linear models will be considered.

A typical example of reflective model is the measurement of intelligence of a person. In that case, it is the 'intelligence level' that influences the answers to a questionnaire for measuring attitude, and not vice versa. Hence, if the intelligence of a person increased, this would be accompanied by an increase of correct answers to all questions (Simonetto 2012).

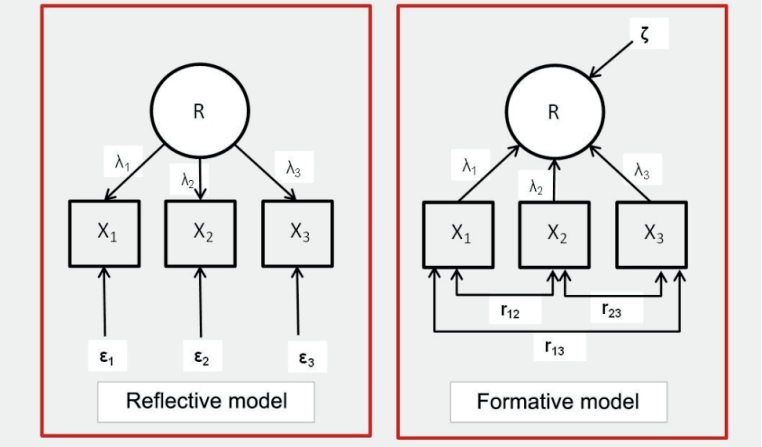


Figure 1: Alternative measurement models

The second approach is the formative model, according to which individual indicators are causes of an underlying latent variable, rather than its effects. Therefore, causality is from the indicators to the concept and a change in the phenomenon does not necessarily imply variations in all its measures. In this model, the construct is defined by, or is a function of, the observed variables.

The specification of the formative model is:

$$R = \sum_i \lambda_i X_i + \zeta \qquad (2)$$

where λ_i is a coefficient capturing the effect of X_i on R, and ζ is an error term[1].

In this case, indicators are not interchangeable (omitting an indicator is omitting a part of the underlying concept) and correlations between indicators (r_{ij}, $i \neq j$) are not explained by the measurement model (high correlations between indicators are possible, but not generally expected). So, in a formative model, polarities and correlations are independent and individual indicators can have positive, negative or zero correlations.

A typical example of formative model is the measurement of socioeconomic status of a person. It depends on education, income, occupation and residence, and not vice versa. So, if any one of these factors improved, the socioeconomic status would increase (even if the other factors did not change). However, if a person's

[1] Some authors exclude the error term so that Equation (2) reduces to a weighted linear combination of the X_i (Diamantopoulos 2006).

socioeconomic status increased, this would not necessarily be accompanied by an improvement in all factors.

In Figure 1, the two different approaches are graphically represented. Traditionally, the reflective model is applied in the development of scaling models for subjective measurement (e.g. attitude or satisfaction scale construction), whereas the formative model is commonly used in the construction of composite indices based on both objective and subjective indicators (Maggino and Zumbo 2012).

Note that (1) is a system of simple regression equations where each individual indicator is the dependent variable and the latent variable is the explanatory variable; whereas (2) represents a multiple regression equation where the latent variable is the dependent variable and the indicators are the explanatory variables. Hence, the correct interpretation of the relationships between indicators and latent variable allows the procedure aimed at aggregating individual indicators to be correctly identified[1] (Maggino 2017).

The properties of the two models are shown in Table 1. Although the reflective view dominates the psychological and management sciences, the formative view is common in economics and sociology (Coltman et al. 2008).

Table 1: Properties of the reflective and formative models

Topic	Reflective Model	Formative model
Nature of construct	Latent construct is existing	Latent construct is created
Direction of causality between indicators and construct	Causality from construct to indicators	Causality from indicators to construct
Characteristics of indicators used to measure the construct	Indicators are manifested by the construct, they share a common theme and they are interchangeable	Indicators define the construct, they need not share a common theme and they are not interchangeable
Indicators intercorrelation and polarity (a)	Indicators with equal polarities must be positively correlated, whereas indicators with opposite polarities must be negatively correlated	Polarities and correlations are independent and indicators can have any pattern of intercorrelation
Measurement error	Error term is clustered at the indicator level	Error term is clustered at the construct level

[1] Note that some multivariate statistical techniques, such as Factor Analysis and Principal Component Analysis are, explicitly or implicitly, based on a reflective measurement model (Mazziotta and Pareto 2019).

3 Aspects of well-being and measurement models

The well-being of a society can be defined as a benefit for all people in the society, implying accomplishment of adequate economic development (objective dimension of well-being) and the resulting positive perception of people towards the proper stage in the society, i.e. the quality of life (subjective dimension of well-being).

So, when we wish to measure well-being, we need to distinguish between objective and subjective well-being (Ivković et al. 2014). Whereas objective social indicators are measures of social reality not filtered by perceptions and independent from personal assessments (e.g., counting the occurrences of given phenomena), subjective indicators are measures supposed to explicitly revealing subjective states, such as feelings, evaluations and preferences (e.g. self-rating scales of satisfaction).

In Figure 2, these two different aspects are graphically represented.

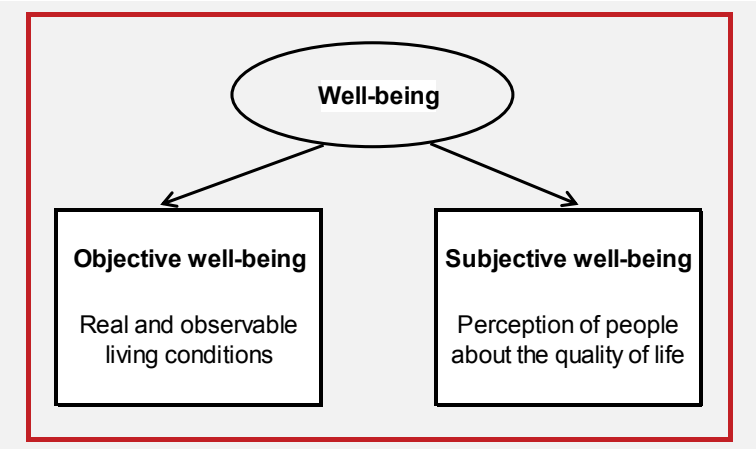


Figure 2: Alternative aspects of well-being

In this paper, we consider two typical cases of measurement of well-being: one based on objective data from official statistics and the other based on subjective data from survey questionnaires.

3.1 The Human Development Index

When the unit of analysis is a country, a region or any geographical unit and individual indicators represent different factors, such as health, income, occupation, services and environment, we can assume that well-being depends on (or is a function of) these factors. In such case, individual indicators do not share a common theme, they are not interchangeable and they can be incorrelated. So, a formative measurement model should be adopted.

One of the oldest and most famous formative composite indices is the Human Development Index (HDI) by United Nations Development Programme (UNDP 1990, 2010). It is a composite measure of human development of countries that includes three theoretical dimensions: (i) health, (ii) education, and (iii) standard of living (UNDP 2018). Any change in one or more of these components is likely to cause a change in a country's HDI score, but there is no reason to expect the components are intercorrelated.

The indicators used are listed in Table 1. They are normalized into a 0-1 scale by using the respective minimum and maximum values, as *goalposts*[1].

Table 1 List of individual indicators of the HDI

N.	Description	Minimum value	Maximum value
	HEALTH		
1	Life expectancy at birth (years)	20	85
	EDUCATION		
2	Expected years of schooling (years)	0	18
3	Mean years of schooling (years)	0	15
	STANDARD OF LIVING		
4	Gross National Income (GNI) per capita (2011 PPP $)	100	75,000

The HDI is the geometric mean of three dimensional indices:

$$\mathrm{HDI}_i = \left(y_{i1} \cdot y_{i5} \cdot y_{i4}\right)^{1/3}$$

where

$$\mathrm{y}_{i5} = \left(y_{i2} + y_{i3}\right)/2$$

and y_{ij} is the normalized value of indicator j for country i.

The main characteristic of this methodology is that it assumes imperfect substitutability across all dimensions of human development (i.e., the HDI is a partially compensatory composite index). It thus addresses one of the most serious criticisms of the linear aggregation formula, which allowed for perfect substitution across dimensions (UNDP 2010).

[1] Note that the logarithm of income is used for the GNI per capita normalization.

3.2 The Personal Well-being Index

When the unit of analysis is a person and individual indicators represent answers to a questionnaire on people's feelings, we can assume that the answers reflect (or are manifestations of) well-being of the person. In such case, individual indicators share a common theme, they are interchangeable and they are intercorrelated. So, a reflective measurement model should be adopted.

An important example of reflective composite index is the Personal Well-being Index (PWI) by Australian Centre on Quality of Life (International Well-being Group 2013). It is a composite measure of well-being of persons that includes seven items related to satisfaction with different life domains: (i) standard of living, (ii) health, (iii) achievements, (iv) relationships, (v) safety, (vi) community-connectedness, and (vii) future security.

Respondents indicate their level of satisfaction with each life domains using an 11-point end defined scale (0 = 'no satisfaction at all'; 10 = 'completely satisfied'). Raw scores are then transformed in normalized indicators by rescaling the values into a 0-100 point scale.

The PWI is the arithmetic mean of seven dimensional indices.

$$\mathrm{PWI}_i = \frac{\sum_{j=1}^{7} y_{ij}}{7}$$

where y_{ij} is the normalized value of indicator j for unit i.

Using the arithmetic mean allows perfect substitutability across all domains of quality of life (i.e., the PWI is a full compensatory composite index).

Factor analysis showed that these domains consistently represent a single latent factor – say 'Satisfaction with Life as a Whole' – that accounts for about 50% of the variance of individuals in Australia and other countries (Cummins et al. 2003).

4 Conclusions

In general, the choice of a formative versus a reflective model depends on the causal priority between individual indicators and latent variable (Bollen 1989).

More specifically, existing constructs, such as "personality", "attitude", "perception" and other intrinsic characteristics of the statistical unit, are typically viewed as underlying factors that give rise to something that is observed. Their indicators tend to be realized and then the model is reflective.

On the other hand, when constructs are created and conceived as explanatory combinations of indicators, such as "Life expectancy" or "GNI per capita", the model should be formative (Diamantopoulos and Winklhofer 2001).

In order to obtain a valid and reliable measure, it is absolutely essential to define the theoretical framework with an appropriate measurement model. This paradigm should always be considered when the aim of the research is to measure a

multidimensional phenomenon through composite indices. And this is even more valid if the phenomenon to be measured is well-being.

The paper is the result of the common work of the authors: in particular, M. Mazziotta has written Sections 2 and 4, and A. Pareto has written Sections 1 and 3.

References

1. Bollen, K.: Structural Equations with Latent Variables. New York, John Wiley & Sons (1989)
2. Borsboom, D., Mellenbergh, G.J., Heerden, J.V.: The theoretical status of latent variables. Psychological Review, 110, 203-219 (2003)
3. Coltman, T., Devinney, T.M., Midgley, D.F., Venaik, S.: Formative versus reflective measurement models: Two applications of formative measurement. Journal of Business Research, 61, 1250-1262 (2008)
4. Cummins, R.A., Eckersley, R., Pallant, J., Van Vugt, J., Misajon, R.: Developing a national index of subjective wellbeing: the australian unity wellbeing index. Social Indicators Research, 64, 159-190 (2003)
5. Diamantopoulos, A.: The error term in formative measurement models: interpretation and modeling implications. Journal of Modeling in Management, 1, 7-17 (2006)
6. Diamantopoulos, A., Winklhofer, H.M.: Index Construction with Formative Indicators: An Alternative to Scale Development. Journal of Marketing Research, 38, 269-277 (2001)
7. Diamantopoulos, A., Riefler, P., Roth, K.P.: Advancing formative measurement models. Journal of Business Research, 61, 1203-1218 (2008)
8. International Wellbeing Group: Personal Wellbeing Index - 5th Edition. Melbourne, Australian Centre on Quality of Life, Deakin University (2013)
9. Ivković, A.F., Ham, M., Mijoč, J.: Measuring Objective Well-Being and Sustainable Development Management. Journal of Knowledge Management, Economics and Information Technology, Vol. IV, Issue 2 (2014)
10. Jarvis, C.B., Mackenzie, S.B., Podsakoff, P.M.: A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. Journal of Consumer Research, 30, 199-218 (2003)
11. Maggino, F.: Developing Indicators and Managing the Complexity. In: F. Maggino (eds.), Complexity in Society: From Indicators Construction to their Synthesis, pp. 87-114. Cham, Springer (2017)
12. Maggino, F., Zumbo, B.D.: Measuring the Quality of Life and the Construction of Social Indicators. In: K.C. Land, A.C. Michalos, M.J. Sirgy (eds.), Handbook of Social Indicators and Quality-of-Life Research, pp. 201-238. Dordrecht, Springer (2012)
13. Mazziotta, M., Pareto, A.: Use and Misuse of PCA for Measuring Well‑Being. Social Indicators Research, 142, 451-476 (2019)
14. Simonetto, A.: Formative and reflective models: State of the art. Electronic Journal of Applied Statistical Analysis, 5, 452-457 (2012)
15. UNDP: Human Development Report 1990. New York, Oxford University Press (1990)
16. UNDP: Human Development Report 2010. New York, Palgrave Macmillan (2010)
17. UNDP: Human Development Indices and Indicators. 2018 Statistical Update (Technical notes). New York, United Nations Development Programme (2018)

Automated Content Analysis of Destination Image: a Case Study

Automated Content Analysis dell'immagine della destinazione: studio di un caso

Antonino Mario Oliveri[1] and Gabriella Polizzi[2]

Abstract Automated content analysis has become one of the most used approaches to extract "hidden" dimensions from text corpora over the last years.
One of the data analysis techniques belonging to this approach is topic modeling, which can be fruitfully used to analyse complex phenomena like tourist destination image.
With this aim in mind, this paper discusses the use of topic modeling to identify the main components of the image of cruise holidays spread through a specific type of visual text, i.e. the Television commercial.
In order to achieve this goal, the paper presents the methodology and main results of a study carried out over a sample of TV commercials, which have recently been broadcast on Television by four of the major cruise lines operating in Italy.
Abstract *Negli ultimi anni l'Automated content analysis è diventata uno degli approcci più seguiti per estrarre dimensioni interpretative "nascoste" dall'interno di corpora testuali.*
Una delle tecniche di analisi dei dati appartenenti a questo approccio, il topic modeling, può essere utilmente impiegata per investigare fenomeni complessi come l'immagine di destinazioni turistiche.
A tale scopo, questo lavoro discute l'impiego del topic modeling per identificare le principali componenti dell'immagine di vacanze in crociera, come diffuse attraverso un particolare tipo di testo visuale: lo spot televisivo.
Per raggiungere questo obiettivo, il lavoro presenta metodologia e principali risultati di uno studio condotto su un campione di spot televisivi, che sono stati recentemente trasmessi in televisione da quattro delle principali compagnie crocieristiche operanti in Italia.

Key words: Automated content analysis, Topic modeling, Destination image, Television commercials, Cruise lines.

---

1 Department Cultures and Society, Università degli studi di Palermo, Italy; e-mail: antoninomario.oliveri@unipa.it
2 Faculty of Human and Social Sciences, Università degli Studi di Enna "Kore", Italy; email: gabriella.polizzi@unikore.it

Introduction

Automated content analysis has often been used in many research fields – tourism included – to analyse textual data in recent years (Hagen, 2018; Hu et al., 2019). One of its data analysis techniques, topic modeling, has proved able to "... effectively uncover the hidden structures of short texts (…) and normal long texts..." (Li et al., 2018: 1345).

However, these models have not yet been used over complex texts, like TV commercials.

In the field of tourism, TV commercials are often produced and broadcast to spread destination image, which is one of the elements that most affect tourists' decision-making processes (Baloglu and McCleary, 1999). Given that a distinctive image can differentiate a destination from its competitors, destinations usually compete also via images (Urry, 1990).

In this regard, cruise tourism is no exception, since "image is what sells cruises" (Klein, 2002: 1), which has led to significant growth in advertising activities aiming to promote the cruise ship as *the destination in itself* since the 1970s (Wood, 2004).

Starting from this background, this paper presents the methodology and main results of automated textual analysis carried out on a sample of recent TV commercials broadcast by four of the major cruise lines operating in Italy, i.e. Costa Cruises, MSC Cruises, Royal Caribbean International and Grimaldi Lines.

The ultimate research aim is to identify the main components of the TV image of cruise holidays.

The Use of Topic Modeling for Automated Content Analysis

The basic assumption behind topic modeling is that "it exists a latent topic level beyond the observable word level, where each topic is a multinomial distribution over the vocabulary" (Li et al., 2018: 1345). As a consequence, unsupervised or supervised algorithms can be applied to textual data in order to extract relevant topics which characterise texts and help to analyse their content.

Unsupervised algorithms model only the words included in the documents under scrutiny, aiming at inferring "... topics that maximize the likelihood (or the posterior probability) of the collection" (Mcauliffe, Blei, 2008: 121).

Supervised algorithms include one or more response variables, for prediction purposes.

A variety of topic modeling algorithms have been proposed so far, such as the Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Jelodar et al., 2019) which is the most commonly used, the Dirichlet Multinomial Mixture (DMM) (Nigam et al., 2000) and the Correlated Topic Modeling (Blei, Lafferty, 2007), among others.

Among the three topic modeling algorithms mentioned above, the LDA algorithm assumes that documents can be interpreted as mixture distributions over latent topics, while topics "(...) are shared by all documents in the collection, but the

topic proportions vary stochastically across documents, as they are randomly drawn from a Dirichlet distribution" (Blei, Lafferty, 2007: 18).

The core of DMM is the idea that several sub-topics can be identified within a topic in terms of different (multinomial) word distributions. The mixture of distributions can therefore better account for the co-occurrence of words which identify different sub-topics.

Correlated Topic Modeling assumes that different topics correlate one another, which the authors consider not taken into adequate account by other algorithms like LDA (Blei, Lafferty, 2007).

Topic modeling has been performed over different corpora so far. For instance, Hagen (2018) implemented LDA over e-petitions, extracting 30 topics and showing high correlation with the results obtained from manual content analysis.

However, to our knowledge applications of topic modeling to the analysis of TV commercials are not common practice in the field of tourism. On the other hand, their implementation seems really promising since it permits to draw information from texts without imposing any constraints over data.

TV commercials are complex texts that combine visual elements (i.e. still or moving images), sounds (i.e. music and spoken words), and written texts (i.e. words than can be read on the TV screen), and are used for many commercial purposes including the diffusion and promotion of specific destination images, as explained in Section 3.

Analyzing the Image Components of a Tourist Destination: the Case of Cruise Ship Holidays

In contemporary cruise tourism, the cruise ship can be considered as *the destination in itself* (Wood, 2004). For this reason, destination image literature and some of its key-frameworks and concepts can be profitably applied to the field of cruise tourism.

Polizzi and Oliveri (2017) carried out a literature review on previous studies about motivations for cruising, cruise experiences and cruise vacation satisfaction, so as to draw those *image attributes of a cruise ship holiday* which have been shown to be the most recurrent. The image attributes emerging from this review were then aggregated into eight *image components*, namely, (i) *escape*; (ii) *relaxation*; (iii) *entertainment*; (iv) *learning*; (v) *prestige*; (vi) *family and social relationships*; (vii) *services and products provided by the cruise line*; and (viii) *environment*.

Looking at the general field of travel decision-making process, *escape* as well as *relaxation* have always been considered *push factors* (Crompton, 1979), i.e. 'socio-psychological motives' internal to the individual which make him/her wish to go on holiday. Similarly, *escape* and *relaxation* have been found as primary reasons that motivate people to take a cruise (Hung, Petrick, 2011).

More specifically, the *escape* image component involves consumers' *active participation and immersion* in experiences that they usually cannot live in their

daily life (Pine, Gilmore, 1998). For example, taking part in a nightly play onboard as actors or singers can be a way for cruisers to escape their ordinary life by immersing themselves in a highly involving activity.

According to Crompton (1979, p. 417), *relaxation* means "taking the time to pursue activities of interest" and refers "to a mental rather than a physical state", which explains why tourists are often willing to experience physical exhaustion or fatigue together with mental refreshment during a holiday.

Since the advent of the Carnival Cruise Lines 'Fun Ships' in the 1970s (Dickinson, 1995; Kwortnik, 2006), *entertainment* has always been a key component of any cruise vacation; it entails *passive participation and absorption* (Pine, Gilmore, 1998) in the different experiences provided onboard or on the mainland, such as nightly shows and casino-style gaming (Hosany, Witham, 2009).

The *learning* image component refers to what several scholars have also called "education", which involves consumers' *active participation and absorption* in certain experiences (Pine, Gilmore, 1998) and can fulfil the need for *exploration and evaluation of self* (Crompton, 1979). For examples, when embarking on a cruise vacation, passengers can participate in a wide range of educational experiences such as direct encounters with people from different cultures, as well as onboard cruise activity programmes (i.e. dancing lessons, cooking, and expert talks about sea life) (Cartwright, Baird, 1999; Hosany, Witham, 2009).

The *prestige* image component refers to the idea of "unique qualities" of a destination or single tourist services (Crompton, 1979). Among cruising motivations, Hung and Petrick (2011) conceptualized *prestige* in terms of *social recognition*, which would increase whenever taking a cruise is a way *to do something that impresses others* as well as *to have a high status vacation*.

As with other types of tourist experiences, *family and social relationships* are fundamental image components of a cruise ship holiday (Huang, Hsu, 2010; Hung, Petrick, 2011; Teye, Paris, 2010), which may be chosen by people who are searching for what Crompton (1979) called *enhancement of kinship relationships* as well as *facilitation of social interaction* with new people.

Services and products provided by the cruise lines can be listed among cruise-specific *pull factors* (Crompton, 1979), i.e. "cultural motives" external to the individual that depend on destinations' specific features. Teye and Paris (2010) refer to them in terms of *convenience/ship-based* factors, which include all the amenities, facilities and services that the cruise line provides onboard as well as on the mainland.

Lastly, the *environment* image component of a cruise ship holiday refers to what Pine and Gilmore (1998) defined as *aesthetics*, which entails consumers' *passive participation and immersion* in the physical environment around them. With regard to the physical environment typical of cruise experiences, Kwortnik (2008, p. 292) used the term "shipscape" to refer to "a context-specific type of servicescape that includes both the man-made physical and social environment in which the cruise service is delivered (the ship), as well as the natural environment (the ocean)".

Research Methodology and Results

With the aim of detecting which of the eight image components of the cruise ship holiday identified in the previous section were actually addressed by the Italian TV commercials, a sample of 33 TV commercials broadcast between January 2011 and May 2015 was analysed, i.e. 11 commercials produced by Costa Cruises, 7 by MSC Cruises, 11 by Grimaldi Lines and 4 by Royal Caribbean International.

As already said, TV commercials are complex texts, which were simplified in this research by "translating" their different components into words. This way, each commercial was transformed into a sequence of words, i.e. a written text. The authors of this paper made independent "translations" of images and music into words, and finally reached an agreement on the words to be included within the written text to analyse

Latent Dirichlet Allocation (LDA) topic modeling was performed to analyse the topics included within the TV commercials. Four topics were extracted under the expectation that some sort of correspondence would be detected between the four cruise companies and the four topics. Based on which words were associated to each topic, topic 1 concerns *Relaxation with family*, which can be related to two of the eight image components of a cruise holiday, i.e. *relaxation* and *family and social relationships*; topic 2 refers to *Entertainment and escapism from ordinary life*, which can be related to the two homonymous image components; topic 3 deals with *Onboard facilities and ports of call*, which corresponds to the *services and products provided by the cruise line* image component; lastly, topic 4 deals with *Aesthetics and physical activities*, which can be related to the *environment* image component.

Table 1 reports some examples of topic contributions to cruise companies' commercials.

Costa's and MSC's promotional campaigns seem to share a multidimensional image of the ideal cruise, which emphasizes the same three topics, i.e. 1, 2, and 4.

Grimaldi Lines adopted a different promotional strategy, which consists of stressing the company's brand name and the "service dimension" (topic 3).

Royal Caribbean International is prevalently focused on topic 4.

Table 1: Examples of topic contributions to cruise companies' commercials

Commercial (text) ID	Company	Top topic		Second-top topic	
		Topic ID	Contribution to text	Topic ID	Contribution to text
1	Costa Cruises	2	0.356	3	0.289
…..	…..	…..	…..	…..	…..
13	Grimaldi Lines	3	0.640	2	0.160
…..	…..	…..	…..	…..	…..
25	MSC Cruises	1	0.568	2	0.223
…..	…..	…..	…..	…..	…..
33	Royal Caribbean	4	0.436	2	0.277

Conclusions and future developments

This study provides analysts with a feasible itinerary to investigate how automated content analysis can help recognise specific image components of cruise ship holidays through complex audio-visual texts such as TV commercials.

More research is required to gain deeper knowledge of the promotional strategies of cruise holidays. In this direction, cruise companies might monitor the images that they impart through information sources other than TV (e.g. brochures, official websites and social networks), so as to control for consistency.

In addition, the tourists' viewpoint should also be taken into account to find out to what extent images of cruise holidays spread by promotional as well as non promotional texts have influenced tourists' images of a cruise holiday, which, in turn, are able to influence the formation of specific travel expectations, finally affecting cruisers satisfaction.

Lastly, cruise markets other than the Italian could be considered.

References

1. Baloglu, S., McCleary, K.W.: A model of destination image formation. Ann. Tourism Res. \textbf{26} (4), 868--897 (1999)
2. Blei, D.M., Lafferty, J.D.: A correlated topic model of Science. Ann. App. Stat. \texbf{1} (1), 17--35 (2007)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Jour. Mach. Learn. Res.\texbf{3}, 993--1022 (2003)
4. Cartwright, R., Baird, C.: The Development and Growth of the Cruise Industry. Butterworth-Heinemann Ltd, New York (1999)
5. Crompton, J.L.: Motivations for pleasure vacation. Ann. Tourism Res. \textbf{6} (4), 408--424 (1979)
6. Dickinson, R.H.: "Fun Ship" marketing philosophy. Hosp. Rev. \texbf{13} (1), Article 1 (1995)
7. Hagen, L.: Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?, Inf. Proc. Manag. \texbf{54} (6), 1292--1307 (2018)
8. Hosany, S., Witham, M.: Dimensions of cruisers' experiences, satisfaction and intention to recommend. Working Paper Series of the School of Management, Royal Holloway University of London, Egham Hill, Egham, UK (2009)
9. Hu, N., Zhang, T., Gao, B., Bose, I.: What do hotel customers complain about? Text analysis using structural topic model. Tourism. Manage. \texbf{72}, 417--426 (2019)
10. Huang, J., Hsu, C.H.C.: The impact of customer-to-customer interaction on cruise experience and vacation satisfaction. Journal of Travel Research \texbf{49}(1), 79--92 (2010)
11. Hung, K., Petrick, J.F.: Why do you cruise? Exploring the motivations for taking cruise holidays, and the construction of a cruising motivation scale. Tourism Manage. \texbf{32}(2), 386--393 (2011)
12. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey, Multimed. Tools Appl. \texbf{78}: 15169–15211 (2019)

13. Klein, R.A.: Cruise Ship Blues: The Underside of the Cruise Industry. New Society Publishers, Gabriola Island, British Columbia, CAN (2002)
14. Kwortnik, R.J.: Carnival cruise lines: burnishing the brand. Cornell Hotel and Rest. A. \texbf{47} (3), 286--300 (2006)
15. Kwortnik, R.J.: Shipscape influence on the leisure cruise experience. *Int. J. Cult. Tourism Hospital. Res*. \texbf{2} (4), 289--311 (2008)
16. Li, X., Zhang, A., Li, C., Ouyang, J., Cai, Y.: Exploring coherent topics by topic modeling with term weighting. Inf. Proc. Manag. \texbf{54}, 1345--1358 (2018)
17. Mcauliffe, J.D., Blei D.M.: Supervised Topic Models. In: Platt, J.C., Koller D., Singer Y., Roweis S.T. (eds.) Advances in Neural Information Processing Systems 20 (NIPS 2007), pp. 121-128. Curran Associates, Inc., Red Hook (2008)
18. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. \texbf{39} (2), 103--134 (2000)
19. Pine, B.J., Gilmore, J. H.: Welcome to the experience economy. Harvard Bus. Rev. \texbf{76} (4), 97--105 (1998)
20. Polizzi, G., Oliveri, A.M.: The Image of Cruise Ship Holidays on Italian Television: A Comparative Analysis. In: Dowling, R., Weeden, C. (eds.) Cruise Ship Tourism, 2nd edition, pp. 274-289. CAB International, Wallingford, UK (2017)
21. Teye, V., Paris, C.M.: Cruise line industry and Caribbean tourism: guests' motivatio
22. Urry, J.: The Tourist Gaze. Leisure and Travel in Contemporary Societies. Sage, London (1990)
23. Wood, R.E.: Cruise Ships: Deterritorialized Destinations. In: Lumsdon, L., Page, S.J. (eds.) Tourism and Transport: Issues and Agenda for the New Millennium, pp. 133-145. Elsevier, Amsterdam (2004)

Evaluating the determinants of environmentally-significant behaviour in Europe

Le determinanti dei comportamenti sostenibili in Europa

Gennaro Punzo, Demetrio Panarello, Margherita Maria Pagliuca, and Rosalia Castellano[1]

Abstract This paper explores how citizens' perceived values and felt responsibility affect pro-environmental behaviours in the EU-28 countries, grouped according to the quartile distribution of the EU Eco-Innovation Index 2017. An SEM is tested on Eurobarometer data (2017), revealing interesting differences across the four groups which can help policy-makers in implementing well-designed strategies.

Abstract *Questo lavoro esplora come i valori percepiti e il senso di responsabilità dei cittadini influenzino i comportamenti pro-ambientali nei Paesi UE, raggruppati secondo la distribuzione quartilica dell'EU Eco-Innovation Index 2017. I risultati di un modello ad equazioni strutturali testato su dati Eurobarometro (2017) evidenziano interessanti differenze tra i quattro gruppi, utili per l'individuazione di adeguate strategie di sensibilizzazione ai comportamenti pro-ambientali.*

Key words: Pro-environmental behaviour, Perceived values, Felt responsibility, Structural equation modelling, Eco-Innovation Index

1 Introduction

The awareness that many environmental problems originate from human actions could become a keystone for driving citizens' pro-environmental behaviours (PEBs), defined as individual behaviours contributing to environmental sustainability [9,13].

[1] Gennaro Punzo and Demetrio Panarello, University of Naples Parthenope, Department of Economic and Legal Studies; email: gennaro.punzo@uniparthenope.it, demetrio.panarello@uniparthenope.it. Margherita Maria Pagliuca and Rosalia Castellano, University of Naples Parthenope, Department of Management and Quantitative Studies; email: margherita.pagliuca@uniparthenope.it, lia.castellano@uniparthenope.it.

Several theoretical frameworks have been developed to explore the determinants of PEBs [e.g., 17,19], stressing the role of psychological constructs in their prediction.

Adding to this strand of literature, our idea lies in a theoretical framework that links perceived values, felt responsibility, and PEBs. More precisely, we use values as perceived by citizens to be embodied by the EU, assuming that citizens' personal values are influenced by situational factors [15]; we refer to felt responsibility as the extent to which citizens feel compelled to take useful action towards the environment [6]; finally, we adopt a multidimensional set of environmentally-friendly practices to express PEBs. In this context, we perform a comparative analysis on all EU countries, divided into four groups, investigating how PEBs are affected by perceived values – directly, indirectly through the mediation of felt responsibility, or either way. Therefore, by using a Structural Equation Modelling approach (see Figure 1), we test the following hypotheses:

- H_1: Values (VA) positively affect Actions (AC);
- H_2: Felt Responsibility (FR) positively affects AC;
- H_3: VA positively affect FR;
- H_4: FR mediates the effect of VA on AC.

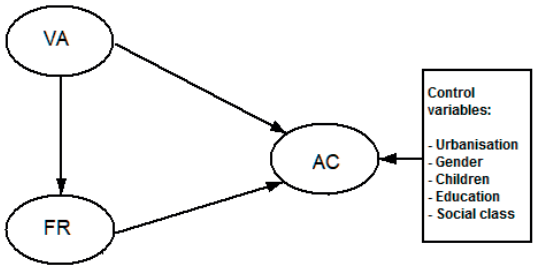


Figure 1: Conceptual model

2 Data and Methods

We use data from the Special Eurobarometer public opinion survey on the environment [5], carried out between September and October 2017 to look at EU-28 citizens' environmental behaviours, perceptions of environmental issues and opinions regarding the role of the European Union in environmental protection.

We group the EU-28 Member States based on the quartile distribution of the 2017 version of the EU Eco-Innovation Index[1], which illustrates eco-innovation performance across EU countries as captured by 16 indicators grouped into a composite index. In more detail, we consider the following groups: Low Performers (countries scoring from 38 to 73), made up by Bulgaria, Cyprus, Poland, Estonia, Hungary, Romania, and Latvia; Middle Performers (from 74 to 86), composed of Slovakia, Croatia, Greece, Lithuania, Czech Republic, Belgium, and Malta; High

[1] https://ec.europa.eu/environment/ecoap/indicators/index_en.

Performers (from 88 to 113), including Netherlands, France, Ireland, Portugal, United Kingdom, Spain, Austria, and Italy; Leaders (from 117 to 144), comprising Slovenia, Denmark, Germany, Luxembourg, Finland, and Sweden. As can be seen, Scandinavian countries, Germany and Luxembourg lead the ranking, while most eastern nations who joined the EU in 2004 and 2007 rank far below the EU average.

We consider 27 variables, which define the latent constructs of VA, FR, and AC. We present a detailed list of such variables in Table 1, along with basic descriptive information. As shown in the table, VA is made up of 7 binary items, expressing whether or not the citizen considers certain values to be embodied by the EU; FR consists of 5 items about the role of individuals in keeping the environment safe, expressed on a 4-point Likert scale (from 1 meaning "totally disagree/not at all important" to 4 meaning "totally agree/very important"); AC is composed of 15 binary items, concerning various specific environmentally-friendly actions implemented by the respondent, in which 1 means that the action is put in place by the individual, and 0 otherwise. In addition, we employ 5 socio-demographic dummy control variables that are assumed to affect pro-environmental actions.

Table 1: Latent constructs and indicators: percentages by group of countries

Construct	***Item***	***Low***	***Middle***	***High***	***Leaders***
		Percentage*			
Values embodied by the EU (VA)	Respect for nature and environment	36.0	37.1	37.0	51.6
	Social equality and solidarity	39.5	42.2	42.1	60.0
	Peace	43.7	47.0	43.3	59.0
	Progress and innovation	28.6	28.6	26.7	31.8
	Freedom of opinion	39.4	41.7	43.5	61.0
	Tolerance and openness	39.7	39.8	40.8	52.2
	Respect for history	32.1	34.9	33.9	39.1
Felt Resp. (FR)	Protection of environment: personal importance	55.9	55.2	56.8	64.9
	Environment: role of the individual	38.3	38.8	50.2	53.7
	Environment: polluters' responsibility	68.1	64.0	65.8	76.3
	Environment: worried about impact of plastic-made everyday products	50.2	45.5	46.2	54.7
	Environment: worried about impact of chemicals in everyday products	54.0	51.9	49.2	54.1
Actions (AC)	Chosen a more ecological way of travelling – last 6 months	26.6	23.8	25.1	34.8
	Avoided buying over-packaged products – last 6 months	19.4	21.2	25.1	31.1
	Avoided single-use plastic products (bags excluded) – last 6 months	26.1	27.9	34.7	46.1
	Separated most waste for recycling – last 6 months	44.5	63.0	68.2	77.1
	Cut down water consumption – last 6 months	23.0	27.5	29.6	27.1
	Cut down energy consumption – last 6 months	29.1	26.6	36.4	42.0

	Bought products marked with an environmental label – last 6 months	13.0	14.3	18.8	39.4
	Bought local products – last 6 months	46.3	45.0	42.1	54.9
	Used the car less by avoiding unnecessary trips – last 6 months	11.5	14.0	18.6	23.9
	Cut down use of plastic bags	56.0	60.6	81.0	76.8
	Replaced heating system with a low-emission one – last 2 years	10.7	12.8	13.2	17.0
	Replaced electrical appliances with high-efficiency ones – last 2 years	28.4	31.8	32.4	41.5
	Frequently used public transport or a bicycle, or chosen to walk instead of talking the car – last 2 years	34.0	34.0	34.1	43.9
	Bought an electric vehicle or a low-emission car – last 2 years	6.5	6.5	8.7	10.6
	Bought low-emission fuels for barbecue or fireplace – last 2 years	7.8	11.0	10.3	16.5
Control variables	Urbanisation (0: urban; 1: rural)	30.2	31.7	34.7	31.0
	Gender (0: female; 1: male)	39.9	43.6	47.5	48.0
	Children (0: no; 1: with children)	39.6	37.2	36.2	29.5
	Education (0: completed education at age 19 or earlier; 1: after age 19)	32.7	28.4	31.9	48.4
	Social class (0: working class or lower; 1: middle class or upper)	55.2	55.1	55.3	69.5

* Reported values refer to the value 1 for dummy variables (VA and AC constructs, control variables) and to "totally agree/very important" for Likert scale variables (FR construct).

The constructs' internal consistency and reliability is evaluated by using the Cronbach's alpha index, computed by construct and group (Table 2). Most alphas fall between the recommended range of acceptability, generally assumed to be between 0.70 and 0.90. Even though the alphas for the AC construct fall slightly below the recommended range – which may depend on the variables' binary encoding [20] – lower figures of alpha estimates do not necessarily lead to questioning the constructs' reliability [1].

Table 2: Cronbach's alpha for each construct and group of countries

Construct	***Low***	***Middle***	***High***	***Leaders***
Values (VA)	0.8546	0.8716	0.8774	0.8755
Felt Responsibility (FR)	0.7521	0.7392	0.7854	0.7779
Actions (AC)	0.6910	0.6939	0.6640	0.6493

By means of the Mplus latent variable modelling program, we perform a Structural Equation Model (SEM), employing the robust Mean- and Variance-adjusted Weighted Least Squares estimator (WLSMV) for parameter estimation. SEM assesses the relationships between latent constructs, each measured by a set of manifest variables [8]. SEM consists of a measurement model, which evaluates the relationship between observed and latent variables in terms of reliability and validity, and a structural model that determines causal relationships between latent variables,

which can simultaneously be dependent in some equations and independent in some others. We estimate SEMs separately by group of countries. Informative measures about the ability of the models to fit the data – Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) – fall inside the range of acceptability [12]: as CFI and TLI are greater than 0.90 and RMSEA is lower than 0.05 for each investigated group of countries, all the models show a high level of fit.

3 Main results

In the measurement models, most factor loadings are highly significant and have a relatively high value on their construct, implying a relationship between each observed indicator and its respective latent variable[1]. The structural models capture significant relationships among the latent constructs of VA, FR, and AC (Table 3).

Table 3: Structural model - estimates by group of countries

Path	***Low***	***Middle***	***High***	***Leaders***
			Coefficient	
VA=>FR	0.058***	0.033**	0.038***	0.052**
FR=>AC	0.331***	0.391***	0.344***	0.398***
VA=>AC (Direct effect)	0.023*	0.013	0.019**	0.058***
VA=>AC (Indirect effect)	0.019***	0.013**	0.013***	0.021**
rural=>AC	-0.085***	0.011	0.006	-0.042**
male=>AC	-0.091***	-0.079***	-0.049***	-0.126***
children=>AC	0.055***	0.054***	0.009	0.061***
education=>AC	0.220***	0.197***	0.227***	0.229***
socialclass=>AC	0.065***	0.133***	0.144***	0.110***

*** p-value < 0.01; ** p-value < 0.05; * p-value < 0.10.

H_1 hypothesis is tested to investigate the relationship of perceived values on PEBs, confirming a positive and highly significant linkage for the two more eco-innovative groups of countries. By contrast, in the two lower-performing groups, values provide less or no guidance for behaviour. This may occur when citizens exhibit a stronger feeling of belonging to the EU, which leads them to behave according to their group identity rather than to their personal characteristics [14,18]. H_2 is consistently verified: FR has a significant impact on AC for each group of countries. Therefore, people always tend to feel morally obliged to care about the environment for their own and the others' interests [10]. H_3 is as well verified for each group, pointing out a positive relationship of VA on FR. There is also a significant indirect impact of VA on AC through FR, though it is reduced in absolute size with respect to the direct effect. This mediated effect has an all-encompassing

[1] For sake of brevity, we omit the measurement models estimates, but they are available upon request.

role for Middle Performers. Specifically, the proportions of the direct effect on the total one become more relevant for the more eco-innovative groups of countries (High Performers and Leaders), controlling for 59.4 and 73.4 percent of the overall effect, respectively. However, the assumption that VA indirectly impact on AC (H_4) is supported for every group of countries, consistently with most research on the field (see [11]). As regards control variables, being woman, well-educated and from a high social class positively affects PEBs in each group of countries, while the presence of children positively affects PEBs in three groups out of four. This is in line with most literature that observes a greater concern of women – especially those with children – for the environment [16] and that identifies highly-educated, middle- and upper-middle-class individuals to be more environmentalist [7].

We perform a multi-group analysis to compare the path coefficients between the four groups of countries. The differences in the structural parameters between the groups can only be verified if the measurement model is invariant. To test measurement invariance, we compare the model (M1) in which factor loadings and thresholds are held free across groups (measurement non-invariance) with the model (M2) with all the parameters constrained to be equal across groups (measurement invariance). We evaluate the invariance models on the basis of their goodness-of-fit and of the model comparison results. In our case, there is evidence of measurement invariance, since when constrains are introduced, the model fit does not get worse than the one of the unconstrained model; indeed, the absolute differences between the fit indices estimated on the two models are not higher than 0.010 [4]. Therefore, we can investigate the differences among the three structural coefficients that link VA, FR and AC between the four groups through pairwise comparisons (Table 4).

Table 4: Differences across groups

Group comparison	***Path***	***Difference***	***Standard Error***
Low vs. Middle	VA=>FR	0.026	0.020
	FR=>AC	-0.060**	0.026
	VA=>AC	0.009	0.016
Low vs. High	VA=>FR	0.020	0.019
	FR=>AC	-0.014	0.023
	VA=>AC	0.004	0.014
Low vs. Leaders	VA=>FR	0.006	0.030
	FR=>AC	-0.068**	0.027
	VA=>AC	-0.036	0.022
Middle vs. High	VA=>FR	-0.005	0.017
	FR=>AC	0.047**	0.023
	VA=>AC	-0.005	0.013
Middle vs. Leaders	VA=>FR	-0.019	0.029
	FR=>AC	-0.007	0.027
	VA=>AC	-0.045**	0.021
High vs. Leaders	VA=>FR	-0.014	0.028
	FR=>AC	-0.054**	0.024
	VA=>AC	-0.039**	0.020

*** p-value < 0.01; ** p-value < 0.05; * p-value < 0.10.

The influence of FR on AC is stronger for eco-innovation Leaders and Middle Performers compared to Low Performers, for Middle Performers compared to High Performers, and for Leaders with respect to High Performers. Moreover, the effect of VA on AC is significantly higher for Leaders compared to Middle Performers and High Performers. Even though the effect of VA on FR is significant for each group (H_3 hypothesis), it does not seem to differ among the four groups taken two by two.

4 Conclusions

Our results show that the largest contribution to the prediction of AC comes from FR, that is, the stronger felt responsibility is for citizens, the stronger their feelings of moral obligation are to act pro-environmentally. On the other hand, the direct impact of VA on AC, with respect to the total one, is the strongest for the eco-innovation leaders, while the indirect effect of VA on AC is stronger for Low Performers and even has an all-encompassing role for Middle Performers. Therefore, values play a key role in PEBs, confirming that citizens who consider the EU institutions to be embodying them feel responsible to act on the environment in line with such values.

These results may help policymakers in implementing specific measures to make citizens act in a pro-environmental way through the feeling of responsibility, especially when their perceived values are weak. Indeed, by looking at the estimates from structural models (Table 3), felt responsibility impacts on PEBs in an extensively stronger way than values, both directly and indirectly, do.

In order to efficiently act on felt responsibility, policy should concentrate on providing well-meaning information on the effects of individual behaviours on the environment. Media could impact on environmental perceptions by spreading the EU guiding principles and thus reinforcing citizens' feeling of belonging to Europe. Moreover, economic measures, such as taxes and subsidies, can induce people to engage in PEBs. For instance, disposal fees and recycling subsidies have been found to lead individuals to increase separate waste collection [3]. To achieve the best performance, the participation of a significant portion of the population is required, and the institutions must inform citizens on why collective actions are necessary [2].

However, even when people are willing to engage in environmentally-responsible behaviour, the lack of infrastructures may make them unable to do so. For example, high prices of 'green' goods and low access to public transport in a given locality may constrain the adoption of PEBs. Environmental programs should be planned by taking the citizens' perspective into account, and the use of participatory methods could lead to them being more motivated to work for the programme's goals.

In brief, although cross-country heterogeneity within each of the four groups should also be investigated, in order to capture the different dynamics characterising each country, well-designed strategies could turn into a general increase in PEBs, aimed at reducing car use, harmful emissions or plastic products usage, increasing waste recycling, optimising water and energy usage, and orienting purchasing choices towards more environmentally-friendly goods.

References

1. Bonett, D.G., Wright, T.A.: Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. Journal of Organizational Behavior 36(1), 3-15 (2015)
2. Castellano, R., Musella, G., Punzo, G.: The effect of environmental attitudes and policies on separate waste collection: the case of Insular Italy. Letters in Spatial and Resource Sciences 12(1), 63-85 (2019)
3. Cecere, G., Mancinelli, S., Mazzanti, M.: Waste prevention and social preferences: the role of intrinsic and extrinsic motivations. Ecological Economics 107, 163-176 (2014)
4. Chen, F.F.: Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal 14(3), 464–504 (2007)
5. European Commission: Special Eurobarometer 468, Attitudes of European citizens towards the environment (2017) Avail. at http://data.europa.eu/euodp/en/data/dataset/S2156_88_1_468_ENG
6. Fuller, J.B., Marler, L.E., Hester, K.: Promoting felt responsibility for constructive change and proactive behavior: Exploring aspects of an elaborated model of work design. Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior 27(8), 1089-1120 (2006)
7. Gifford, R., Nilsson, A.: Personal and social factors that influence pro-environmental concern and behavior. A review. International Journal of Psychology 49(3), 141-157 (2014)
8. Goldberger, A.S.: A course in econometrics. Harvard University Press, Cambridge (1991)
9. Jebli, M.B., Youssef, S.B., Ozturk, I.: Testing environmental Kuznets curve hypothesis: The role of renewable and non-renewable energy consumption and trade in OECD countries. Ecological Indicators 60, 824-831 (2016)
10. Kaiser, F.G., Shimoda, T.A.: Responsibility as a predictor of ecological behaviour. Journal of Environmental Psychology 19(3), 243-253 (1999)
11. Marquart-Pyatt, S.T.: Explaining environmental activism across countries. Society & Natural Resources 25(7), 683-699 (2012)
12. Marsh, H.W., Kit-Tai, H., Zhonglin, W.: In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers. In Overgeneralizing Hu and Bentler's (1999) Findings. Structural Equation Modeling 11(3), 320-41 (2004)
13. Mesmer-Magnus, J., Viswesvaran, C., Wiernik, B.M.: The role of commitment in bridging the gap between organizational sustainability and environmental sustainability. In: Jackson, S.E., Ones, D.S., Dilchert, S. (eds.) Managing Human Resources for Environmental Sustainability, pp. 155-186. Jossey-Bass/Wiley, San Francisco (2012)
14. Punzo, G., Panarello, D., Pagliuca, M.M., Castellano, R., Aprile, M.C.: Assessing the role of perceived values and felt responsibility on pro-environmental behaviours: A comparison across four EU countries. Environmental Science & Policy 101, 311-322 (2019)
15. Steg, L., Bolderdijk, J.W., Keizer, K., Perlaviciute, G.: An integrated framework for encouraging pro-environmental behaviour: The role of values, situational factors and goals. Journal of Environmental Psychology 38, 104-115 (2014)
16. Thomas, G.O., Fisher, R., Whitmarsh, L., Milfont, T.L., Poortinga, W.: The impact of parenthood on environmental attitudes and behaviour: a longitudinal investigation of the legacy hypothesis. Population and Environment 39(3), 261-276 (2018)
17. Turaga, R.M.R., Howarth, R.B., Borsuk, M.E.: Pro-environmental behavior. Annals of the New York Academy of Sciences 1185(1), 211-224 (2010)
18. Turner, J.C., Onorato, R.S.: Social identity, personality, and the self-concept: A self-categorization perspective. In: Postmes, T., Branscombe, N.R. (eds.) Key readings in social psychology. Rediscovering social identity, pp. 315-339. Psychology Press, New York (2010)
19. Van Poeck, K., Vandenabeele, J.: Learning from sustainable development: Education in the light of public issues. Environmental Education Research 18(4), 541–552 (2012)
20. Voss, K.E., Stem, D.E., Fotopoulos, S.: A comment on the relationship between coefficient alpha and scale characteristics. Marketing Letters 11(2), 177-191 (2000)

Financial transaction data for early estimates of macroeconomic indicators for Services in Italy: value added and turnover index

I dati di pagamenti elettronici per le stime anticipate di indicatori macroeconomici dei Servizi in Italia: valore aggiunto e indice del fatturato

Alessandra Righi, Guerino Ardizzi, Alessandro Gambini, Filippo Moauro, Nazzareno Renzi

Abstract We provide first results of real-time nowcasting of quarterly value added and turnover index in Services by autoregressive distributed lag models based on innovative datasets built on electronic payment transaction data managed by Bank of Italy and on anti-money laundering aggregate reports managed by the Italian Financial Intelligence Unit. The accuracy of estimates is improved by new series with respect to the performance of both the autoregressive benchmarks and the alternative specifications based on official short-term statistics indicators.

Abstract *Sono presentati i risultati del nowcasting in tempo reale delle stime trimestrali del valore aggiunto e dell'indice del fatturato dei Servizi, realizzate con modelli ADL e nuovi indicatori basati sui dati delle transazioni elettroniche di pagamento di fonte Banca d'Italia e sui dati delle segnalazioni anti-riciclaggio aggregate dell'UIF. L'accuratezza delle stime previsive viene migliorata dalle serie dei nuovi indicatori sia rispetto ai benchmark autoregressivi sia alle alternative basate su indicatori congiunturali ufficiali.*

Key words: financial transaction data, big data, nowcasting, ADL model

[1]Alessandra Righi, ISTAT; email: righi@istat.it

Guerino Ardizzi, Bank of Italy; email: Guerino.Ardizzi@bancaditalia.it

Alessandro Gambini, Bank of Italy; email: Alessandro.Gambini@bancaditalia.it

Filippo Moauro, ISTAT; email: moauro@istat.it

Nazzareno Renzi, UIF-Banca d'Italia; email: Nazzareno.Renzi@bancaditalia.it

1 Introduction

The introduction of big data use for nowcasting and forecasting macroeconomic indicators has been recently enhanced by institutions, researchers and policy makers at the European level. The statistical office of the European Union has devoted an increasing commitment to the study of these issues testified by the organization of conferences and publication of various volumes (Baldacci et al. 2016; Kapetanios et al. 2017a, 2017b and 2018). Big data assumed a relevant role also in the Central banks due to the opportunity to get new information aimed at increasing the accuracy of forecasts at short delay (Panetta 2018). Given the fast digitalisation in retail payment systems, financial transaction data show suitable features to track the short-term economic trends and consumer confidence (Ardizzi et al. 2019). In Italy, a recent contribution by Aprigliano et al. (2017) has addressed for the first time the attention to electronic payments, since their link to traditional economic measures finds solid justifications by the economic theory.

The Italian national institute of statistics (Istat) and Bank of Italy decided to set an institutional collaboration agreement to study the potentiality of new data sources in term of forecasting power. The joint Bank of Italy-Istat working group produced datasets including information on electronic payment transactions processed by regulated payment systems and on suspicious transactions listed in the anti-money laundering aggregate reports (SARA) that financial intermediaries file monthly to the Financial intelligence unit (UIF), a separate branch of the Bank of Italy. Similar activities were conducted by the joint Istat-Bank of Italy working group in the context of the Eurostat ESSnet Big Data Project on early estimates of economic indicators. The main achievement of this project was to demonstrate that a combination of multiple big data sources, administrative and official statistical data can be usefully used in producing macroeconomic early estimates (Luomaranta et al 2018).

Our paper aims at describing the data production of the new indicators and presenting the main results of two experimental case studies where the new indicators have been used for the nowcasting of the quarterly national accounts estimates referred to the Services sector, with reference to the cases of the turnover index and the value added.

2 Data sources, data access and data editing

Wholesale payments (handling large-value transactions, more connected with financial markets flows and banks' refinancing operations with national central banks) and retail payments (mostly related to individual payment activity) are two main categories of payments in the Payment System. The second set of payment transactions is the most interesting for nowcasting and forecasting macroeconomic

indicators since we can use them to proxy economic activity in terms of consumption and real investments.

The Bank of Italy operationally manages the retail payments and, thus, has daily data availability on two payment systems: (i) BI-COMP, the clearing system which clears the domestic retail electronic payments on a multilateral net basis; (ii) the Italian component of TARGET2 (Trans-European Automated Real-Time Gross settlement Express Transfer system; T2-IT), whose retail branch (henceforth T2-retail) is used by banks to settle large amounts urgent payments on behalf of their customers. In both cases data are collected on a daily basis, while aggregated monthly series are available within few days after the end of the month; no information is available on the customer.

The retail non-cash payments settled through BI-COMP and T2-retail add up to about 60% of the total value of retail payments in Italy (see Aprigliano et al. 2017). In term of relevance, the total amount of payment flows settled through the two payment systems is about three time the Italian GDP on annual basis, and in terms of number of transactions, payment system transactions data account for 40-50% of total transactions.

As for the data access, elementary data have been aggregated in compliance with the existing General Data Protection Regulation (GDPR) in order to produce time series for our experiments. This resulted in the production of about twenty monthly time series extracted from BI-COMP and TARGET2 retail circuits (in amount and in number of transactions). The BI-COMP series, reconstructed from January 2000 to November 2017, are broken down by payment instruments (namely, credit transfers, direct debits, payment cards as POS and ATM, cheques). The TARGET2 payments series refer to the total of customer transactions and to the only cross-border transactions.

A preliminary phase of the study was devoted to the understanding of the links between macroeconomic phenomena (with their indicators) and the available financial transaction series. It was rapidly clear that the new payments series needed a greater granularity in terms of sectors of economic activity involved in the transactions to better represent the standard economic indicators. Thus, we decided to introduce in our project another big data source referred to the anti-money laundering aggregate reports (SARA) filed by financial intermediaries to the Financial intelligence unit (UIF). The database consists of anonymous data relating to all the transactions of an amount equal to or greater than EUR 15,000 made by private customers. The over 100 million yearly records, monthly collected, contain information on the bank branch where the transaction took place, on the kind of transaction (give or take) and summary information of the customers performing the transaction (among the others, the sectors of activity interested by the transaction broken down by manufacturing industry, retail trade, wholesale trade, market services, households aggregate and other). We produced about fifty monthly series referring both to bank wire transfers (national or abroad) and cash payments, broken down by type of transaction and sector of activity, made available since January 2001 (e.g., market services and trade series), or since later dates (all other series).

The data editing allowed to verify that during the period from January 2000 to December 2017 the system of payments series are fully comparable except for some

changes in the regulation of the system. The pre-treatment and inspection phase highlighted some periods starting from December 2013 showing particular perturbations for the BI-COMP series due to the migration to the Single Euro Payments Area (SEPA) and to the concurrent increased relevance of the privately owned and managed Pan-European inter-bank clearing and settlement system STEP2-T. This private payment system has initially greatly eroded the representativeness of the BI-COMP series in the processing of the new SEPA-compliant credit transfers and direct debits, but this evidence is gradually getting back.

The SARA series revealed some perturbations in the period from October 2002 to December 2003, due to problems caused by the technology used for reporting the series of cash or debit amounts. An outliers' detections phase was needed to correct these breaks in the series before starting to use them in the experiments.

3 Impact assessment of the use of financial transaction data on early estimates of macroeconomic indicators for Services sector

The Istat early estimate of quarterly GDP is officially released at 30 days by the end of reference quarter since May 2018. This is compiled using the same sources and methods adopted for the full compilation of quarterly national accounts (which is released in a second round at 60 days). The quarterly estimate is indirectly calculated by aggregation of more than 50 sub-components of value added obtained through temporal disaggregation and extrapolation methods (as genuine data sources are available only at annual frequency and after 18 months). The method is of a Chow and Lin (1971) type and makes use of short-term indicators as related information (Istat 2008). Short-term indicators involved in the estimation of GDP come from different sources, including official data (industrial production, turnover indexes, social surveys) as well as a long list of administrative data. This approach guarantees high quality standards to GDP estimate, since flash estimates show low revisions with respect to full-informative releases, even if several indicators are either totally or partially missing at the T+30 round. Missing data are estimated though a pure time series forecasting methods, rarely using alternative non-official indicators.

The lack of data afflicts less the manufacturing industry sector since a preliminary version of the industrial production index (IPI) is provided by Istat just in time for the GDP flash estimate thanks to recent improvements. A provisional version of quarterly IPI is built on the elementary information provided by respondents at around t+28 days by the end of the month. Results of a preliminary experiment over recent years have shown very low revisions of data from the preliminary round with respect to the official release at T+40 days (Istat 2015).

Conversely, no indicator is sufficiently timely for early estimates for the Services. This is why the availability of new financial transaction series is

particularly relevant to test a one-step-ahead forecasting exercise of the value added in Services.

Autoregressive distributed lag models (ADL, in particular, first-order ADL), conducted at a quarterly frequency, are used for the evaluation of the new time series. We used the following ADL(1,1) model

$$\Delta^l \Delta_4^m y_t = \phi \Delta^l \Delta_4^m y_{t-1} + \alpha + \beta_0' x_t + \beta_1' x_{t-1} + \gamma t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad t = 1, \dots, T-1 \qquad (1)$$

where:

y_t is the target variable, in our case the value added (in natural levels or logarithms);

Δ^l l = 0.1 is the simple difference operator $\Delta^l y_t = y_t - y_{t-1}$ for the $l = 1$ or absence of differentiation for $l = 0$;

Δ_4^m $m = 0.1$ is the seasonal difference operator such that $\Delta_4^m y_t = y_t - y_{t-4}$ for $m = 1$ and no seasonal difference for $m = 0$;

ϕ is the autoregressive coefficient such that $-1 < \phi < 1$;

x_t is the vector of the regressors given by one or more system of payment series to which the same differences of y_t operator is generally applied and which are considered with a lag equal to 0 or 1;

β_0 and β_1 are the regression coefficients associated respectively to x_t and x_{t-1};

α is the constant term;

γ is the coefficient associated with the linear trend t;

ε_t are the white noise residuals of the regression (assuming presence of homoscedasticity, absence of autocorrelation) whose variance is equal to σ_ε^2).

A recursive estimate referring to an increasingly longer period, starting from the period 2000Q1-2011Q4 up to the period 2000Q1-2015Q4, is used for the rolling forecasting exercise. The one-step-ahead forecasts and the forecast errors (evaluated with respect to the T+60 releases based on the full set of indicators) refer to 16 quarters over the period 2012Q1-2015Q4.

For comparative evaluation, we conducted also an exercise of pure extrapolation on data of value added in Services through a first-order autoregressive model. This provides the benchmark values to which refer to in studying the performance of the new indicators in terms of mean error (ME), mean absolute errors (MAE), and root means squared errors (RMSE). Main results show that there are several differences when using seasonally adjusted series or unadjusted data: the data specification in logarithms and the difference-order 1 prevail for seasonally adjusted data, whereas the specification in levels and the difference-order 4 for unadjusted data. Moreover, lower values of MAEs and RMSEs are observed for seasonally adjusted data.

The forecasting performance of the system of payment series for value added in Services shows that some series reduce the mean error statistics of the AR(1) benchmarks. This is true only if a careful intervention analysis for outlier removal (with Tramo-Seats software) is preliminary conducted. The reduction of the error in the nowcasting the quarterly growth rates ranges around 12%, this gain is in line with the improvement obtained using traditional short-term indicators, such as the industrial production index (Table 1).

Table 1. Comparison of mean errors (MAE and RMSE) for the forecasting (one-step forward, period 2011q4-2015q4) of value added in Services for models using different BI-COMP – TARGET2 series or IPI series as regressor

	ME liv	ME d4	MAE liv	MAE d4	RMSE liv	RMSE d4
AR(1) model	43.23	0.03	1535.43	0.57	2250.48	0.84
ADL(1,1) model with IPI series	116.13	0.06	1326.84	0.50	1975.23	0.74
ADL(1,1) model with BI-COMP - TARGET2 series	285.42	0.12	1357.61	0.51	2183.30	0.82

A similar forecasting exercise conducted over the Services turnover index gave also encouraging results; the use of T2-retail and BI-COMP series as regressors produce new estimates showing on average 7% gains of in terms of MAE with respect to the AR(1) benchmarks (Table 2). Particularly positive are the gains due to TARGET2 Total series (12%), BI-COMP direct debit series (8%) and BI-COMP debit cards series (6%).

Around 16 SARA time series by types of payment and sector are used (after being linearized with the Tramo-Seats) to generate the one-step forward forecast of the Services turnover index in the real-time exercise[1]. The gains in terms of MAE with respect to the AR(1) benchmarks are of 10.4% (Table 2). The best performers are the series referring to the bank transfer from the retail trade (21%) and the total series from market services (13%).

Table 2. Comparison of mean errors (MAE and RMSE) for the forecasting (one-step forward, period 2011q3-2017q2) of Services turnover index for models using different BI-COMP - TARGET2 series, SARA series or IPI series as regressor

	ME liv	ME d4	MAE liv	MAE d4	RMSE liv	RMSE d4
AR(1) model	0.0411	0.0782	1.1209	1.1313	1.3955	1.4108
ADL(1,1) model with IPI	-0.2585	-0.2422	0.7701	0.7716	0.9949	0.9871
ADL(1,1) model with BI-COMP - TARGET2 series						
BI-COMP Total	0.2736	0.3070	1.0522	1.0681	1.3250	1.3570
BI-COMP Direct debit	0.0299	0.0558	1.0142	1.0380	1.2781	1.3195
BI-COMP Debit cards	0.2408	0.2729	1.0631	1.0674	1.3751	1.3759
TARGET2 total	-0.1260	-0.1016	0.9771	1.0005	1.2744	1.3095
ADL(1,1) model with SARA (take) series						
Bank transfer from Wholesale trade	0.4167	0.4615	1.0008	1.0296	1.4016	1.4670
Bank transfer from Retail trade	0.5396	0.5858	0.8587	0.8927	1.3343	1.4249
Total from Wholesale trade	0.0137	0.0436	0.9771	0.9990	1.2108	1.2534
Total from Other sector	0.3666	0.4081	0.9956	1.0228	1.3707	1.4308

[1] Since SARA series are available at around T + 60, the services turnover index for the last month is not available for the quarterly estimate T + 30 and it is necessary to forecast it.

5 Conclusion

A first systematic evaluation of the set of financial transaction data coming from T2-retail, BI-COMP and SARA databases for forecasting of the Service sector indicators gave promising results. A relevant general conclusion of the study is that an accurate pre-treatment of the outliers is needed before these new series could effectively be used in nowcasting activities. This is due to the relevant perturbations showed by the series when changes in regulations or laws ruling the sector of electronic transactions intervened. These derived fluctuations in observed series are larger than those observable in standard short-term economic indicators used for official early estimates in national accounts framework.

The informative contribution that comes from these new series seems to be relevant and it can actually lead to improvements in the nowcasting of Services indicators. The exercise on value added in Services showed that the size of the reduction of the mean absolute error due to the use of the new series is in line with that obtained using in the nowcasting the most informative traditional short-term indicators (e.g., IPI) as regressors, Similar results also come from the nowcasting of Services turnover index using a combination of T2-retail and BI-COMP as well as SARA series.

Further developments of our study for the enhancement of forecasting with financial transaction data concern the extension of the experiment of the test-series on other target economic indicators and the widening of the modelling strategy, on the one side; the enrichment of the set of transaction data with the acquisition of further data flows from private providers, on the other side.

References

1. Aprigliano V., Ardizzi G., Monteforte L.: Using the payment system data to forecast the Italian GDP, Banca d'Italia, Temi di discussione (Working papers), 1098 – February (2017)
2. Ardizzi G., Emiliozzi S., Monteforte L., Marcucci J.: News and consumer card payments, Banca d'Italia, Working Papers series (forthcoming)
3. Baldacci E., Buono D., Kapetanios G., Krische S., Marcellino M., Mazzi G. L., Papailias F.: Big data and macroeconomic nowcasting: from data access to modelling, Eurostat statistical books (2016)
4. Chow, G.C., Lin A.: Best linear unbiased interpolation, distribution and extrapolation of time series by related series. Rev.Econ.Stat. 53, 372–5 (1971)
5. Istat: Quarterly national accounts inventory, sources and methods of Italian quarterly accounts (2008), available at https://ec.europa.eu/eurostat/documents/24987/4253464/IT-QNA-Inventory-ESA95.pdf/5f7d98d8-2734-448c-9c5f-75845e648bc1
6. Kapetanios G., Marcellino M., Papailias F.: Big data conversion techniques including their main features and characteristics, Eurostat statistical working papers, July (2017a)
7. Kapetanios G., Marcellino M., Papailias F.: Filtering techniques for big data based uncertainty indexes, Eurostat statistical working papers, November (2017b)
8. Kapetanios G., Marcellino M., Petrova K.: Analysis of the most recent modelling techniques for big data with particular attention to Bayesian ones, Eurostat statistical working papers (2018)
9. Luomaranta H., Puts M., Grygiel G., Righi A., Campos P., Grahonja Č., Špeh T.: Deliverable 6.6: Report about the impact of one (or more) big data source (and other) sources on economic

indicators, Work Package 6 - Early estimates of economic indicators, ESSnet Big Data specific grant agreement 2 (SGA-2), pp.36-42 (2018)
10. Panetta F.: Big data and machine learning technology for central banking, speech at the Conference Harnessing big data & machine learning technology for central banking, Roma, March (2018)

Education, Women and Empowerment: the case of India

Istruzione, donne ed empowerment: il caso indiano

Azzurra Rinaldi and Fabiana Sciarelli

Abstract Data show us that where education is poor, the basic health standards are not met, population lives on a low income, development is slow and unequal. This is the reason why we focused on the human development in India, with a particular focus on the link between the development of education and gender biases. The process of building a gender equal education system in a country is based on an Education Development Management Model structured in three main phases: the definition of objectives; the identification of the solutions; the implementation of the identified solutions. We provide a hybridized methodology between economics and management so as to allow each country to plan and manage their own education process and overcome the gender bias in education.

Abstract *I dati ci dimostrano che, nei paesi in cui vi è scarsa istruzione, gli standard sanitari di base non vengono rispettati, le popolazioni vivono con redditi bassi e lo sviluppo è lento ed iniquo. Per questo motivo, abbiamo condotto l'analisi partendo dallo sviluppo umano in India, con un focus particolare sul legame tra lo sviluppo dell'istruzione e le questioni di genere. Il processo di costruzione di un sistema di istruzione che garantisca l'equità di genere si basa su un Modello di gestione dello sviluppo dell'istruzione che si basa sulla definizione degli obiettivi, l'identificazione della soluzione e l'implementazione della soluzione identificata. Nel nostro lavoro, forniamo una metodologia ibrida tra economia e management che consente ad ogni paese di gestire l'istruzione e superare il gender bias.*

Key words: Gender Equality, Development, Education, Education Development Management Model.

1. Introduction

The correlation between education and women empowerment is well-known. Many studies testify, indeed, the extent in which participation in higher education offers

empowerment to women in different countries (Malik & Courtney, 2011; Omwami, 2015).
In our paper, the process of building a gender equal education system in a country is based on the Education Development Management Model (Sciarelli, Rinaldi, 2018) , which is structured in three main phases: the definition of objectives; the identification of the solutions; the implementation of the solutions that have been previously identified. In the first phase, we analyzed the current situation of the Indian education system. In the second phase, we tried to isolate the problems that impede to Indian women to follow a complete educational path and to develop feasible solutions and their implementation Finally, in the third phase we provided some suggestions that could be useful to fix this situation.

2. Methodology

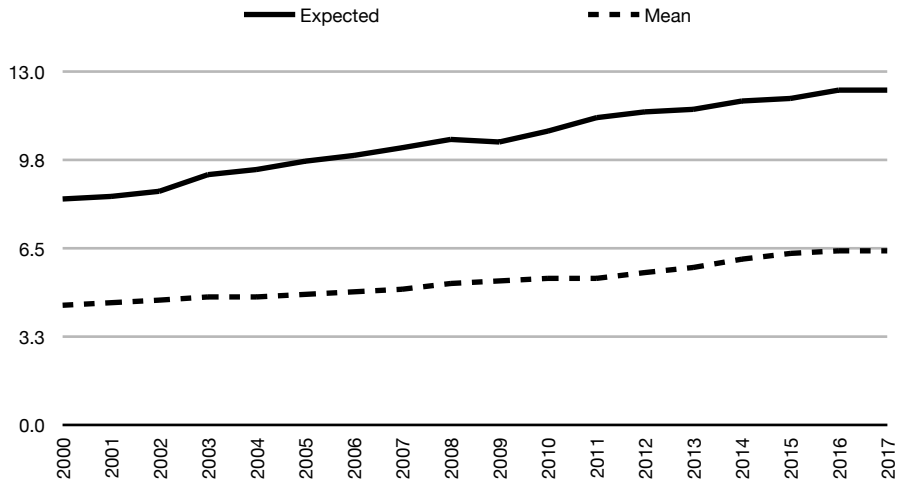
In our research, we followed different objectives. First of all, studying the school system of India and observing the existing gender gaps, also thanks to the field research we carried on. During this phase, we conducted numerous interviews with UN, UNDP and WB representatives. With this aim, we also carried on about 100 interviews with a random sample of ordinary citizens of the country. Another aim of the study has been the identification of the actors and the characteristics of the learning process that are most common in the country. These steps allowed us to define some tailor-made proposals for the development of the educational system in India. To do so, we also analyzed each and every Indian strategic plan and development program, even using the World Bank, UNICEF, etc. data.

3. Results

3.1.Education in India

Many governments focus on the objective of economic growth, rather than trying to reach an adequate social, human and sustainable development. Indeed, relevant literature shows that economic development is not sustainable in the long term if it is not supported by a general development strategy and also that education may be one of the main drivers of a long-term growth (Aghion & Howitt, 1996; Mincer, 1996; Madsen & Murtin, 2017)). Although during the last years India has shown significant values of the GDP growth rate, its development process has been very unbalanced: according to the 2018 Human Development Index, the country ranks 130th out of 189 countries in terms of human development. And, while some parts of the country have made huge socio-economic progress, other regions still lag behind, even in two crucial fields such as education and health. For example, the

percentage of women giving birth in the hospitals varies between states (99.9% in Puducherry, 32.8% in Nagaland), despite the fact that, at a national leve, it increased from 38.7% in 2005-06 to 78.9% in 2015-16. Since a sustained growth process may be accomplished only together with a concomitant transformation in the human development of every part of the country, the Indian government recently launched the Aspirational Districts Programme (ADP), along with the Twelfth Five-Year Plan. Education is presented as one of the key factors of development in both the national plans.



Indeed, it remains an important issue. As we may see in Figure 1, from 2000 both the expected and the mean years of schooling increased, yet the gap between these two indicators is increasing, too.

Figure 1: Expected vs. Mean Years of Schooling, 2000-2017

As we mentioned before, the Twelfth Plan addresses the strategies dedicated to the education sector as an important element for the development. The need for an improvement in the educational sector is also testified by two more programs: the Sarva Shiksha Abhiyan (SSA) and the Right of Children to Free and Compulsory Education Act or Right to Education Act (RTE). Both of them, indeed, have increased basic education.

3.2.Women's Empowerment through Education

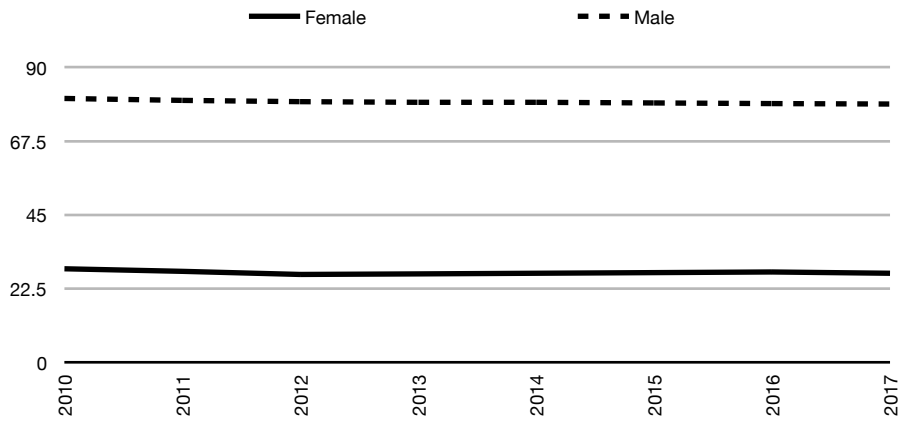
Gender gap in India is still quite high: the Gender Inequality Index is 0.524, that means that India ranks 127th at a global level. Income per capita in 2017 is 2,722 Purchasing Power Parity $ (2011) for female and 9,729 for male. Even in the field of the attained education, we observe huge differences between sexes. Indeed, while the mean years of schooling for men are 8.2, for women they are just 4.8.

Mean Years of Schooling	2010	2011	2012	2013	2014	2015	2016	2017
Female	3.6	3.9	4.2	4.5	4.8	4.8	4.8	4.8
Male	7.2	7.4	7.6	7.8	8.0	8.2	8.2	8.2

Figure 3: Mean Years of Schooling - Female and Male, 2010-2017

Unsurprisingly, the proportion of illiterate adult women is significantly higher than that of illiterate adult men. This gap also reflects on the labour market. Indeed, the labour force participation rate of women is steadily much lower with respect to the rate of men: in 2017 - the latest data available -, the labour force participation of men is 78.8, while for women it is just 27.2 (Fig. 4).

Figure 4: Labour Force Participation Rate - Female and Male, 2010-2017

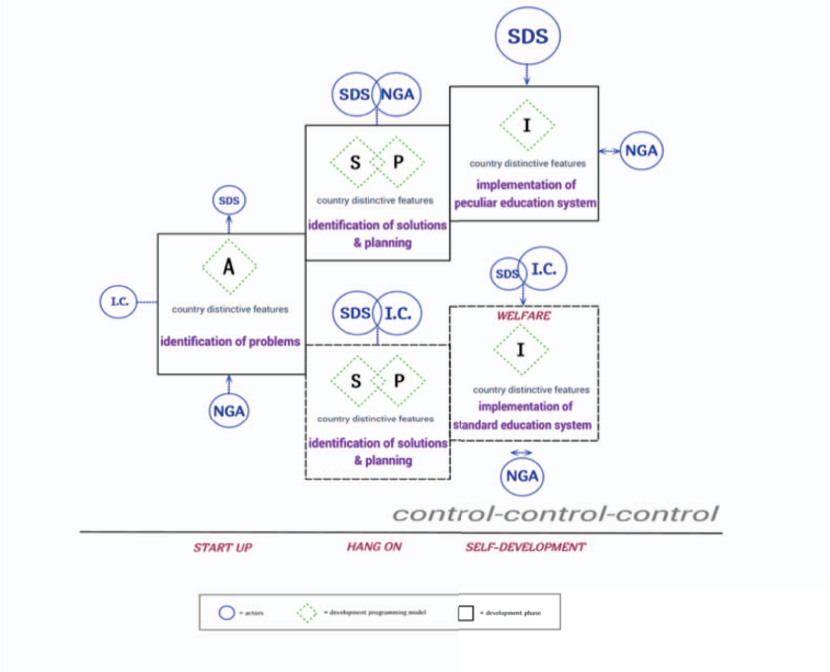


Many reasons may explain the gender gap in education (that leads to gender gap in the labour force). One of them is linked to the labor market discrimination: since employers value women's education less than men's education, the economic incentives to educate girls are lower (Kingdon 1998). Moreover, educated women tend to remain more in paid work after delivery with respect to the less skilled women because they face higher opportunity costs of leaving their jobs (Becker, 1991).) In societies in which men have higher incomes than women, having babies causes a reduction in women's paid work, since women take care of children and

use their time in unpaid care work (Becker, 1991; Nakamura & Nakamura, 1992). Educational attainment (together with other factors, such as family situation, work experience and the husband's income) can affect the employment decisions of women (Killingsworth & Heckman, 1986; Mroz, 1987; Blundell & MaCurdy, 1999).
The Indian government is trying to address this issue: in 1978, it launched the National Adult Education Program (NAEP), followed by the Rural Functional Literacy Programme (RFLP), the National Literacy Mission (NLM) and the Saakshar Bharat Abhiyan, just to mention some of the attempts to reverse the situation of Indian women within the society.

3.3. The Development Model

As we assessed before, an adequate strategy of growth should be the key element of the development process of each country. We used a hybridized methodology between Economics and Management, using tools coming from the traditional Project Management, the Logic Framework Approach and the Result Based



Management. The objective is that each country may be able to plan and manage its own education development process.

Figure 5: The Education Development Model.

The Education Development Model represents a tool box for a growth path that can be used by the governments in order to achieve development. The model is based on the hybridization of different approaches, where the strategic analysis is a pillar of development and a two-way procedure, both top-down and bottom-up. It is divided in three phases which are temporally differentiated: Start Up, Hang On, Self Development. These three phases represent three fundamental steps, namely the identification of the problems, the identification of the solutions and the planning, the implementation of peculiar (or standard) education systems.. Finally, its objective is the decreasing dependence of the national governments from international public actors (Sciarelli, Rinaldi, 2017), that characterizes the implementation of national autonomy and the capacity of countries to define their own unique education process.

4. Conclusions

Designing an education model that aims to measurable results also in terms of gender empowerment needs a great attention to the real functioning of the system. The in-depth analysis of the peculiarities of the country, as well as the design of a strategic path for the achievement of the objectives require some factors, such as: a consistent motivation; the vision of education as a primary element of every balanced and sustainable development process such as the Overall Development; the political foresight to wait for many years for the results and the sincere search for change because, as Nelson Mandela said, "Education is the most powerful weapon that can be used to change the world" and it must be used to help women.

References

1. Aghion, P., Howitt, P. J. (1996). Research and development in the growth process, *Journal of Economic Growth*, 1(1), pp. 49-73.
2. Becker, G. S. (1991). *A Treatise on the Family: Enlarged Edition, Cambridge.* MA: Harvard University Press.
3. Blundell, R. and MaCurdy, T. E. (1999) 'Labor Supply: A Review of Alternative Approaches'. In Ashenfelter, O. and Card, D. (eds) Handbook of Labor Economics, Vol 3, Amsterdam, Elsevier Science, pp. 1559–1696.
4. Killingsworth, M. R., Heckman, J. J. (1986). Female Labor Supply: A Survey. In Ashenfelter, O. C. and Layard, R. (eds) Handbook of Labor Economics, Vol 1, Amsterdam, North-Holland, pp. 103–204.
5. Kingdon G. G., (1998). Does the Labour Market Explain Lower Female Schooling in India? *Journal of Development Studies* 35(1), pp. 39–65.
6. Krueger A. B., Lindhal M. (2001). Education for growth: Why and for whom?, *Journal of Economic Literature*, 39(4), pp. 1101-1136.

7. Madsen, J.B., Murtin, F. (2017). *Journal of Economic Growth*, 22(3), pp. 229-272
8. Malik S., Courtney K. (2011). Higher education and women's empowerment in Pakistan, *Gender and Education*, 23(1), pp. 29-45.
9. Mincer, J. (1996). Economic Development, Growth of Human Capital, and the Dynamics of the Wage Structure. *Journal of Economic Growth, 1*(1), pp. 29-48.
10. Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica*, 55, pp. 765–799.
11. Nakamura, A., Nakamura, M. (1992). The Econometrics of Female Labour Supply and Children, *Econometric Reviews*, 11, pp. 1–71.
12. Omwami, E.M. (2015). Intergenerational comparison of education attainment and implications for empowerment of women in rural Kenya, Gender, Place and Culture, 22(8), pp. 1106-1123.
13. Rammohan, A., Vu, P. (2018). Gender Inequality in Education and Kinship Norms in India. *Feminist Economics*, 24(1), pp. 142-167.
14. Sciarelli F., Rinaldi A., (2017). Development Management of Transforming Economies. Theories, Approaches and Models for Overall Development; Palgrave Macmillan: London, UK.
15. Sciarelli, F., Rinaldi, A., (2018). Il macro-management per le aree deboli del mondo. Economia e politiche di gestione dello sviluppo, Milano: Franco Angeli.
16. Sciarelli F., Rinaldi A., (2018). Education Management Process in Emerging Countries, Published in INTED2018 Proceedings, Valencia 5-7 March 2018.
17. Tomboy M., Fort L., (a cura di)(2008), Girls Education in 21st Century, World Bank, http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099080014368/DID_Girls_edu.pdf
18. Thévenon O., Nabil A., Adema W., Salvi del Pero A., (2012). Effects of Reducing Gender Gaps in Education and Labour Force Participation on Economic Growth in the OECD. OECD Social, Employment and Migration Working Papers 138, Organisation for Economic Co-operation and Development (OECD).

LOCAL AUTHORITIES AND TOURIST USE OF THE TERRITORY
ENTI LOCALI E USO TURISTICO DEL TERRITORIO

Sara Sergio[1]

Abstract: The development of a territory cannot ignore the attention that institutions (and primarily local authorities) place in it. Indeed the tourism is connected to the government of the territory, to the protection of the landscape, to the protection and enchancement of the cultural heritage. The tourism development of a territory therefore requires a territorial planning policy to guarantee a more balanced use of the places and a collaboration between institutions to guarantee the construction of infrastructures and communication networks that allow a place to become a tourist destination and a source of wealth.
Abstract in Italian: Lo sviluppo turistico di un territorio non può prescindere dall'attenzione che le istituzioni locali ripongono in esso. Infatti il turismo è strettamente collegato al governo del territorio, alla tutela del paesaggio, alla tutela e alla valorizzazione dei beni culturali. Lo sviluppo turistico di un luogo richiede quindi una politica di pianificazione territoriale per garantire un uso più equilibrato dei luoghi e una collaborazione tra istituzioni per garantire la realizzazione di infrastrutture e reti di comunicazione che consentano ad un luogo di divenire meta turistica e fonte di ricchezza.

Keywords: local authorities – tourism – development – use of territory.

1. Tourist use of the territory

When we talk about tourism, natural is the reference to the territory and to the community that lives in it[2].
The territoriality of tourism is an expression of the tourism phenomenon understood from an international point of view, characterized by the need to overcome local and national boundaries, as tourism cannot be excluded from the study of the government of the territory and from the need to safeguard the cultural and landscape heritage. natural.
The tourist use of the territory is regulated by rules governing tourism and by those concerning the use of the territory[3].

---

[1] Reasercher of Administrative Law, University of Rome "Unitelma Sapienza" - Department of Legal and Economic Sciences.
[2] The close link between tourism and the territory makes it possible to involve as much as possible the public and private operators in relation to tourism initiatives that affect the territory.

Regulatory framework on the subject of tourist use of the territory is disparate, as it attracts state and regional sources with relative differentiations of competences[4], with a central role undoubtedly covered by the Tourism Code.

Use of the territory highlights «*the sociological dimension of urban planning, completing with the reference that follows the needs of the community what legally is the activity of government of the territory*»[5].

An important doctrine highlights a «renewed strength to the advantage of that idea of territory as a meeting place for all public and private interests which, without trespassing further on the outskirts of a pesticide, has characterized part of the first regionalism»[6].

In this way a new idea of the use of the territory is born, in which the Administrations merge different factors between them, from the protection and promotion of the territory to the valorization, guaranteeing a use of the territory conforming to its nature as a public resource, less and less constructible, but increasingly demanding of an adequate government: in this way an integration between urban planning and territorial planning becomes necessary[7].

The close link between tourism and the territory is analyzed not only from an economic or social point of view, but centrality is certainly entrusted to the legal profile.

This means that the link between tourism and the territory requires the satisfaction of two opposite needs: in particular, the development of tourism as a resource capable of producing wealth and protecting and conserving the territory.

It becomes so clear that any action aimed at regulating tourism in a particular place can only be accomplished by protecting the treasures of that territory[8].

Consequently, the use of tourism in the territory requires that any form of intervention on a territory must take into account the principles underlying the administrative action, namely, reasonableness, good progress and proportionality[9].

---

[3] The rules are not only those that govern the government of the territory in detail, but also those that relate to the enhancement and protection of the landscape and cultural heritage.
Relevant in the matter of the tourist use of the territory are certainly the regional urban laws, since they are the starting point for identifying the urban planning tools for the use of the territory.

[4] In other words, reference made, for example, to exclusive state jurisdiction with regard to the protection of cultural and environmental assets; to the concurrent competence of the Regions instead with regard to the government of the territory or to the exclusive regional one for what concerns the subject of tourism.

[5] M. GOLA, *Pianificazione urbanistica e attività economiche cibo e spazio urbano: urbanistica e mercati agroalimentari*, in *Riv. Giur. Ed.*, 3, 2016, p. 209 ss..

[6] M. GOLA, *Pianificazione urbanistica e attività economiche cibo e spazio urbano: urbanistica e mercati agroalimentari*, *cit.*, p. 209 ss..

[7] M. GOLA, *Pianificazione urbanistica e attività economiche cibo e spazio urbano: urbanistica e mercati agroalimentari*, *cit.*, p. 209 ss..

[8] That is, an accommodation facility cannot be built at the expense of cultural heritage or archaeological heritage.

[9] The administrative functions "which are assigned to the different levels of the order to achieve, based on the normative catalog, the development and regulation of tourism, must be exercised in accordance with the principles of administrative action. This means that both public entities called in various capacities to perform these functions and private parties that are eligible to share the exercise must be oriented to these principles ", S. VASTA, *Uso turistico del territorio*, in A. CICCHETI, M. GOLA, A. ZITO (by), *Amministrazione pubblica e mercato del turismo*, Santarcangelo di Romagna, 2012, p. 140.

2. Tourist use of the territory in the Tourism Code and local tourism systems

Combination of tourism and territory allows us to address the issue in various ways, first and foremost that of the urban planning regulations in force in a given territory in order to create accommodation facilities.

The study of tourism cannot therefore disregard the study of the territory from a legal point of view, starting from the provisions set forth in the Tourism Code[10], without neglecting the town planning and building regulations.

In particular, tourist use of the territory is taken into consideration by the articles 22 and 23 of the Tourism Code: art. 22, National circuits of excellence in support of the tourist offer and of the Italian system identifies a different notion of territory, based on thematic areas.

And indeed, that provision provides for the creation of national circuits or homogeneous thematic itineraries based on the tourist specificities of the territories: in this way the different forms of tourism are identified, from the cultural to the spa or wellness.

The standard then divides the geographical areas regardless of the territorial boundaries and instead linking the places between them - even far away geographically - but joined by similar tourist aspects[11].

Well, with the Tourism Code the territory is considered differently compared to the classical notion, identifying itself «*with places connected to each other for a thematic factor that creates elements of territorial cohesion*»[12].

In other words, the territory is considered from a thematic point of view and no longer only from a purely geographical one[13].

Thanks to the rule in question, we should be able to create methods of promoting tourism in the area, through coordinated action between local authorities and regions in order to guarantee a unitary tourism development of Italian places, regardless of regional boundaries.

The following art. 23 of the Tourism Code[14], then, establishes the methods of legal regulation of the territory from a local point of view.

---

[10] In this regard, it is important to recall that tourism - following the reform of Title V of 2001 - falls within the areas of full legislative competence of the Regions.

Nevertheless, we have seen that the legislative competence of the State in the field of tourism is not absolutely ruled out, as this matter interferes with transversal matters of State competence (civil law) or with other matters of exclusive state competence or competitor (such as tourist professions, for example).

However, this should not have justified the adoption of a Tourism Code (Legislative Decree May 23, 2011, No. 79), since the provisions contained therein could be part of the consumer code from the private law, while those concerning administrative actions assigned to the State, on a subsidiary basis, should have been the exception.

[11] Also, M. GOLA, *Pianificazione urbanistica e attività economiche cibo e spazio urbano: urbanistica e mercati agroalimentari*, *cit.*, p. 209 ss..

[12] S. VASTA, *Uso turistico del territorio*, in *Amministrazione pubblica e mercato del turismo*, *cit.*, p. 142.

[13] Limit prescribed by the art. 22 in order to identify the circuits of excellence is that they must correspond to homogeneous tourist environments or representatives of similar realities, even if they are non-contiguous territories.

[14] That article today provides that «*as part of its programming functions and to favor the integration between tourism policies and policies of territorial government and economic development, the regions shall, pursuant to Chapter V of Title II of the part I of the Consolidated Law on Local Authorities,*

First of all, it should be noted that with the entry into force of the Tourism Code, the legislator has tried to stimulate a dialogue between local authorities and private subjects in order to coordinate their activities in the promotion and tourist information.

It is from this purpose that the expression local tourism system is born, where attention is placed not so much on the organizational models to be used, but rather on goods of tourist importance to be presented in a unified manner.

In this regard, it was noted that «*the relevant legal requirement is that local legal systems also move from the identification of homogeneous territorial areas, which, as has already been observed with respect to art. 22, do not necessarily correspond to traditional geographical territorial boundaries, but adhere to a different meaning*»[15].

Local tourism system relates to the tourist portion of the territory within which different organizational structures for the promotion and tourist information can operate in a coordinated way, thus making the local tourist system an element with which to promote a place from a tourist point of view, being able to «*create an intense tourist image that is, at the same time, unitary (relative to a specific area) and variously articulated (for the cultural heritage to visit, for the itineraries to follow, for the typical products to be tasted, for the events to which take part)*»[16].

Local tourist systems then provide a conception of the territory as «*data consistent with the multiplicity of tourist offers that it produces or that it can be incentivized to produce and this also occurs regardless of the territorial boundaries of a given territory or portion of it*»[17].

From this it derives also that being different the concept of territory to which the local tourist system refers, also the administrative functions exercised *ex* art. 23 of the Tourism Code can produce effectiveness beyond the territorial boundaries and this because the activity of tourist promotion of a territory can be exercised through forms of collaboration between more public and private subjects: they are in this way to create cooperations in the exercise of administrative functions, which produce legal effects even beyond the territorial boundaries of each Administration.

The basic idea of the local tourism system is therefore the leading role of local authorities and Municipalities in the first place in the promotion of tourism in application of the principles of horizontal and vertical subsidiarity.

In other words, the local tourism system becomes a form of collaboration between public subjects and public and private subjects to carry out a specific tourism development project for a territory[18].

---

pursuant to Legislative Decree 18 August 2000, n. 267, and of the title II, chapter III, of the legislative decree 31 March 1998, n. 112, to recognize the local tourist systems referred to in this article».

[15] S. VASTA, *Uso turistico del territorio*, in *Amministrazione pubblica e mercato del turismo*, *op. cit.*, p. 143.

[16] M. MALO - A. PERINI, *Promozione, informazione, accoglienza turistica*, in *Manuale di diritto del turismo*, Torino, 2013, p. 83.

[17] S. VASTA, *Uso turistico del territorio*, in *Amministrazione pubblica e mercato del turismo*, *op. ult. cit.*, p. 143.

[18] Relevant are regional laws to local tourism systems. For example, l.r. Lombardia, 16 luglio 2007, n. 15, *Testo unico delle leggi regionali in materia di turismo*, in BURL del 19.7.2007, n. 29, S.O. n. 2, This regional law defines local tourism systems as «*the set of programs, projects and services oriented to the*

And in particular, the cooperation between several public subjects in the exercise of administrative functions requires the use of the known associative forms between local authorities, as foreseen by the TUEL, through which they can guarantee the integration between tourism policies and policies of territorial governance and economic development, as required by paragraph 3 of art. 23 of the Tourism Code, thus achieving the promotion of tourism in the territory through the regulation of territorial government.
The tourist use of the territory and the related tourist promotion activity, as foreseen by the Tourism Code, inevitably requires to take into strict account the town planning and building legislation that determines the possibility of realizing accommodation facilities in a territory or the possibility of realizing infrastructures to reach a certain place or even the possibility of creating tourist facilities, such as museums.
It is therefore necessary to regulate the use of tourism in a territory by applying urban planning which determines how a territory can also be used for tourism purposes.
And this requires careful planning of the territory by the Municipalities, which starting from the current situation are called to plan their own territory with a look to the future, having to interpret the needs of that particular place and understand the methods of tourism development[19].
The link between tourism and territorial governance is tight: only thanks to a careful territorial planning activity can a territory be guaranteed tourism development and also guarantee economic growth.
Well, since the territory and its juridical configuration require choices made by the Municipalities, it is important to observe that «*to implement the promotion and development of tourism-related interventions involving the transformation of the territory of several Municipalities and, in the case, also of several Provinces or Regions, it is necessary to carry out a planning concertation among all the administrations involved*»[20].
And this can certainly happen by making use of negotiated territorial planning and planning tools that allow the implementation of the principles of administrative action, ie, that of simplifying the activity of the p.A. as well as that of participation in order to guarantee the best tourist use of a territory.
Ultimately, the activities of transformation of the territory for tourist use must be carried out in full compliance with the provisions of the Tourism Code as well as the provisions on territorial governance (both state and regional), which at the same time are required to comply with the regulations on the protection of the environment, the landscape and cultural heritage which in a broad sense concern the

---

development of tourism in the territory and to the integrated offer of cultural, environmental and tourist attractions, including typical products of production and food and wine local».
[19] It should be noted that it would be advisable for the Municipalities to equip themselves, in addition to the classic spatial planning tool, such as the general development plan or the urban plan, also of the tourism regulatory plan, which for example foresees public and private areas defined as tourist value, indicating where accommodation facilities can be located.
[20] S. VASTA, *Uso turistico del territorio*, in *Amministrazione pubblica e mercato del turismo*, *op. ult. cit.*, p. 146.

legal regulation of the territory, an expression of its transformation, as will be seen below.

3. Concluding remarks

Tourist use of the territory - in the perspective of development and a new conception of territory - requires the involvement of public and private actors in a virtuous comparison, in order to guarantee the growth of the tourism sector in Italy.
Territorial development activity and the promotion of tourism in our country belong to the local administrations, mainly and then to private individuals, entrepreneurs, who increasingly create networks to make tourist facilities accessible and usable through the enhancement of the place where they find[21].
Therefore a coordinated activity is needed between public and private subjects that concretely achieve the set objectives, also through a territorial planning procedure that gives the right emphasis to the tourist area intended.
The main objective is to «*encourage the aggregation of local actors (entrepreneurs, administrators, employees in every sector and more) with the aim of producing territorial and communication networks around natural, cultural, gastronomic and artisan fields, to innovate and start new tourist models, such as to create conditions of territorial evolution*»[22].
Secondly, to encourage tourism it would be advisable for local authorities to make the best use of the specific resources of each territorial area: that is, it would be necessary to highlight the various types of tourism and make the most of them, even in an associated form by the Municipalities[23].
If a territory - also understood in a broad sense, not limited to municipal boundaries - enjoys the presence of numerous cultural assets, it would be necessary to aim at territorial development by enhancing cultural tourism[24].
Only in this way will Italy be able to reach a tourism development that will allow it at the same time the economic development it deserves.

[21] Administration of tourism by programs cannot be decontextualized, but must be inserted in the process of change that has characterized the p.A. both for the innovations in the organization and for the growing emergence of an idea of valorisation.
[22] F. DALLARI, *Il progetto del territorio: gli scenari turistici della sostenibilità*, in *Geografia del turismo*, F. BENCARDINO - M. PREZIOSO (by), Milano, 2007, p. 1 ss..
[23] Conversely, if a territory is known for the presence of thermal waters it is there that it is necessary to dwell attention and create a network of services, from infrastructures, to accommodation facilities that enhance this element, enhancing thermal tourism, also in conjunction with the Contiguous territorial administrations.
[24] And indeed, cultural tourism is today to be considered no longer as a segment, but rather as a large portion for the purposes of seasonal adjustment of the tourist flow and the development of the territory.
In order for the link between tourism and culture to become effective, synergies needed between public and private operators.

References:

1. Barbati C., Governo del territorio, beni culturali e autonomie: luci e ombre di un rapporto, Aedon, 2009.
2. Cicchetti A., Il vincolo "turistico-alberghiero": strumento di conservazione o trasformazione del territorio?, in Riv. Giur. Ed., 4, 2014.
3. Dallari F., Il progetto del territorio: gli scenari turistici della sostenibilità, in F. Bencardino – M. Prezioso (a cura di), Geografia del turismo, 2007.
4. De Carlo M., La valorizzazione delle destinazioni: cultura e turismo, Milano, 2008.
5. Foà S., L'accordo di programma quadro tra il Ministero per i beni culturali e la regione Piemonte, in Aedon, 2001.
6. Gola M., *Pianificazione urbanistica e attività economiche cibo e spazio urbano: urbanistica e mercati agroalimentari*, in *Riv. Giur. Ed.*, 3, 2016, p. 209 ss..
7. Grossi R., Introduzione, in Cultura & Turismo, Locomotiva del Paese, Formez PA, 2014.
8. Malo M. – Perini A., *Promozione, informazione, accoglienza turistica*, in *Manuale di diritto del turismo*, Torino, 2013, p. 83.
9. Monari P., Il turismo culturale nuovi orizzonti di sviluppo economico e sociale, in Turismo culturale: nuovi orientamenti di sviluppo economico-sociale, in www.beniculturali.it.
10. Morra G., Homo turisticus, in P. Guidicini – A. Savelli (a cura di), Il turismo in una società che cambia, Milano, 1988.
11. Renna M., Al via la concertazione in materia di beni culturali: l'accordo di programma quadro tra ministero e regione Lombardia, in Aedon, 1999.
12. Rizzi P. – Scaccheri A., Promuovere il territorio. Guida al marketing territoriale e strategie di sviluppo locale, Milano, 2006.
13. Sau A., Turismo culturale: alcune considerazioni a margine della nuove competenze del Mibact, in federalismi.it.
14. Urbani P., Riflessioni in tema di pianificazione territoriale regionale, in Riv. trim. dir. pubbl., 1986.
15. Vasta S., *Uso turistico del territorio*, in A. CICCHETI, M. GOLA, A. ZITO (a cura di), *Amministrazione pubblica e mercato del turismo*, Santarcangelo di Romagna, 2012, p. 140.

An index for crowdsourced data on multipoint scales in tourism services evaluation

Un indice per dati crowdsourced su scale multipoint nella valutazione dei servizi turistici

Venera Tomaselli and Giulio Giacomo Cantone

Abstract When statistical approaches for customer satisfaction are employed in larger digital applications, longitudinal structured Big Data are produced. Result of this 'crowd rating' is not independent from interaction between users and online platforms. Given an empirical case study, we propose our interpretation on employment of parametric and not parametric indexes for rankings' construction from data collected from rating online platforms.

Abstract *Quando i metodi statistici della customer satisfaction sono impiegati in larghi contesti digitali si producono Big Data longitudinali strutturati. Il risultato di questo "crowd rating" non è indipendente dalle interazioni tra consumatori e piattaforma. Alla luce di un caso studio, proponiamo le nostre interpretazioni sull'impiego di indici parametrici e non parametrici per la costruzione di ranking di punteggio per mezzo di dati ottenuti da piattaforme di rating.*

Key words: tourism evaluation**,** customer satisfaction, crowd rating, ranking.

1 Evaluation from crowd rating

Crowd rating is a data gathering process to collect opinions on a topic. A common application of crowd rating in tourism is for evaluation of perceived quality:

[1] Venera Tomaselli, Department of Political and Social Sciences, University of Catania, IT; email: tomavene@unict.it.
Giulio Giacomo Cantone, email: prgcan@gmail.com.

platforms and businesses relying on recommender systems (*Facebook, Amazon Group, TripAdvisor*, etc.) are common examples.

A contributing factor of enthusiasm for crowd rating is the generally low cost to achieve acquisition of large structured datasets [7]. We found four relevant reasons to adopt crowd rating for organization of people's opinion:

- to build trust in digital communities (e.g., *eBay*)
- to display to the public a massive flux of information (e.g., to rearrange Big Data into sorted rankings)
- to develop matching algorithms for recommender systems
- to lock-in and select users, as after they 'scored' a desirable reputation in a platform, it's less likely that they will leave the platform for a competitor, so as not to lose their previous 'score' [4].

We noticed an affinity with established practices in customer satisfaction [15], but we relate this methodology to historical Galton's experiment [6]. The British polymath showed that, challenged 787 totally unknown people to estimate the weight of an ox, the difference between the median of crowd's opinions and the exact value was lesser than 1%. In particular, we will develop this intuition about the employment of median for measuring people's opinion.

Table 1: Differences between experimental design of research and implemented rating systems

Controlled research design	***Implemented design on websites***
An exact value exists and it's approximate by a metric	Supplies the lack of unit of measures for features like 'taste'
The experiment has a fixed end, and until then, other's people opinion is secret	Public crowd rating websites run with no end times and no secrecy of what is trending
No competition among subjects of measurement	Enables a competition to get better positions in future rankings, or to influence recommender systems

The following features highlight structural complexity in data production in crowd rating:

- lack of exact measure: while Galton asked people to estimate *weights*, crowd rating often aims to estimate latent features like *quality* or *satisfaction*. An established method to evaluate an inter-subjective value in perceptions employs ordinal multipoint scales. We commonly observe this method in rating systems

- secrecy of opinions: While open platforms vigorously enforce secrecy on how their algorithms are 'hardcoded', their business model is still based on making public the monitored data that includes reviews and ratings
- competition: when a new technology enables to rank products under a common criterion of interrogation ('query'), to keep a ranking position online becomes a primary target for any of those products, and in particular to be in the first visualized webpage of any related *query* on search engines [17].

Adopting definition of this new frontier of tourism as a micro-social system in Jeacle and Carter [8], a quantitative evaluation of satisfaction should not ignore the following statistical biases commonly associated to non-experimental studies on public opinions:

- non-independency of observations: earlier ratings influence late ratings. Experimental studies [16] and empirical findings [9] on crowd rating suggest that, in the absence of secrecy of trends, judgements over products converge towards a strong modal class of answers ('herding'). Research on platforms *Amazon* and *Yelp* [2] confirmed the hypothesis of the existence of a social mechanism of 'herding' that ensures that earlier ratings are more likely to influence future 'popularity' of products than later ones
- survivorship bias: competition of subjects reflects competition for survival in a market [5]. By this struggle for survival, some subjects disappear from the market others show up. Not only subjects in the same *query* or list have different lifespans, but their data can be retroactively censored by platforms too. Platforms do not desire to host an inactive subject in their online rating service. This could be a misleading factor in analysis because it censors those subjects where it is more likely that 'unpopularity' and *weaknesses* will be observed. More generally, it skews the distribution of ratings into higher numerical values [13]
- frauds and optimization strategies: platforms monitor data which are voluntarily submitted. Sometimes they lack clear procedures to confirm the general *sincerity* of the submitted data. While technologies to improve *fake detection* are constantly in development[2], frauds are usually a consistent factor of skewness in reviews [14, 12] A further reflection is necessary: while a subject who actually manipulates a ranking by the submission of *fakes* may be held responsible of crime under a variety of legislation, *TripAdvisor* states that 'optimization' and anything that does not involve a 'payment' to fake a review is not against its Terms of Service[3]. We could conclude that 'asking gently' to submit a max-scored rating should be considered a legitimate strategy of optimization of reputation and awareness, but it is made clear that material incentives in exchange for max-scored ratings is inadmissible behaviour under ToS[4]. Thus, those ratings will be

[2] https://www.tripadvisor.com/TripAdvisorInsights/w3703
[3] *ibidem.*
[4] https://www.tripadvisor.com/TripAdvisorInsights/w591

subjected to censor, introducing another bias in observed results between ratings already revised and those not.

Inequalities and biases can be extremely relevant for an effective rating. In presence of these, even with high number of reviews, many standard assumptions are not a suggestible approach of time series data analysis.

2 Web-scraped Data

We sampled a list of 60 web pages of active restaurants on *TripAdvisor.com*, the first operating since October 10th, 2009. We define 'time-point' an interval of time, which marks and gives an information about the time when the recorded review was submitted When is not explicated otherwise, we adopt 'day' as standard time-point.

A restaurant with at least one recorded review at a time-point considered an 'active subject since that time-point' ('active subjects') and we expect that the restaurant was operating at least since that time-point. Restaurants are considered 'inactive subjects' until the time-point they receive the first review.

For all the restaurants, the sampling criteria were:

- addressed in the tourism city of Catania, IT
- not less than 20 reviews at August 5th, 2018, from a total of 3204 days of activity
- 'pizza' in the menu

With a web scraping script in *R* framework, we collected metadata from the reviews in the sample (N=26.888), in particular we recorded only the following variables of metadata from reviews:

- day of submission, in the range of 3204 days, as *t* timing
- uniquely associated ID of subject restaurant on *TripAdvisor*
- recorded class of review scores, within the ordinal scale of 1 through 5.

While the number of active subjects grows linearly, the number of collected reviews does not. Even taking into account survivorship bias, which obscures data from subjects active in the past but inactive at August 5th, 2018, this does not explain the difference between the two growth ratios. The maximal divergence between the two growth ratios of (i) active subjects and (ii) collected reviews is reached on Day 1513th (December 18th, 2013) of 3204 (47%), when 34 of 60 subjects (57%) were already active.

Daily relative cumulative (until the last day) frequencies of classes of scores $F(x) = n_x / N$ were stable most of the time. The modal class, indeed, was always $x = 5$, floating around a median frequency of .441. Data are consistent with results from previous studies on Italian cities on *TripAdvisor* [1]. Frequencies were stable and $x = 5$ scored almost half of total reviews, hence we supposed that weekly $x = 5$ should have been distributed around a central value of .445 (therefore $x \neq 5$ around .555).

After we aggregated daily data-points into weekly data-points by summing all the reviews with 7 days between a Sunday and its subsequent Saturday, starting from August 7th, 2011 and ending August 4th, 2018, for a total of 365 weeks in 7 years, we found the aforementioned hypothesis to be coherent with our data: ***weekly***

$x = 5$ had a geometric mean of .4372, an average of .4487, a median of .4487, with a standard deviation of .0983, confirming the stable value.

We framed the time series from the starting week (August 7th, 2011) because this is the first week that satisfies this condition: every subsequent week had at least one $x = 5$ and one $x \neq 5$ reviews. Another noteworthy property of this starting week is that at least 10 subjects were already active on that day. By doing this frame, we excluded .005 of total recorded reviews ('N') and .202 of total recorded time-points ('days').

3 Ratings estimation in a ranking system

Although the debate between mean and median as estimators of the central value of records from multipoint scales is an open controversy [11, 18], we will argue that for low amounts of classes of score, the mean must be adopted.

We noticed that, in ordinal scales, the robustness towards the extreme values of the median as estimator of central value is of no utility because the values are enclosed in a finite domain. Its lack of sensitivity towards small differences, on the other hand, is a disadvantage in cases where these differences, even the smallest, are decisive in sorting. In particular, when the estimated parameter is argument of a rank-function for a benchmark. This incongruence is exacerbated in a longitudinal context: the median as an argument of a rank-function is sensible to factors such as skewness in frequencies of classes of values, as in our case. In particular, we observe that a minimal increase of the median of scores of an item after t causes a big 'jump' or permutation of rank [3] of the item in the ranking towards the first positions. This property seems undesirable because, under ideal conditions, every permutation of ranks after t should be imputable much to a mutation of the measured performance, less to random or structural error in the model.

More specifically, the sum of amounts of $x = 4$ and $x = 5$ was always over .7 of the total, both in our data and in another study [1]. Hence, to rank subjects by the median always produces a binary classification, which is of no use for ranking purposes for the aforementioned reasons. To estimate rating of subjects, we came to conclusion that a normalized average:

$$\frac{\frac{\sum n_{x*x}}{N} - \min(\xi)}{\max(x) - \min(\xi)} \tag{1}$$

may be the viable solution when subjects are sorted in a ranking.

For those situations where we can be confident to detect strong skewedness toward the highest ("max") class of scores, we noticed that the simpler non parametric ratio:

$$f(x = \max) = \frac{n_{max(x)}}{N} \tag{2}$$

will lead into a more stable over time, ranking. We don't assume that the feature of lesser variability over time of a ranking is valuable by default in statistical analysis, but it can be in some cases.

A variant that helped further comparison of ratings between (1) and (2) for our case study reckons on adding the 4th class of score (the second highest) to the numerator of the ratio

$$f(x = 5) + f(x = 4) \tag{3}$$

4 Conclusions

One of the results of the normalization in (1) is to enclose estimations in the dominion (0,1). This is valuable for practical uses because allows to make further reflections and compute systemic differences (i.e. stability) between rankings sorted by parametric and non-parametric estimators of rating, i.e. for items rated with different scales of scores.

The issue of developing a tool to evaluate different estimators of ratings for ranking is still open. Our suggestion is to take in consideration the rigid mathematical structure of a ranking, which is a succession of natural number, where the distance among ranks is linear. Therefore, the more the estimated ratings associated to ranks fit the assumption of linearity, proceeding from $rank_{min} -> rating = 1$ through $rank_{max} -> rating = 0$, the more that estimator fits the purpose of ranking the empirical sample and, we assume, the target population.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Multi-facet Rating of Product Reviews. In: Proceedings of 31th European Conference on IR Research on Advances in Information Retrieval, pp. 461-472. (2009).
2. Bai, T., Zhao, X., He, Y., Nie, J. Y., Wen, J. R.: Characterizing and Predicting Early Reviewers for Effective Product Marketing on E-Commerce Websites. IEEE Transactions on Knowledge and Data Engineering, 30(12), 1-14 (2018).
3. Corain, L., Arboretti, R., Bonnini, S.: Ranking of multivariate populations: a permutation approach with applications, CRC Press, Boca Raton, FL (2016).
4. Dellarocas C.: Designing Reputation Systems for the Social Web. In: Masum, H., Tovey, M. (eds) The Reputation Society, pp. 3-12. MIT Press, Cambridge, MA (2011).
5. Farmer R.: Web Reputation Systems and the Real world. In: Masum, H., Tovey, M. (eds) The Reputation Society, pp. 13-24. MIT Press, Cambridge, MA (2011).
6. Galton, F.: Vox Populi, Nature, **75**, 450-451 (1907).
7. Geiger, D., Schader, M., Rosemann, M., Fielt, E.: Crowdsourcing information systems - definition, typology, and design. In: Proceeding of International Conference on Information Systems, pp. 9-11 (2012).
8. Jeacle, I., Carter, C.: In TripAdvisor we trust: Rankings, calculative regimes and abstract systems, Accounting, Organizations and Society, **36(**4/5), 293-309 (2011).

9. Lee, Y.J., Hosanagar, K., Tan, Y.: Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. Management Science, **61**(9), 2241-2258 (2015). doi:10.1287/mnsc.2014.2082
10. Lewis, J. R.: Multipoint scales: mean and median differences and observed significance levels. Int. J. Hum.-Comput. Interact. **5**, 382–392 (1993).
11. Lewis, J.R., Sauro, J.: Quantifying the user experience: Practical statistics for user research, : Morgan Kaufmann , Cambridge, MA (2016).
12. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a General Rule for Identifying Deceptive Opinion Spam. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics pp. 1566–1576, Baltimore, MD (2014).
13. Mangel, M., Samaniego, F.: Abraham Wald's work on aircraft survivability. Journal of the American Statistical Association. **79,** 259–267 (1984).
14. Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities, Proceedings of the 21st international conference on World Wide Web, 201-210 (2012).
15. Pizam, A., Shapoval, V., Ellis, T.: Customer satisfaction and its measurement in hospitality enterprises: a revisit and update. International Journal of Contemporary Hospitality Management, **28**(1), 2–35 (2016).
16. Salganik, M. J., Dodds, P. S., Watts, D. J.: Experimental study of inequality and unpredictability in an artificial cultural market. Science, **311**, 854–856 (2006).
17. Varian, H. R.: The economics of Internet search. In: Bauer, J., Latzer, (eds) Handbook on the Economics of the Internet (pp. 385–394), Edward Elgar Publishing, Cheltenham UK (2016).
18. Velleman, P. F., Wilkinson, L.: Nominal, ordinal, interval, and ratio typologies are misleading. American Statistician, **47**(1), 65–72 (1993).

Testing consistency in preference orderings

Un test sulla consistenza tra ordinamenti di preferenze

Amalia Vanacore, Maria Sole Pellegrino, Yariv N. Marmor and Emil Bashkansky

Abstract The analysis of individual preferences is widespread in many disciplines, ranging from psychology to politics. A critical issue in the analysis of preferences is assessing and testing the consistency among two or more groups of subjects ranking a set of alternatives according to some criteria. An inferential procedure for testing preference consistency across different groups of subjects has been recently proposed by the authors. In this paper the statistical power of the proposed inferential procedure is investigated via a Monte Carlo simulation under several scenarios, differing for group size, number of alternatives and systems of hypothesis.

Abstract *L'analisi delle preferenze è diffusa in molte discipline, che vanno dalla psicologia alla politica. Una fase critica di tale analisi riguarda lo studio della consistenza tra due o più gruppi di soggetti che esprimono le loro preferenze rispetto a molteplici alternative utilizzando uno o più criteri di valutazione. Una procedura inferenziale per verificare la consistenza tra ordinamenti di preferenze forniti da diversi gruppi di soggetti è stata recentemente proposta dagli autori. Questo lavoro si propone di analizzare, attraverso uno studio in simulazione Monte Carlo, la potenza statistica della suddetta procedura inferenziale in condizioni sperimentali che differiscono per dimensione del gruppo dei soggetti, numero di alternative e sistema d'ipotesi.*

Key words: Angular distance, Preference consistency, Power analysis, Monte Carlo simulation

Amalia Vanacore and Maria Sole Pellegrino
Dept. of Industrial Engineering, University of Naples "Federico II",
e-mail: amalia.vanacore@unina.it, mariasole.pellegrino@unina.it

Emil Bashkansky and Yariv N. Marmor
ORT Braude College, Karmiel, Israel e-mail: ebashkan@braude.ac.il, myariv@braude.ac.il

1 Introduction

A critical issue in the analysis of preference data is investigating the diversity among two or more groups of subjects ranking the same set of alternative objects, namely intergroup dissimilarity, or at the opposite *consistency*, stability or agreement, as commonly referred to in the specialized literature (e.g. [2, 3, 9]). From an operative point of view, the problem is to test whether different groups of subjects are ranking in a similar manner the same set of alternatives according to some criteria (e.g. liking).

Testing for preference consistency requires to define a proper metric to measure similarity (or dissimilarity) over individual preferences expressed as weak orderings (hereafter, Preference Chains, PCs) and adopt a suitable statistical procedure to test whether the degree of similarity or dissimilarity differs across groups of subjects expressing their preferences over the same set of alternatives.

The degree of similarity or dissimilarity between two or more PCs can be evaluated through proximity or distance measures. Many distance measures have been proposed in the literature and the choice of the best one is not an exact science and depends on the data and particularly on the field of application or scientific area of interest. A widespread approach uses similarity measures based on correlation coefficients for ranking data and derives distance measures via linear transformations; typical examples are Kendall τ rank correlation, Spearman rank correlation and Pearson correlation similarity. A different approach stems from multidimensional geometry, according which each PC can be represented in a multidimensional Euclidean space as a multidimensional vector so that, more consistently with the human idea of distance, the similarity/dissimilarity between any two PCs can be measured via the angular distance between the corresponding vectors; some well-known examples are cosine similarity, Cook distance and angular distance metric. An extensive listing of such measures can be found in Deza and Deza [5].

Preference consistency can be tested through the analysis of variation. According to Orloci [11], "a comparison between groups can be achieved on the basis of the average distance between groups", estimated as mean value of the distances computed over all pairs of subjects belonging to two different groups. The same approach has been adopted by Vanacore et al. [12, 13] who formulated the variation among multiple PCs as the average pairwise distance between the observed PCs and suggested a procedure for testing preference consistency based on the indicator of segregation power, already proposed by Gadrich and Bashkansky [7].

The performance of the testing procedure has not yet been studied and this paper aims to overcome this gap by investigating the statistical power of the testing procedure via a Monte Carlo simulation study under different scenarios, differing for group size, number of alternatives and systems of hypothesis.

PCs can be generated adopting probability models for ranking data, which can be broadly grouped into 4 categories: 1) *order statistics models*, which assume that the PC is determined by the ordering of alternatives' utilities; 2) *models induced from paired comparisons*, which assume that the PC is deduced from a set of $n(n-1)/2$ arbitrary paired comparison probabilities; 3) *distance-based models*, which assume

that the probability of each PC depends on its distance to a chosen modal PC, to which it is expected most of PCs are close (i.e. PCs nearer to the modal PC have a higher probability of occurrence); and 4) *multistage models*, which decompose the ranking process into a sequence of $n-1$ decision stages whose probabilities depend only on the stage and not on the alternatives to rank at that stage. An overview of models for ranking data can be found in Diaconis [6] and Critchlow et al. [4].
In our study, different levels of preference consistency are simulated adopting the framework of distance-based models developed by Diaconis [6].

The remainder of this paper is organized as follows: in Section 2 the variation measures are introduced and the testing procedure is briefly described; Section 3 is devoted to the design and the main results of Monte Carlo simulation; in Section 4 a real case study aimed at illustrating the applicability and usefulness of the proposed approach is described; finally, conclusions are summarized in Section 5.

2 An inferential procedure to test preference consistency

A PC (i.e. $A_2 > A_1 > A_4 > \cdots > A_n$) over a set of n alternatives can be summarized in a pairwise comparison matrix of dimension $n \times n$: the generic element of the (i, j) cell is coded as "1" if the alternative A_i precedes the alternative A_j; viceversa, it is coded "-1"; "0" is used when A_i and A_j are indistinguishable alternatives. The pairwise comparison matrix, in turn, can be represented as a multidimensional vector in Euclidean space with a space dimension n^2. Dissimilarity between any two such vectors, say $\boldsymbol{a}$ and $\boldsymbol{b}$, can be measured via the normalized pairwise angular distance metric $L(\boldsymbol{a}, \boldsymbol{b})$ [12] expressed as follows:

$$L(\overrightarrow{a}, \overrightarrow{b}) = \frac{\hat{\theta}_{\overrightarrow{a}, \overrightarrow{b}}}{\pi} \frac{\arccos\left(\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}\right)}{\pi} \tag{1}$$

The normalized pairwise angular distance metric ranges between 0 and 1, where 0 means adjacent vectors (i.e. no distance and thus identical PCs), whereas 1 means opposite vectors (i.e. maximum distance and thus opposite PCs).

The PCs provided by M subjects can be arranged into a matrix of distances $\boldsymbol{L}$ (see Table 1) of dimension M $\times$ M where the generic element L_{ij} is the distance between the PCs provided by subjects i and j.

According to the definitions provided by Gadrich and Bashkansky [7] and Bashkasky et al. [1], Vanacore et al. [12, 13] estimated the variation of PCs simultaneously provided by M subjects by averaging the corresponding pairwise distances (computed according to Eq. 1) as follows:

$$V = \frac{1}{\mathrm{M}^2} \cdot \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} L_{ij} \tag{2}$$

The same dispersion measure can be adopted for assessing the total variation among the PCs provided by K groups of subjects prioritizing the same set of n

Table 1: M × M pairwise distance matrix $\boldsymbol{L}$ between M preference chains

	PC 1	⋯	**PC j**	⋯	**PC M**	***Row Sum***
PC 1	L_{11}	⋯	L_{1j}	⋯	$L_{1\mathrm{M}}$	$\sum_{j=1}^{\mathrm{M}} L_{1j}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
PC i	L_{i1}	⋯	L_{ij}	⋯	$L_{i\mathrm{M}}$	$\sum_{j=1}^{\mathrm{M}} L_{ij}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
PC M	$L_{\mathrm{M}1}$	⋯	$L_{\mathrm{M}j}$	⋯	L_{MM}	$\sum_{j=1}^{\mathrm{M}} L_{\mathrm{M}j}$

alternatives. In such case, the whole set of PCs can be arranged into K matrices of dimension M×M (one for each group, whose variation V_k is assessed via Eq. 2) or in a single matrix of dimension MK×MK, and the total variation over all K groups (V_{TOT}) can be assessed by averaging the corresponding pairwise distances as reported in the following Eq. 3:

$$V_{\mathrm{TOT}} = \frac{1}{(\mathrm{MK})^2} \cdot \sum_{i=1}^{\mathrm{MK}} \sum_{j=1}^{\mathrm{MK}} L_{ij}, \quad df_{\mathrm{TOT}} = \mathrm{KM} - 1 \tag{3}$$

According to the analysis of variation [7, 8], V_{TOT} can be split into variation within group V_{WG} with df_{WG} degrees of freedom and variation between groups V_{BG} with df_{BG} degrees of freedom, respectively formulated as:

$$V_{\mathrm{WG}} = \sum_{k=1}^{\mathrm{K}} V_k/\mathrm{K}; \qquad df_{\mathrm{WG}} = \mathrm{K} \cdot (\mathrm{M} - 1) \tag{4}$$

$$V_{\mathrm{BG}} = V_{\mathrm{TOT}} - V_{\mathrm{WG}}; \qquad df_{\mathrm{BG}} = \mathrm{K} - 1 \tag{5}$$

If all PCs come from the same population, it can be assumed that there is the same portion of expected variation per every degree of freedom:

$$\frac{E[V_{\mathrm{TOT}}]}{df_{\mathrm{TOT}}} = \frac{E[V_{\mathrm{WG}}]}{df_{\mathrm{WG}}} = \frac{E[V_{\mathrm{BG}}]}{df_{\mathrm{BG}}} \tag{6}$$

This means that the consistency can be tested through the Indicator of Segregation Power I_{SP}, defined as the ratio of the normalized V_{BG} component (Eq. 5) to the normalized V_{TOT} (Eq. 3; [7]):

$$I_{SP} = \frac{V_{\mathrm{BG}}/df_{\mathrm{BG}}}{V_{\mathrm{TOT}}/df_{\mathrm{TOT}}} \tag{7}$$

When the prioritization process significantly differs across groups the I_{SP} significantly diverges from 1.

3 Simulation study

In order to simulate different levels of preference consistency, the PCs are generated using the distance-based models developed by Diaconis [6]. The probability of each PC is controlled by a dispersion parameter λ: $\lambda = 0$ means uniform distribution of PCs, whereas larger values of λ mean that the PCs distribution is more concentrated around the modal PC. Let λ_k be the dispersion parameter for the generic k^{th} group of subjects, by varying it across $k = 1, 2, \cdots$, K groups, different samples of PCs can be simulated: higher differences across λ_k produce a lower consistency across the PCs provided by different groups of subjects.

The case of K = 3 has been considered and 36 different scenarios have been simulated. The simulated scenarios differ for group size (M = 10, 20), number of alternatives (n = 5, 10) and system of hypothesis. Specifically, the dispersion parameters λ_k have been chosen in the range $[1 \div 30]$ and 9 combinations of different values of λ_k have been considered so as to represent as many alternative hypotheses ($\Lambda_1, \cdots, \Lambda_9$ in Table 2) of decreasing consistency levels across groups (i.e. increasing heterogeneity across groups). Since lower levels of consistency can be obtained by increasing the differences across λ_k, the simulated alternative hypotheses of heterogeneity have been defined by choosing combinations of λ_k values characterized by increasing values of maximum difference ($\Delta\lambda$).

For each scenario, $r = 2000$ data sets of PCs have been generated. The simulation algorithm has been implemented using Mathematica (Version 11.0, Wolfram Research, Inc., Champaign, IL, USA).

Table 2: Values of λ identifying distance-based models under alternative hypotheses $\Lambda_1, ..., \Lambda_9$

	Λ_1	Λ_2	Λ_3	Λ_4	Λ_5	Λ_6	Λ_7	Λ_8	Λ_9
λ	(1,2,4)	(1,5,6)	(1,4,8)	(1,5,10)	(5,10,18)	(5,10,20)	(1,5,20)	(5,10,30)	(1,5,30)
$\Delta\lambda$	3	5	7	9	13	15	19	25	29

The critical value I_{cr} of the indicator I_{SP} is the $(1-\alpha)$ percentile of the empirical distribution of I_{SP} built under the assumption of homogeneity. For $\alpha = 0.05$, the values of I_{cr} for 4 different null hypotheses ($\lambda_1 = \lambda_2 = \lambda_3 = \boldsymbol{\lambda}_0$) are reported in Table 3, whereas the power curves are plotted in Figure 1.

Simulation results highlight that in the case of a medium number of alternatives, like $n = 5$, the testing procedure is adequately powered even for very small group size; in scenarios with a fairly large set of alternatives, like $n = 10$, statistical power worsens but it can be improved by increasing the group dimension.

Table 3: I_{cr} of the I_{SP} for the significance level $1-\alpha = 0.95$ considering K = 3 groups of M subjects prioritizing n alternatives for different values $\boldsymbol{\lambda}_0$

	$n=5$		$n=10$	
$\lambda_1=\lambda_2=\lambda_3=\boldsymbol{\lambda}_0$	**M = 10**	**M = 20**	**M = 10**	**M = 20**
$\lambda_1=\lambda_2=\lambda_3=\mathbf{1}$	1.53	1.53	1.28	1.28
$\lambda_1=\lambda_2=\lambda_3=\mathbf{5}$	1.52	1.53	1.28	1.28
$\lambda_1=\lambda_2=\lambda_3=\mathbf{10}$	1.54	1.56	1.28	1.28

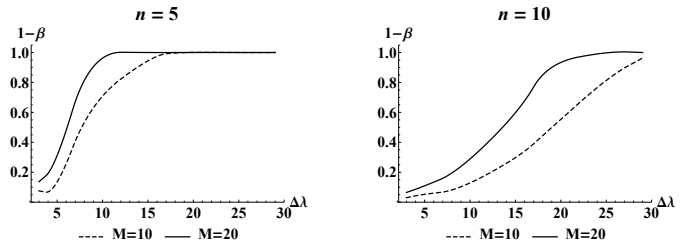


Fig. 1: Statistical power curves when testing alternative hypotheses $\Lambda_1,...,\Lambda_9$ for different values of n and M

4 An illustrative example

The testing procedure for preference consistency proposed in [12, 13] is fully exploited through the application to a real data set containing full rankings given by 5000 Japanese consumers over alternative types of sushi [10]. Each sushi is associated with seven features (i.e. style, major and minor group, heaviness, consumption frequency, normalized price and sell frequency) whereas each consumer is represented by a set of features namely gender, age, geographical and regional information.

In order to emulate the scenarios analyzed in our Monte Carlo simulation study, the preference consistency is tested for two consumer features (i.e. consumer living region and age) across K = 3 groups of M consumers ranking $n = 10$ alternatives of sushi. The 3 regions under investigation are middle, northwestern and southern-central part of Japan's main island Honshu; the 3 age classes are 20-29, 30-39 and 40-49 years old. For each group, the normalized pairwise angular distances L_{ij} for every pair of the M PCs are calculated according to Eq. 1 as well as all the dispersion measures described in Section 2.

The first step of our analysis consists in testing preference consistency across all consumers living in the Honshu Island grouped either by living region or age class. The results, reported in Table 4, reveal that, except for consumers living in the middle region for whom the age is not a significant distinguishing factor, there

is always evidence for rejecting the null hypothesis of homogeneity since the I_{SP} exceeds the critical value.

A further analysis has been conducted on groups stratified by gender. The results reveal that the significance/not significance of the whole group of both male and female consumers is generally confirmed also in the sub-stratified groups; there are only two exceptions: the region is not a significant distinguishing factor for male of 20-29 years old and for female of 40-49 years old. Moreover, the factor age appears a more segregating factor for male consumers, vice-versa the region appears a more segregating factor for female consumers.

Table 4: I_{SP} across subjects stratified either by region (on the left) or by age (on the right), for both male and female (All) consumers and for sub-stratified groups into male (M) and female (F) consumers

	Region | Age			**Age | Region**		
	Age=20-29	*Age=30-39*	*Age=40-49*	*Reg.=South*	*Reg.=Middle*	*Reg.=North*
All	1.761	2.022	1.558	2.454	**1.077**	2.902
M	**1.202**	1.312	1.361	2.234	**1.159**	2.376
F	1.739	1.798	**1.188**	1.343	**0.891**	1.365

The preference consistency test has been replicated on 1000 samples of M = 10, 20 consumers, randomly drawn from the sushi dataset. The percentages of samples for which there is evidence for rejecting the null hypothesis of preference consistency are assessed and reported in Table 5, for groups stratified by region of each age class (on the left) or stratified by age for each living region (on the right).

Table 5: Percentage of samples for which the null hypothesis of homogeneity is rejected. The samples are stratified by region (on the left) and by age (on the right)

	Region | Age			**Age | Region**		
M	*Age=20-29*	*Age=30-39*	*Age=40-49*	*Reg.=South*	*Reg.=Middle*	*Reg.=North*
10	15.6%	12.7 %	**25.6 %**	14.9 %	**43.3 %**	9.2 %
20	32.1%	23.7 %	**58.5 %**	28.4 %	**84.4 %**	15.1 %

The study reveals that in groups of M $\leq$ 20 consumers the region is less often a significant distinguishing factor in determining consumer preferences about sushi; the highest frequency of significant results is obtained for consumers living in the middle region of Honshu Island grouped by age; the age class for which the region is more frequently a significant distinguishing factor is 40-49 years old. Specifically, when testing preference consistency across groups of 10 consumers of 40-49 years old grouped by region, the evidence for rejecting the null hypothesis of homogeneity is obtained for the 25.6% of samples; for groups of different ages living in the

middle part of the Honshu Island, instead, the null hypothesis of homogeneity can be rejected for 43.3% of samples. The statistical significance for rejecting the null hypothesis of homogeneity approximately doubles for groups of 20 consumers.

5 Conclusions

The research work aims at investigating the statistical properties of a procedure to test preference consistency across multiple groups of subjects prioritizing the same set of alternatives. The Monte Carlo simulation study reveals that the statistical power of the testing procedure worsens as the subject number decreases and the dimension of the set of alternatives increases. It is adequately powered when small (i.e. $M \geq 10$) groups of subjects rank as few as $n = 5$ alternatives; otherwise, in scenarios with a fairly large set of alternatives, like $n = 10$, it can be improved by increasing the group dimension. Further developments are going to investigate the statistical behaviour of the testing procedure with larger group dimensions and/or different number of alternatives to rank in order to provide useful guidelines for its adoption.
The critical values of the Indicator of Segregation Power I_{SP} under the null hypothesis of homogeneous preferences across groups have been obtained and applied to test the homogeneity in sushi preferences with respect to demographic and geographical consumer features.

Although we deal with consistency across different groups of subjects (i.e. consistency as cohesiveness), the same procedure can be applied for test for consistency over time of the same group of subjects prioritizing the same alternatives during different sessions over time (i.e. consistency as stability over time).

References

1. Bashkansky, E., Gadrich, T., Kuselman, I.: Interlaboratory comparison of test results of an ordinal or nominal binary property: analysis of variation. Accreditation and Quality Assurance **17**(3), 239–243 (2012)
2. Brouwer, R.: Constructed preference stability: a test–retest. Journal of Environmental Economics and Policy **1**(1), 70–84 (2012)
3. Chuang, Y., Schechter, L.: Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. Journal of Development Economics **117**, 151–170 (2015)
4. Critchlow, D.E., Fligner, M.A., Verducci, J.S.: Probability models on rankings. Journal of mathematical psychology **35**(3), 294–318 (1991)
5. Deza, M.M., Deza, E.: Encyclopedia of distances. In: Encyclopedia of Distances, pp. 1–583. Springer (2009)
6. Diaconis, P.: Group representations in probability and statistics. Lecture notes-monograph series **11**, i–192 (1988)
7. Gadrich, T., Bashkansky, E.: Ordanova: analysis of ordinal variation. Journal of Statistical Planning and Inference **142**(12), 3174–3188 (2012)
8. Gadrich, T., Bashkansky, E., Zitikis, R.: Assessing variation: a unifying approach for all scales of measurement. Quality & Quantity **49**(3), 1145–1167 (2015)

9. González-Artega, T., de Andrés Calle, R., Peral, M.: Preference stability along time: the time cohesiveness measure. Progress in Artificial Intelligence **6**(3), 235–244 (2017)
10. Kamishima, T.: Nantonac collaborative filtering: recommendation based on order responses. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 583–588. ACM (2003)
11. Orloci, L.: An agglomerative method for classification of plant communities. The Journal of Ecology pp. 193–206 (1967)
12. Vanacore, A., Marmor, Y.N., Bashkansky, E.: Some metrological aspects of preferences expressed by prioritization of alternatives. Measurement **135**, 520–526 (2019)
13. Vanacore, A., Pellegrino, M.S., Marmor, Y.N., Bashkansky, E.: Analysis of consumer preferences expressed by prioritization chains. Quality and Reliability Engineering International **35**(5), 1424–1435 (2019)

A simulation study to investigate the paradoxical behaviour of inter-rater agreement coefficients in non-asymptotic conditions

Studio in simulazione del comportamento paradossale dei coefficienti di accordo inter-valutatore in condizioni non asintotiche

Amalia Vanacore and Maria Sole Pellegrino

Abstract The ability to measure how closely raters agree when providing subjective evaluations is a need common to many fields of research. The assessment and characterization of the extent of rater agreement is generally obtained through a two-steps procedure: firstly, the degree of agreement is assessed through a κ-type coefficient, then the extent of agreement is characterized by comparing the value of the adopted coefficient against a benchmark scale. Two main criticisms are often overlooked in agreement studies: 1) some κ-type coefficients depend on the frequency distribution of ratings over classification categories; 2) the straightforward benchmarking procedure neglects the influence of experimental conditions on the estimated coefficient. The robustness of two inferential benchmarking procedures adopted for different κ-type coefficients has been investigated via a Monte Carlo simulation study. Several scenarios have been analyzed, differing for sample size, rating scale dimension, number of raters, frequency distribution of ratings and pattern of agreement among raters.

Abstract *Misurare il livello di accordo nelle valutazioni soggettive fornite da un gruppo di valutatori è un'esigenza comune a diversi ambiti di ricerca. La valutazione e caratterizzazione del grado di accordo avviene generalmente attraverso due step: il grado di accordo viene stimato attraverso coefficienti di tipo κ e successivamente caratterizzato attraverso un confronto diretto con valori di riferimento. Molti studi in cui viene valutato l'accordo tra valutatori non tengono conto di due criticità: alcuni coefficienti κ dipendono dalle distribuzioni di frequenza marginali delle valutazioni; la procedura di caratterizzazione con confronto diretto non tiene conto delle condizioni sperimentali. Uno studio in simulazione Monte Carlo è stato condotto per analizzare la robustezza di due procedure inferenziali per caratteriz-*

Amalia Vanacore
Dept. of Industrial Engineering, University of Naples "Federico II", e-mail: amalia.vanacore@unina.it
Maria Sole Pellegrino
Dept. of Industrial Engineering, University of Naples "Federico II", e-mail: maria-sole.pellegrino@unina.it

zare il livello di accordo tra più valutatori. Lo studio è stato condotto per diversi coefficienti κ. Gli scenari di simulazione analizzati differiscono per: dimensione campionaria, numero di livelli della scala di valutazione, numero di valutatori, distribuzione di frequenza delle valutazioni e grado di accordo tra i valutatori.

Key words: Inter-rater agreement, κ-type coefficients, Paradoxical behaviour

1 Introduction

In many contexts of research (e.g. medical, behavioural, social, industrial) evaluation processes rely on small groups of human raters, either field experts (e.g. clinicians, visual inspectors, physicians) or untrained operators (e.g. consumers), who are asked to evaluate a set of items (or generally subjects) according to some properties (e.g. faulty classification by defect type) and/or perception aspects (e.g. quality, comfort, pain).
When providing subjective evaluations, the raters act as measurement instruments thereby, because of the common premise that only reliable raters can provide fair evaluations, the assessment of rater reliability gains importance.

Measurement System Analysis (MSA) procedures (ISO 5725) [1, 2] estimate the system reliability as its ability to provide both accurate and consistent results, where accurate means that repeated measurements are close to the true value, whereas consistent means that measurements repeated under the same conditions are close to each other. Actually, by definition, subjective evaluations miss a true reference value so that it is not possible to assess their inter-rater accuracy and thus subjective evaluations can be evaluated only in terms of their consistency.

Inter-rater agreement is generally assumed as a suitable proxy of consistency: the more raters agree on the evaluations they provide, the more comfortable we can be that their evaluations are consistent and thus trustworthy [12]. The κ-type coefficients are widespead measures of inter-rater agreement, suitable when two or more raters provide their evaluations on a categorical rating scale [3, 11].
Despite their popularity, some κ-type coefficients have been long criticized because of their dependency on the marginal frequency distribution of the ratings over classification categories. This behaviour, defined *paradoxical* by Feinstein and Cicchetti [7, 8], makes κ-type coefficients misrepresent the level of agreement in the case of skewed frequency distributions. Specifically, when the frequency distribution is balanced (i.e. the ratings are well distributed across the scale categories), the coefficients agree with one another; but, when the frequency distribution becomes more skewed, the coefficients diverge and it is unclear what they are truly measuring.
A second criticism regards the interpretation of the agreement coefficients, generally accomplished by a straightforward characterization procedure that does not account for the influence of experimental conditions.

This research work aims at investigating, via a Monte Carlo simulation study, the robustness of two non-parametric benchmarking procedures for the characterization

of the extent of inter-rater agreement estimated using three different κ-type coefficients: Fleiss' K [9], s^* coefficient proposed by Marasini et al. [17] and Gwet's AC_2 [10].

The remainder of this paper is organized as follows: in Section 2 the κ-type coefficients are introduced; Section 3 is devoted to benchmarking procedures for the characterization of the extent of agreement; in Section 4 the coefficients together with the benchmarking procedures are applied to a real data set; in Section 5 the Monte Carlo simulation is described and the main results are discussed; finally, conclusions are summarized in Section 6.

2 κ-type agreement coefficients for ordinal data

The κ-type coefficients are rescaled measures of the observed proportion of agreement corrected with the proportion of agreement expected by chance:

$$\kappa_w = \frac{p_{a_w} - p_{a|c_w}}{1 - p_{a|c_w}} \tag{1}$$

The κ-type coefficients share the same formulation of observed agreement but differ in the notion of agreement expected by chance. Let n be the number of items rated by R raters on an ordinal rating scale with $k > 2$ categories; the data, denoted Y_{lr}, with l indexing items and r indexing raters, can be arranged into a $n \times k$ table $(r_{li})_{n\times k}$, where the generic (l,i) cell contains the number of raters r_{li} who classify item l into category i. Being $R(R-1)/2$ the total number of pairs of raters, the weighted proportion of agreement for the generic item l is given by:

$$p_{a_{w|l}} = \frac{\sum_{i=1}^{k} r_{li}\left(\sum_{j=1}^{k} r_{lj} w_{ij} - 1\right)}{R(R-1)} \tag{2}$$

where w_{ij} is the symmetrical weight associated to each pair (i,j) of ratings (with $i,j = 1,...,k$) introduced in order to account that on ordinal rating scale disagreement on two distant categories are more relevant than disagreement on neighbouring categories. The symmetric (i.e. $w_{ij} = w_{ji}$) agreeing weights here adopted are the linear weights [6] formulated as follows:

$$w_{ij}^{L} = 1 - \frac{|i-j|}{k-1} \tag{3}$$

The overall weighted proportion of agreement, p_{a_w}, can be estimated by averaging the values $p_{a_{w|l}}$ over all n items:

$$p_{a_w} = \frac{1}{n} \cdot \sum_{l=1}^{n} p_{a_{w|l}} \tag{4}$$

The weighted proportion of agreement among multiple raters expected by chance defined by Fleiss [9], Marasini et al. [17] and Gwet [10] are respectively given by:

$$p^{K}_{a|c_w} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} r_i r_j \tag{5}$$

$$p^{s^*}_{a|c_w} = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} \tag{6}$$

$$p^{AC_2}_{a|c_w} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{w_{ij}}{k(k-1)} \cdot \sum_{i=1}^{k} r_i(1-r_i) \tag{7}$$

where r_i is the estimate of the probability of classifying an item into i^{th} category and is given by:

$$r_i = \frac{\sum_{l=1}^{n} r_{li}/R}{n} \tag{8}$$

Specifically, Fleiss defines the probability of chance agreement as the probability that any pair of raters classify an item into the same category under the assumption of independent classifications [9]; Marasini et al. [17] assume that the probability of chance classification is uniform across categories; Gwet [10], instead, defines the probability of agreement by chance as the probability of the simultaneous occurrence that one rater provides a random rating and agreement among raters.

3 Benchmarking procedures

The magnitude of agreement coefficient is commonly related to the notion of extent of agreement by a straightforward comparison of the estimated coefficient against a benchmark scale. Among the several benchmark scales proposed in the literature over the years, the one here adopted is that proposed by Landis and Koch [16] consisting of six ranges of values corresponding to as many categories of agreement: Poor, Slight, Fair, Moderate, Substantial and Almost perfect agreement for coefficient values ranging between -1 and 0, 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively.

Although commonly adopted, the straightforward characterization can be misleading for two main reasons: it fails to consider that an agreement coefficient, as any other sampling estimate, is exposed to sampling uncertainty; moreover, it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions (i.e. number of rated items, number of categories or distribution of items over categories).

In order to account for sampling uncertainty, a proper characterization of the extent of agreement should be based upon an inferential procedure that allows to identify a suitable neighbourhood of the truth (i.e. the true value of agreement). The

easiest and most common procedure is comparing against a benchmark scale the lower bound of a confidence interval (CI) [15].

Under asymptotic conditions, κ-type coefficients are normally distributed [11] and the parametric CI can be suitable adopted, but under the most realistic scenarios of non-asymptotic conditions, bootstrap CIs are recommended [5, 14, 20, 21]. The widespread adoption of percentile bootstrap (hereafter, p) CI is due to its low computational complexity; however, for severly skewed distribution, the Bias-Corrected and Accelerated bootstrap (hereafter, BCa) CI is recommended. Given a $(1-2\alpha)\%$ confidence level, the lower bounds of the two sided p CI and BCa CI are respectively given by:

$$\mathrm{LB}_p = G^{-1}(\alpha); \qquad \mathrm{LB}_{\mathrm{BCa}} = G^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 - a(z_\alpha - b)}\right)\right) \tag{9}$$

where G is the bootstrap distribution function for κ, Φ the standard Gaussian distribution function, z_α the α percentile of the standard normal distribution, b the bias correction parameter and a the acceleration parameter.

4 Motivating example

In order to show the implication of the paradoxical behavior of the κ-type coefficients as well as the usefulness of the described benchmarking procedure, a real case study is hereafter presented.

The analyzed data set, one of the most common in agreement studies, was originally published by Holmquist et al. [13] and then analyzed by many other authors [16, 4, 19]. It contains the classifications of $n = 118$ images/slides into $k = 5$ ordinal categories (i.e. (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma in situ, (4) squamous carcinoma with early stromal invasion, (5) invasive carcinoma) of uterine cervix carcinoma made by $R = 7$ independent pathologists on the basis of the dimension and type of lesions. In order to emulate non-asymptotic conditions, the original data set has been dropped so as to retain only the classifications made by $R = 5$ pathologists about $n = 10$ randomly chosen images/slides.

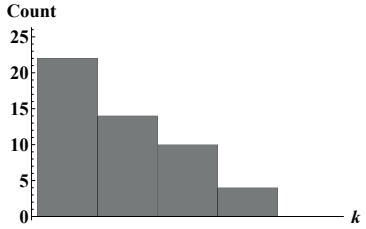


Fig. 1: Marginal distribution of pathologists' classifications

The marginal distribution of pathologists' classifications are reported in Fig. 1: the unbalanced (positively skewed) marginals represent the condition for occurring paradoxical behaviour in κ-type coefficients, so it is expected that the coefficients diverge to each other.

The inter-rater agreement is assessed via the three analyzed κ-type coefficients. For comparison purpose, the extent of agreement is characterized by benchmarking the lower bound of both p and BCa CIs against the agreement categories of Landis and Koch benchmark scale (Fig. 2). Although the observed agreement, computed according to Eq. 4, is equal to 0.835, as foreseen the skewness of the marginal distribution of classifications makes the proportions of agreement expected by chance diverge and thus the coefficients of inter-rater agreement differ from each other, being $K = 0.367$, $AC_2 = 0.662$ and $s^* = 0.587$.

According to the non-parametric benchmarking procedure (Fig. 2), the extent of inter-rater agreement can be characterized as Moderate for s^* and AC_2 and Slight or Fair (depending on the adopted bootstrap benchmarking procedure) for K.

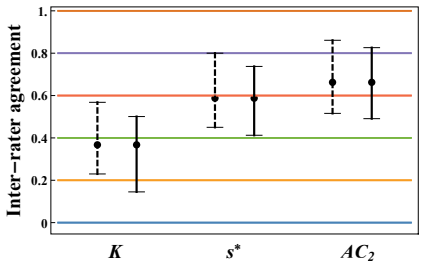


Fig. 2: Coefficients' estimates, BCa (dashed line) and p (solid line) CI obtained for all agreement coefficients ($1 - 2\alpha = 0.95$)

5 Simulation

The benchmarking procedure can be defined robust if it gives roughly the same results for a fixed level of agreement among raters whatever the frequency distribution (FD) of ratings over classification categories. The robustness of the proposed bootstrap benchmarking procedures has been investigated via a Monte Carlo simulation study, developed in accordance to the framework suggested by Quarfoot and Levine [18]. The simulated data sets are the classifications simultaneously provided by R raters on the same set of n items adopting a k-point ordinal rating scale, considering that the generic i^{th} rater provided her/his own classifications according to a given FD and all the other $R-1$ raters agree with her/him according to a given pattern of agreement among raters (i.e. Agreement Distribution, AD).

The simulation study has been designed taking into account five multi-level factors: k, n, R, FD and AD. Specifically, FDs are all special cases of the beta-binomial dis-

tribution with different values of the shape parameters (a,b): FD 1 (0.25, 0.25), FD 2 (1, 1), FD 3 (2, 2), FD 4 (50, 50), FD 5 (25, 50), FD 6 (5, 50). AD is a binomial distribution scaled on k and centered on the evaluation of the generic i^{th} rater. Considering 3 levels for R (5, 10, 20), 2 levels for n (5, 10), 3 levels for k (3, 5, 7), 6 levels for FD and 1 level for AD, the simulation design counts $3 \times 2 \times 3 \times 6 \times 1 = 108$ different scenarios, and for each scenario, $r = 1000$ Monte Carlo data sets have been generated and for each Monte Carlo data set the bootstrap CIs have been built on 1000 bootstrap runs. The robustness of the benchmarking procedures has been evaluated in terms of the mean agreement range $\overline{\text{AR}}$ over the 6 different FDs; the simulation results are represented in Figure 3. The simulation algorithm has been implemented using Mathematica (Version 11.0, Wolfram Research, Inc., Champaign, IL, USA).

The results highlight that the robustness of the bootstrap benchmarking procedures improves with increasing sample size and number of raters. Moreover, depending on differences across FDs, the extent of rater agreement may be characterized as belonging to at most 2 adjacent categories (e.g. either Substantial or Almost perfect) for s^* and AC_2 and even from 2 to 3 categories (e.g. from Fair to Substantial) when benchmarking the lower bound of the BCa CI for K.

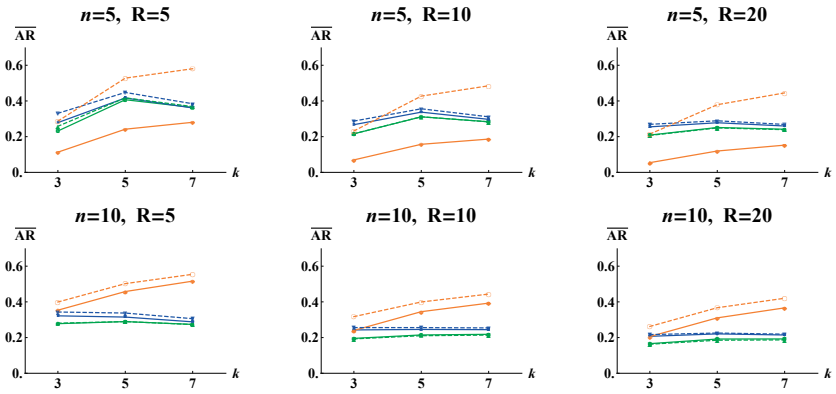


Fig. 3: Mean Agreement Range of LB_p (solid line) and LB_{BCa} (dashed line) obtained for 3 agreement coefficients (Orange: K, Green: s^*, Blue: AC_2) for different combinations of k, n and R

6 Conclusions

The robustness of two inferential benchmarking procedures for characterizing the extent of agreement with either small or moderate sample sizes has been investigated via a Monte Carlo simulation. Our findings suggest that both the number of raters and the sample size affect the robustness of the benchmarking procedures. Moreover, in presence of agreement, the benchmarking procedures for K are strongly influenced by the frequency distributions of items across categories, whereas bench-

marking procedures for s^* and AC_2 are more robust to changes in frequency distributions.

It is worthy to read the obtained results in light of their practical implication. Indeed, the degree of inter-rater agreement is characterized as belonging at most into adjacent categories (e.g. from Substantial to Almost perfect) when adopting s^* and AC_2 and as belonging to categories of 3-steps apart (e.g. from Slight to Substantial), when adopting K, especially when benchmarking the lower bound of the BCa CI.

Our results confirm the paradoxical behaviour of K which is difficult to trust in the presence of agreement among raters, despite its popularity among practitioners; conversely s^* and AC_2 seem to represent a hopeful step forward in this regard since they perform quite consistently across various frequency distributions, being less influenced by paradoxical behaviour: their use should be recommended among practitioners in order to improve robustness of agreement studies.

References

1. International Organization for Standardization (ISO) (1994). Accuracy (Trueness and Precision) of Measurement Methods and Results - Part 1: General Principles and Definitions (5725-1). Geneva, Switzerland: ISO
2. International Organization for Standardization (ISO) (1994). Accuracy (Trueness and Precision) of Measurement Methods and Results - Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method (5725-2). Geneva, Switzerland: ISO
3. Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond kappa: A review of interrater agreement measures. Canadian journal of statistics **27**(1), 3–23 (1999)
4. Becker, M.P., Agresti, A.: Log-linear modelling of pairwise interobserver agreement on a categorical scale. Statistics in medicine **11**(1), 101–114 (1992)
5. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in medicine **19**(9), 1141–1164 (2000)
6. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. American Journal of EEG Technology **11**(3), 101–110 (1971)
7. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. Journal of clinical epidemiology **43**(6), 551–558 (1990)
8. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. The problems of two paradoxes. Journal of clinical epidemiology **43**(6), 543–549 (1990)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378–382 (1971)
10. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology **61**(1), 29–48 (2008)
11. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
12. Hayes, A.F.: Statistical methods for communication science. Routledge (2009)
13. Holmquist, N., McMahan, C., Williams, O.: Variability in classification of carcinoma in situ of the uterine cervix. Archives of Pathology **84**(4), 334–345 (1967)
14. Klar, N., Lipsitz, S.R., Parzen, M., Leong, T.: An exact bootstrap confidence interval for κ in small samples. Journal of the Royal Statistical Society: Series D (The Statistician) **51**(4), 467–478 (2002)
15. Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M., Streiner, D.L.: Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International journal of nursing studies **48**(6), 661–671 (2011)

16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
17. Marasini, D., Quatto, P., Ripamonti, E.: Assessing the inter-rater agreement for ordinal data through weighted indexes. Statistical methods in medical research **25**(6), 2611–2633 (2016)
18. Quarfoot, D., Levine, R.A.: How robust are multirater interrater reliability indices to changes in frequency distribution? The American Statistician **70**(4), 373–384 (2016)
19. Saraçbaşi, T.: Agreement models for multiraters. Turkish Journal of Medical Sciences **41**(5), 939–944 (2011)
20. Vanacore, A., Pellegrino, M.S.: Benchmarking rater agreement: Probabilistic versus deterministic approach. SERIES ON ADVANCES IN MATHEMATICS FOR APPLIED SCIENCES **89**(1), 365–374 (2019)
21. Vanacore, A., Pellegrino, M.S.: Inferring rater agreement with ordinal classification. New Statistical Developments in Data Science pp. 91–101 (2019)

The effect of noise factors in experimental studies on aircraft comfort

L'effetto dei fattori di disturbo sulla valutazione sperimentale del comfort aereo

Amalia Vanacore[1] and Chiara Percuoco[1]

Abstract This paper describes a strategy adopted for the analysis of aircraft seat comfort data collected in laboratory experiments. A crossover study was planned to investigate whether the noise factors related to inter-individual variability and timing impact on seating comfort perceptions. The data analysis strategy is based on cumulative link mixed models (CLMMs). The results confirm the necessity to control the noise factors in order to obtain a diagnostic assessment.

Abstract *In questo lavoro è descritta una strategia per l'analisi delle valutazioni sul comfort dei sedili aerei raccolte mediante esperimenti in laboratorio È stato pianificato uno studio di tipo crossover per indagare se i fattori di disturbo, ovvero la variabilità dei passeggeri e il tempo, influenzano la percezione del comfort di seduta aereo. La strategia di analisi proposta si basa sui cumulative link mixed models (CLMM). I risultati confermano la necessità di controllare i fattori di disturbo al fine di ottenere una valutazione diagnostica del comfort.*

Key words: seat comfort assessment, crossover design, cumulative link mixed models

1 Introduction

Commercial aviation is the most global of businesses: it is a growth market with more than 60% growth over the last ten years. Since 1990, both aircraft movements and the number of destinations have doubled. For the next 20 years, Airbus GMF [4]

[1] Amalia Vanacore and Chiara Percuoco
Dept. of Industrial Engineering, University of Naples "Federico II", e-mail: amalia.vanacore@unina.it, chiara.percuoco@unina.it

forecasts a 4.4% global annual air traffic growth. Despite this impressive growing demand, airlines are still one of the lowest-scoring industries in the American Customer Satisfaction Index [1] with the aircraft seat rated as the most unsatisfying aspect of flying. In this context, the improvement of aircraft seat comfort can provide a concrete opportunity for airlines to improve passenger satisfaction and loyalty and thus gain competitive edge in aircraft industry [15].

Experimental studies are the easiest way to learn more about comfort experience and collect diagnostic information to be used for improving product design.
Generally, laboratory experiments for comfort studies [8, 12, 15, 16] involve potential users (*i.e.* participants) to compare different products (*e.g.* aircraft seats and/or cabin interiors) in a simulated environment (*e.g.* equipped room or fuselage). The main advantages of laboratory experiments, compared to survey studies, are that the experimenters can control the environment under which the participants make their evaluations, and moreover, they allow to learn more about comfort experience with a significant reduction in costs and time for data collection [13, 14].

On the other hand, comfort data collected in laboratory experiments may be affected by two well-known noise factors: inter-individual variability and time. The inter-individual variability (*e.g.* anthropometric characteristics, individual history and state of mind) make participants experience different levels of comfort (or discomfort) in identical environments [9, 16]; time may impact on comfort evaluations not only in terms of exposure duration (*i.e.* experiment length) but also in terms of timing, indeed different patterns of discomfort may be experienced by the participants in different days of the week.
In order to characterize and quantify the effects of the above noise factors on the evaluations about aircraft seat comfort, a laboratory experiment was conducted involving a selected group of aircraft passengers. The experimental trials were planned according to a crossover design: each participant evaluated different aircraft seats with the aim of comparing them on a within-subject basis rather than on the group level [10]. Since comfort evaluations could not be assumed independent, collected data were analysed via a cumulative link mixed model (CLMM) [2, 3, 6].

The paper is organised as follows: in Section 2, an overview of the experiment is provided; the data analysis strategy is introduced in Sections 3; the experimental results are reported and discussed in Section 4; finally, conclusions are summarized in Section 5.

2 Overview of experiment

A total of 17 volunteers (8 females and 9 males; aged between 24 and 44 years) were selected to participate in the aircraft seat comfort experiment using the following criteria:

(1) to be free from severe musculoskeletal disorders in the last year;
(2) to have taken at least 2 flights in the last year;
(3) to be economy class flyers.

Procedures for participant recruitment and data collection were defined taking into

account ethical considerations. Before providing the informed consent, participants (hereafter, assessors) were briefed about the type, the number and the duration (40 minutes) of each comfort trial, as well as on the research aims and the treatment of the collected data. The selected group of assessors was representative of the anthropometric variability in the Italian adult population with respect to both weight and height (Table 1).

Table 1. Main anthropometric characteristics of the participants

	Num.	*Age [year] [min-max]*	*Weight [kg] [min-max]*	*Height [m] [min-max]*	*BMI [kg/m^2] [min-max]*
Males	9	[27.0-41.0]	[73.1-101.8]	[1.60-1.90]	[22.8-34.7]
Mean (SD)		34.6 (4.25)	88 (8.53)	1.77 (0.08)	28.03 (3.46)
Females	8	[26-44]	[55.5-75]	[1.55-1.73]	[21.2-27.5]
Mean (SD)		33.9 (5.9)	66 (5.41)	1.66 (0.05)	24.1 (2.08)

The comfort experiments took place in a laboratory environment equipped with two rows of double-seats for regional aircrafts. The assessors sat in the second row with a pitch fixed on 32 inches in order to realistically replicate legroom.

Each assessor evaluated, in different order, the comfort of 3 typical seats (hereafter identified as A, B and C). The seats A and B were baseline configurations whereas seat C was a lightweight seat. The three seats differed from each other in terms of weight, reclining, headrest and dimensions of seat pan and backrest. All the seats were designed for economy class regional aircraft market. Each assessor tested the 3 seats following a crossover design [10] built to investigate 3 main noise factors: the day of the week, the order of testing and the inter-individual variability.

Specifically, 18 test sequences were defined using a 3×3 Greek Latin square with 6 replications; the test sequences were randomly assigned to the assessors. The design was uniform within sequences and periods. Since only 17 assessors entered the comfort experiment, there was a slight imbalance in the crossover [11].

In order to avoid the sensory biases caused by the residual sensations of previously tested seat (*i.e.* carryover effects) a wash-out period of 72 hours was fixed.

In each comfort trial, a trained interviewer asked the assessor to rate the overall comfort perception with the seat as well as the comfort perception related to specific attributes of the seat pan (*i.e.* padding and comfort) and the backrest (*i.e.* padding and support).

3 Data analysis strategy

The analysis of data obtained from the aircraft seat comfort experiment aimed at explaining the variation in subjective responses in terms of treatments (*i.e.* seats), periods (*i.e.* days of the week) and testing order.

Comfort responses collected from the same assessor (*i.e.* replications or repeated measures) are likely to be more similar on average than responses provided by different assessors, thus they cannot be assumed independent.

Cumulative link mixed models (CLMMs) are a powerful and flexible approach to handling replicated ordinal responses [6]. The main features of this approach will be briefly described in the following.

Let y_{is} denote the response provided by subject i for object s; let $x_{1is},\ldots, x_{kis}$ denote the values of the k explanatory variables and let u_i be the random effect for subject i. For response categories $j = 1, 2, \ldots, c-1$, the cumulative logit model with a random intercept is

$$\log it\left[P\left(Y_{is} \le j\right)\right] = u_i + \alpha_j - \beta_1 x_{1is} - \ldots - \beta_k x_{kis} \qquad (0)$$

The model in (0) takes the linear predictor from the marginal model and adds a random effect u_i to the cut-point term α_j. It uses the same random effect for each cumulative probability. Using an overall intercept term of form $u_i + \alpha_j$, the CLMM allows the ordinal scale cut-points to vary across subjects and thus it is a way of accounting for subjectivity in evaluations.

The random effect u_i is unobserved, so its value is unknown. It is usually assumed to vary from subject to subject according to a normal $N\ (0, \sigma_u^2)$ distribution. The variance component σ_u^2 is estimated together with the fixed effects [2, 3].

Likelihood ratio (LR) tests can be used to test fixed-effects model terms in the same way for cumulative link mixed models as in cumulative link models. A LR test of the random-effect term is a bit more complicated. Being the random effect standard deviation non-negative, the test is one-sided. The usual asymptotic theory for the LR statistic dictates that the LR asymptotically follows a χ^2 distribution with one degree of freedom. However, since the σ_u is on the boundary of the parameter space, the usual asymptotic theory does not hold. The LR more closely follows an equal mixture of χ^2-distributions with zero degrees of freedom (a point mass distribution) and one degree of freedom.

For this reason, it is often argued that a more correct interpretation is obtained from the adjusted p-value obtained by halving the p-value produced by the conventional LR test. Wald tests of the variance parameter can also be constructed, but since the profile log-likelihood function is only approximately quadratic, when σ_u^2 is not small and well defined, such tests cannot be recommended [6].

4 Results

Assuming the seat, the period and the testing order as fixed effects and the assessor as a random effect, the CLMM was fitted on the collected comfort responses (i.e overall seat comfort, seat-pan padding, seat-pan comfort, backrest padding and backrest support) using the ordinal package available in R [7].

For each fitted model, the estimates of the cut-point terms αj (j = 1, 2), the fixed effect parameters and their asymptotic standard errors are listed in Table 2 together with the p values obtained for Wald test and LR test.

Table 2. Cumulative logit mixed model fitted to aircraft seat comfort data

		Estimate	*St. error*	*Wald test p-value*	*LR test p-value*
Overall Comfort	α_1	-4.03	1.13		
	α_2	-1.63	0.90		
	Seat B	-1.25	0.95	0.18	**0.02**
	Seat C	-2.82	1.20	**0.01**	
	Order 2	0.90	0.99	0.36	0.26
	Order 3	1.73	1.14	0.13	
	Period 2	0.59	1.03	0.56	0.72
	Period 3	-0.18	0.74	0.80	
Seat-pan padding	α_1	-2.19	1.05		
	α_2	2.83	0.74		
	Seat B	0.54	0.78	0.49	0.80
	Seat C	0.26	0.88	0.76	
	Order 2	-0.53	0.89	0.55	0.38
	Order 3	0.61	0.92	0.50	
	Period 2	-0.14	1.03	0.88	0.78
	Period 3	0.26	0.88	0.48	
Backrest padding	α_1	-1.37	0.78		
	α_2	1.32	0.78		
	Seat B	0.57	0.69	0.40	0.26
	Seat C	-0.69	0.77	0.36	
	Order 2	-0.13	0.77	0.86	0.73
	Order 3	0.41	0.77	0.60	
	Period 2	-0.03	0.85	0.96	0.39
	Period 3	-0.82	0.66	0.21	
Backrest support	α_1	-0.61	0.85		
	α_2	2.08	0.94		
	Seat B	0.20	0.74	0.78	0.46
	Seat C	-0.81	0.83	0.32	
	Order 2	-0.78	0.86	0.57	0.60
	Order 3	0.27	0.82	0.74	
	Period 2	-0.47	0.97	0.62	0.79
	Period 3	-0.46	0.71	0.51	
Seat-pan comfort	α_1	-1.23	0.75		
	α_2	1.65	0.78		
	Seat B	-0.18	0.64	0.79	0.96
	Seat C	-0.16	0.77	0.83	
	Order 2	-1.48	0.82	**0.07**	**0.09**
	Order 3	-0.37	0.78	0.63	
	Period 2	0.05	0.86	0.95	0.93
	Period 3	-0.18	0.64	0.76	

The results in Table 2 highlight that the probability of obtaining a low rating for the overall comfort is significantly higher for seat C than for seats A and B (β_{seatC} = -2.82; *p-value* 0.02); under no circumstances the period (*i.e.* the day of the week) impacts significantly on comfort assessments, on the contrary the testing order impacts significantly on seat pan comfort evaluations (β_{order}= -1.48; *p-value* 0.09).

The estimates of the variance of the random effect reported in Table 3 suggest that the assessor effect is negligible for the responses about overall seat comfort; the same conclusion does not hold for the specific seat comfort attributes, since the results reflect a significant within-subject correlation for backrest support and seat pan padding and a slight correlation for backrest padding and the seat pan comfort.

Table 3. Estimated variance of assessor effect and p-values of LR tests

Comfort attribute	$\hat{\sigma}_u^2$	***LR test p-value***	***LR test adjusted p-value***
Overall comfort	<0.0001	1	0.5
Backrest padding	0.98	0.16	0.08
Backrest support	2.20	0.03	0.015
Seat-pan padding	5.56	0.0005	0.00025
Seat-pan comfort	0.26	0.63	0.32

The results of the LR test in Table 3 show that the assessor effect is significant in all models except for the overall comfort and the seat pan comfort.

The assessor effects, u_i, are not parameters, so they cannot be estimated in the standard way, but a "*best guess*" is provided by the conditional modes. Similarly the conditional variance provides an uncertainty measure of the conditional modes.

The assessor effects given by conditional modes with 95% confidence intervals based on conditional variance are plotted in Figure 1 and Figure 2 for specific comfort attributes related to the seat pan and the backrest, respectively.

Assessors #2 and #17 provided the lowest comfort ratings, whereas Assessors #1 and #16 generally gave the highest comfort ratings. The significant assessor effect indicates that assessors perceived the seat comfort differently. Two natural interpretations are that either the same comfort rating means different things to different assessors, or the assessors actually perceived the seat comfort differently. Possibly both effects play their part.

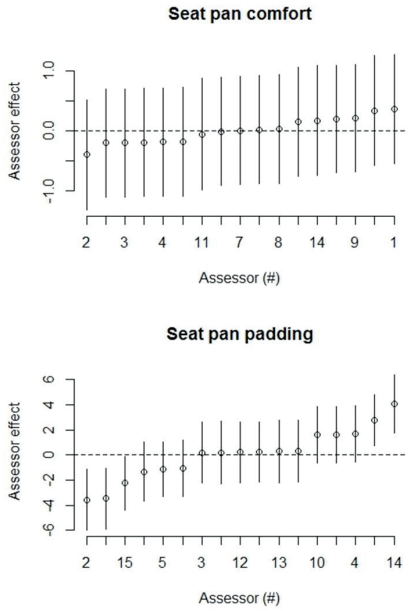


Figure 1. Assessor effect on seat pan attributes given by conditional modes with 95% confidence intervals based on the conditional variance

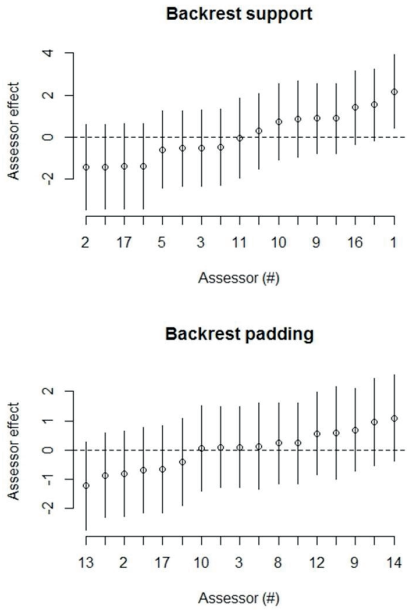


Figure 2. Assessor effect on backrest attributes given by conditional modes with 95% confidence intervals based on the conditional variance

5 Conclusions

The results of our study confirm that comfort experiments involving human assessors are prone to sensory biases due to experimental conditions and inter-individual variability. Well-designed experiments and proper data analysis strategies are thus required in order to obtain reliable information to be used for comfort improvement.

A well-designed experiment allows to control the main noise factors (*e.g.* period and testing order) causing sensory biases; on the other hand, a proper data analysis strategy allows to account for realistic violations of classical assumptions (*i.e.* normality and independence).

The results highlight that the assessor effect resulted negligible for the overall comfort and the seat pan comfort but it necessary to take this effect into account when dealing with subjective comfort perceptions related to specific seat features (*i.e.* seat pan padding, backrest support and backrest padding).

However, since psychological and physiological biases generally affect the subjective assessment in a sample set, assessor effect cannot be disregarded *a priori*.

Further investigations are necessary in order to check the generalizability of our findings outside the laboratory setting.

References

1. ACSI LLC: ACSI Travel Report. https://www. theacsi.org/news-and-resources/customer-satisfaction-reports/reports-2018/acsi- travel-report-2018, Accessed date: 7th March 2019.
2. Agresti A.: Analysis of ordinal categorical data, John Wiley & Sons, vol. 656. (2010).
3. Agresti A., Natarajan R.: Modeling clustered ordered categorical data: A survey. International Statistical Review, 69(3), 345-371 (2001).
4. AIRBUS S.A.S.: Global Market Forecast 2017–2036, "Growing horizons". Reference D14029465, Issue 4 April, 2017 (2017) (http://www.airbus.com/aircraft/market/global-market-forecast.html. Accessed September 9, 2019).
5. Bazley C., Nugent R., Vink P.: Patterns of discomfort. Journal of Ergonomics, 5(1), 1-7 (2015).
6. Christensen, R. H. B., Brockhoff, P. B.: Analysis of sensory ratings data with cumulative link models. Journal de la Societe Francaise de Statistique & Revue de Statistique Appliquee, 154(3), 58-79 (2013).
7. Christensen, R. H. B.: ordinal—regression models for ordinal data. R package version, 28, (2015).
8. Hiemstra-van Mastrigt, S., Meyenborg, I., Hoogenhout, M.: The influence of activities and duration on comfort and discomfort development in time of aircraft passengers. Work, 54(4), 955-961 (2016).
9. Hiemstra-van Mastrigt S., Groenesteijn L., Vink P., Kuijt-Evers L. F.: Predicting passenger seat comfort and discomfort on the basis of human, context and seat characteristics: a literature review. Ergonomics, 60(7), 889-911 (2017).
10. Jones B., Kenward M. G.: Design and analysis of cross-over trials. CRC press (2014).

11. Jones, B., Wang, J.: The analysis of repeated measurements in sensory and consumer studies. Food quality and preference, 11(1-2), 35-41 (2000).
12. Kremser, F., Guenzkofer, F., Sedlmeier, C., Sabbah, O. and Bengler, K.: Aircraft seating comfort: the in-fluence of seat pitch on passengers' well-being. Work, 41(Supplement 1), pp.4936-4942 (2012).
13. Molenbroek, J. F. M., Albin, T. J., Vink, P. : Thirty years of anthropometric changes relevant to the width and depth of transportation seating spaces, present and future. Applied ergonomics, 65, 130-138 (2017).
14. Smulders, M., Berghman, K., Koenraads, M., Kane, J. A., Krishna, K., Carter, T. K., Schultheis, U.: Comfort and pressure distribution in a human contour shaped aircraft seat (developed with 3D scans of the human body). Work, 54(4), 925-940 (2016).
15. Vanacore A., Lanzotti A., Percuoco C., Capasso A., Vitolo B.: Design and analysis of comparative experiments to assess the (dis-) comfort of aircraft seating. Applied ergonomics, 76, 155-163 (2019).
16. Vink P.: Aircraft interior comfort and design. CRC press (2016).

Climate Change and Italian Cities: evidence from Meteo-climatic Statistics and Indices on Extreme Events

Cambiamenti Climatici e Città Italiane: le Statistiche Meteoclimatiche e gli Indici di Estremi Climatici

Donatella Vignani, Francesca Budano, Claudia Busetti[1]

Abstract Increasing information needs on Climatic Change (CC) has favoured statistical frameworks development defined by international institutions to provide standardized and harmonized methodologies for producing data and indicators. In 2018, Istat has disseminated Survey on *Meteo-climatic and Hydrological Data* results, developing statistical tools to analyse meteo-climatic conditions and variability in main Italian cities. Meteo-climatic statistics and indicators provide useful tools to design efficient policies, support adaptation and disaster risk analysis and actions in cities governance. In recent years Italian cities recorded a growing climatic variability and extreme events causing relevant impacts on their territories.

Abstract *Crescenti fabbisogni informativi sui Cambiamenti Climatici (CC) ha favorito lo sviluppo di framework statistici definiti da istituzioni internazionali per fornire metodologie standardizzate per la produzione di dati e indicatori. L'Istat nel 2018 ha diffuso i principali risultati della rilevazione Dati Meteo-climatici e Idrologici, sviluppando strumenti statistici per analizzare condizioni e variabilità del clima nelle città capoluogo di regione. Tali statistiche e indicatori possono supportare scelte di policy per l'adattamento al clima, analisi dei rischi e misure di governance. Negli ultimi quindici anni le principali città italiane hanno registrato una crescente variabilità climatica ed eventi estremi con impatti rilevanti.*

Key words: meteo-climatic statistics and indices, climate variability, urban systems

[1] Donatella Vignani Ph.D. Researcher, Istat Italian National Institute of Statistics, Environmental and Territorial Statistics Department, Rome email: vignani@istat.it

Francesca Budano, Istat Italian National Institute of Statistics, Environmental and Territorial Statistics Department, Rome email: frbudano@istat.it

Claudia Busetti, Istat Italian National Institute of Statistics, Environmental and Territorial Statistics Department, Rome email: busetti@istat.it

1 Climate change: international organization and EU strategy on adaptation

Climate Change (CC) is in place in many areas of the World. CC mitigation and adaptation, disaster risk reduction are priorities objectives for many International Institutions. The United Nations (UN) encourage analyses and research to integrate adaptation into national policies and require all countries improving their knowledge on meteo-climatic and CC data, to evaluate CC impacts and future risks, to adopt response strategies. The World Meteorological Organization (WMO) and United Nations Environment Program (UNEP) set up in 1988 the Intergovernmental Panel on Climate Change (IPCC) to provide policy makers with assessments of CC scientific basis and climate models development [7]. The Covenant of Mayors[1] initiative was launched by the European Commission (EC) in 2008, to engage and support local governments to reach the EU climate and energy targets. In 2013 EC adopted an EU strategy on CC mitigation-adaptation to make Europe more climate-resilient.

In 2015, the National Strategy of Adaptation to Climate was approved also in Italy. The National Adaptation Plan to CC (PNACC) supports national, regional and local public institutions to contain vulnerability of natural, social and economic systems to CC impacts and increase their adaptive capacities. Many Italian regions are defining and developing adaptation strategies and policies. Abruzzo, have approved a Regional Planning Document about the roadmap to reach a Regional Adaptation Plan. Friuli Venezia Giulia, Marche, Molise, Piemonte, Puglia, Valle d'Aosta and Provincia Autonoma of Trento started routes to prepare a Regional Adaptation Strategy to CC. Lombardia, starting from the Regional Strategy for adaptation to CC (SRACC 2014) approved in december 2016 the Regional Action Document. The Sardegna Region in February 2019 adopted the Regional Strategy for adaptation to climate change (SRACC), in which all the actions and objectives are coordinated at a regional level, through the adoption of a specific governance model for i) transferring strategies for adaptation in regional and local planning and ii) programming processes. To be effective, regional policies need to be developed at the local level, combining and integrating top-down and bottom-up approaches. In 2015 Sorradile (in Oristano province) is the first municipality in Sardegna region to have adopted a plan for CC adaptation (PACC). This plan represents - for the first time - a replicable model of approach to specific adaptation for small rural centers in Italy. At local level, Ancona and Bologna approved in 2013 Local Adaptation Plans on risks and response actions. In particular, Bologna identified adaptation measures to contrast drought and water scarcity, heat weaves in urban areas and extreme events of precipitation and hydro-geological risk with the aim of urban renewal to a soil consumption reduction and a regeneration of areas with abandoned buildings

[1] See https://www.eumayors.eu/en/. About 4000 Italian municipalities join the EU Covenant of Mayors and 80% had submitted a Sustainable Energy (and Climate) Action Plan describing its CC mitigation actions.

(brownfields). Ancona plan identifies landslides, coastal erosion and several impacts on infrastructure connectivity and mobility and on cultural heritage as the main CC risks. Management actions aimed to improve knowledge on erosion determinants have been identified. Measures such as the training of specific professionals for assessment and monitoring, for analysis of historical and cultural heritage have been encouraged. Other measures have been developed, such as setting up technological infrastructure actions, improving and extending technologies for monitoring landslides and coasts protection also by moving back from the coastline. Moreover public information campaigns have been regularly made.

2 Meteo-climatic official statistics and Indices on temperature and precipitation provided by Istat

Meteo-climatic statistics and Indices can describe climate variability and changes at spatial scale. Statistics on different thematic areas, such as environment, climate, human health, urban systems, public utilities, can be put in relation with meteo-climatic statistics to provide useful tools for policy makers designing efficient policy, supporting adaptation and disaster risk analysis and actions in cities governance [4].

Standardized, complete and comparable information systems are needed to develop complex analysis. Integration of statistical and geographical information is crucial to build a complete and updated knowledge framework at local scale.

To provide common methodologies and internationally comparable CC Related Statistics and Indicators (CCRSI) referenced statistical frameworks have been developed by international Organizations. National Statistics Offices (NSO's) are so called to enhance their role to provide high quality statistics reliable, consistent, timely, comparable.

For these reasons Italian National Institute of Statistic (Istat) provided in 2018 a new set of data through the Survey on Meteo-climatic and Hydrological Data[1] [1,2]. To update Istat geo-database of meteorological daily observations (time series 1971-2016), the survey have collected data on climatic variables (minimum, maximum and mean temperature, precipitation level) from 65 National Institutions and Agencies managing meteorological stations. Among the main providers, Italian Air Force Meteorological Service RSMC, Italian Institute for Environmental Protection and Research ISPRA, Italian Air Navigation Service Provider ENAV, Council for Research in Agriculture and Analysis of Agriculture Economics CREA-AA, Regional Agencies for Environmental Protection ARPA, Regions, Provinces, Universities). Concerning the 21 Italian Regional Capital Municipalities (IRCM), where 38.9% of Italian population lives, daily measurements are collected from 122 gauging stations located within urban territory. Quality statistical controls have

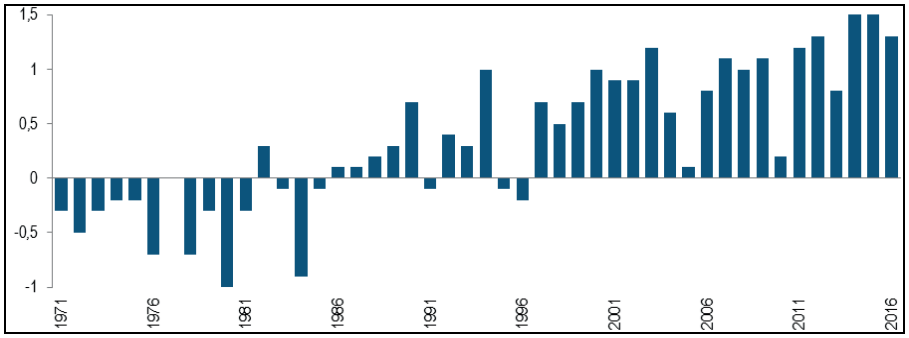
[1] Included in the National Statistical Program (PSN IST-02190).

reduced the number of stations used for analyses. Meteo-climatic conditions and variability in IRCM considered have been analysed, mainly observing *temperature* and *precipitation anomalies,* comparing 2002-2016 mean values with Climatic Normal (CLINO 1971-2000) values. A set of ETCCDI *Indices of climatic extremes*[1], is calculated by city and by year.

3 Evidences from data and Indices on climate extreme events on Italian regional capital municipalities

Considering the 2002-2016 period, in the regional capital municipalities observed the *mean temperature* is on average equal to +15.5°C, +1.0°C with respect to the correspondent CLINO value [3]. Analysing the anomalies of mean temperature of each year of the time series 1971-2016 compared to the 1971-2000 (CLINO) mean temperature value, figures shows in the decade 1971-1981 mean temperature anomalies assume negative values (Figure 1), while in 1985 seems to be change of such previous trend. Moreover, after 1996, anomalies assume always positive and growing values with respect to the CLINO value. In presence of continuous increases in mean temperature, 2014 and 2015 represents the years recording the highest values of the time series observed (16° C), with an anomaly of about + 1.5 ° C with respect to the climatic value. According to data collected, 2016 is the third warmest year since 1971 with a mean temperature of +15.8 ° C and an anomaly equal to + 1.3 ° C.

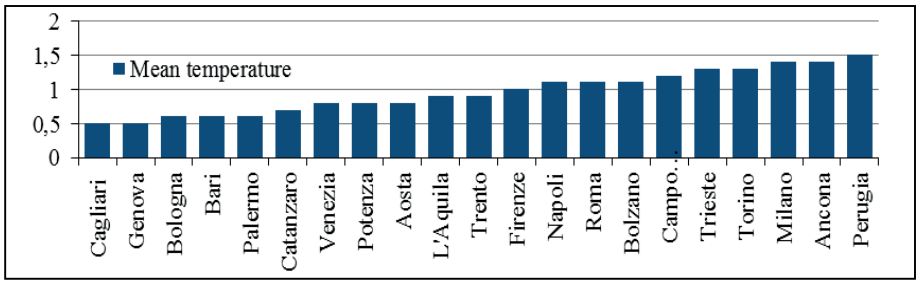
Figure 1: Differences of annual mean temperature (climatic anomalies) of the stations analysed on CLINO 1971-2000 mean value, by year, absolute values in °C



As main Italian cities are concerned, all urban systems register positive values and 43% exceeds the respective CLINO value by +1.0°C (Figure 2). Perugia records the highest anomalies (+1.5°C) and Cagliari and Genova the lowest (+0.5°C).

[1] ETCCDI methodology is developed by the Expert Team on Climate Change Detection and Indices, within the UN World Climate Research Program (WCRP) in collaboration with the World Meteorological Organization (WMO-UN).

Figure 2: Mean temperature average anomalies of the period 2002-2016 compared to CLINO value (1971-2000) by Italian regional capital municipalities (°C)

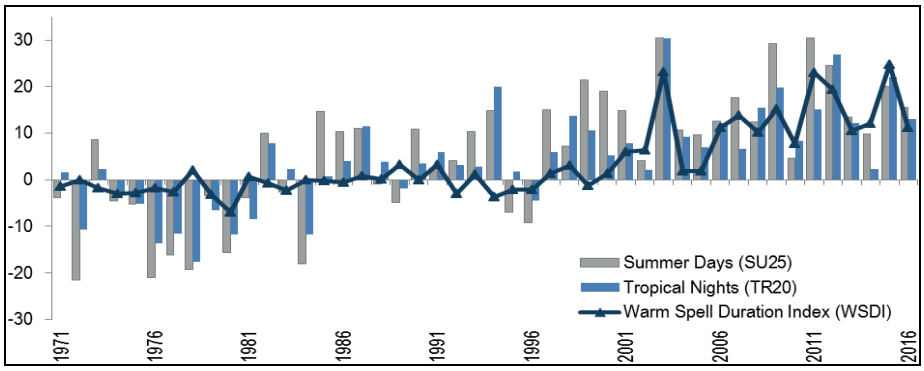


Concerning *Indices on climatic extremes*, results confirm a warming trend in Italian cities in the period observed [5]. The average annual number of *Summer days index* (maximum temperature >25ºC) exceeds by +17 days the CLINO value (93 days). All cities present positive anomalies and about 76% of them record more 10 more days then the CLINO value. Ancona and Perugia measure the highest anomalies of summer days (respectively +34 and +33 days). *Tropical nights index* (minimum temperature does not fall below 20ºC) increases (on average among cities) of +14 nights with respect to the CLINO value (31 nights) and 66.7% of cities register anomalies higher than +10 nights. Napoli records the highest anomaly (+34).

Also the *Warm spell duration index* (a measure of heat waves) value increases in 2002-2016 (+12 days) with respect to 1971-2000 value (11 days). More than ten cities register anomalies higher than +10 days. Perugia (+34 days), Trieste (+31), Ancona (+30) and Roma (+28) register the highest positive anomalies. The mean annual number of *Frost days* (minimum temperature <0°C) reduces by 3 days with respect to the CLINO value (26 days). Ten cities register negative anomalies: Bolzano (-20 days), Trento (-14) and Bologna (-10) measure the highest. L'Aquila (+10 days), Aosta and Firenze (+3), Perugia (+1) are the only ones recording increases by respective CLINO value.

A combined use of such indices allow to analyze interactions and effects between different observed phenomena. For example, analyzing 1971-2016 time series of annual anomalies of *Warm spell duration index WDSI* (Figure 3), since 2000 an increasing trend of number of days characterized by heat waves occurred, showing always positive and growing anomalies with respect to the correspondent CLINO value. Within the time series analyzed, the highest value of the WSDI index is registered in 2011 (+35 days) followed by 2015 (+34). Also 2016, last year considered, has a positive anomaly (+11 days). In all these three years, Perugia, Trieste, Milan and Rome seems to be particularly affected by heat waves events.

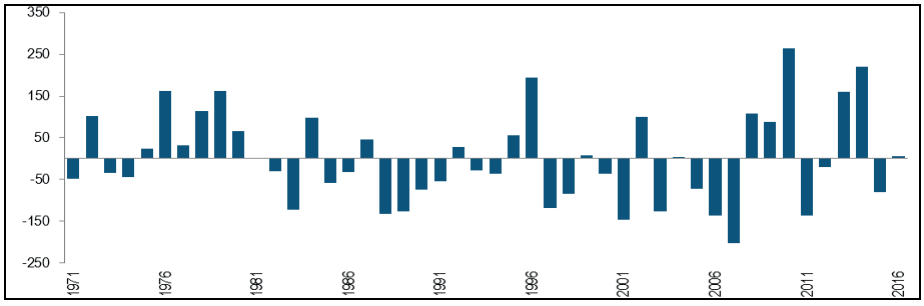
Figure 3: Annual anomalies average years 1971-2016, Summer Days (SU25), Tropical Nights (TR20), Warm Spell Duration Index (WSDI) compared to CLINO 1971-2000 values.



A high spatial and temporal variability seems to characterize precipitation phenomena among IRCM observed. Considering all these cities, the *Total annual precipitation* average value is equal to 778 mm (2002-2016), +1.6% with respect to the CLINO value (765.8 mm).

Analysing 2002-2016 time series (Figure 4), the variability of anomalies is growing and the most and the least rainy years since 1971 are present at the same time. Regarding positive anomalies, the highest value is registered in 2010 (+264 mm) which is the most rainy year since 1971 with a *Total annual precipitation* average value equal to 1,030 mm, followed by 2014 (+221) and average value equal to 987 mm of precipitation. 2007 is the year with the lowest negative anomaly since 1971 (-204 mm) and it is also the least rainy year of the entire series examined (562 mm).

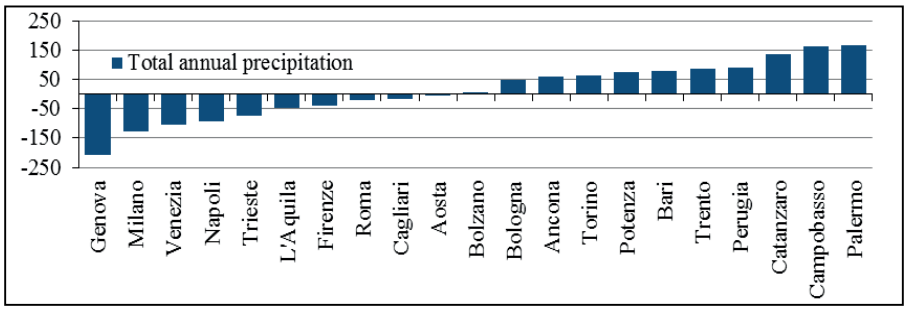
Figure 4: Differences of annual mean precipitation (climatic anomalies) on CLINO 1971-2000 correspondent mean value, by year, absolute values in mm



Ten cities record negative anomalies (Figure 5). In the period 2002-2016, Genova, ranking second about total annual average precipitation among the cities examined, presented the highest difference of total annual average precipitation (-206.2 mm) from the correspondent mean CLINO value recorded. Also Milano (-125) and Venezia (-104.8) record high values of negative anomalies. Differently, in

the South, Palermo (+166.8 mm), Campobasso (+162.1), Catanzaro (+136.8) register the highest positive anomalies.

Figure 5: Total annual precipitation mean anomalies of the period 2002-2016 value compared to CLINO value (1971-2000), by Italian regional capital municipalities, absolute values in mm.



In 2002-2016 *Days with precipitation>1 mm index* on average record 82 annual number of days, in line with the CLINO value. Anomalies variate between -5 days of Venezia to +8 of Palermo. The average annual number of days with precipitation higher than 20 mm (R20 index) also remained steady with respect to the normal value (on average 10 days). Variation ranged from -3 days in Milan and Genova to +3 days in Campobasso, Catanzaro, Palermo and Trento. Also the annual number of *Days with precipitation >50 mm index,* remain steady with reference to the respective CLINO values (on average 1 days). *Total precipitation in very rainy days index* is on average equal to 192 mm, that is to say 24.7% of total annual precipitation.

4 Conclusions and further work developments

Statistics and indicators provided by Istat for IRCM point out relevant variations in temperatures trend and precipitation patterns, strengthening information yet available at a local scale to support adaptation policies and CC action responses. Same statistics calculated for Italian provincial municipalities are being to release [4]. New analyses based on integration of statistics of different domains, annual update of database and strengthening cooperation with other meteo-climatic data producers are further work developed by Istat. Meteoclimatic statistics provided by Istat strengthen information available at local and temporal scale to support capacity building in the area of adaptation policy and CC action responses.

5 References

1. Istat, Statistica Report "Temperatura e precipitazione nelle principali città (Anni 2001-2016)" https://www.istat.it/it/archivio/217402 (2018)
2. Istat, Annuario Statistico Italiano 2018, Capitolo 2 Ambiente ed energia https://www.istat.it/it/archivio/225274 (2018)
3. Istat, Tavole di Dati "Dati meteoclimatici nelle città capoluogo di provincia e città metropolitane Anni 2007-2106" Istat (2019)
4. Ispra, Variazione e tendenze degli estremi di temperatura e precipitazione in Italia (2013)
5. Vignani D., Budano F., Busetti C., An application of Indices on Extreme Climate Events on Italian Regional Municipalities, AISRe Conferenze 2018
6. Ispra, XIV Rapporto scientifico "Gli indicatori del Clima in Italia" (2018)
7. IPCC (United Nations), Special Report Global Warming of 1.5 ºC (2019)