11th Scientific Meeting of the SIS Group
"Statistics for the Evaluation and Quality in Services"

BOOK OF SHORT PAPERS

Editors
Andrea Bucci
Alfredo Cartone
Adelia Evangelista
Andrea Marletta

IES PESCARA 2023

STATISTICAL METHODS FOR EVALUATION AND QUALITY: TECHNIQUES, TECHNOLOGIES AND TRENDS (T^3)

Edizioni Il Viandante

IES 2023 - Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)

BOOK OF SHORT PAPERS

Editors: Andrea Bucci, Alfredo Cartone, Adelia Evangelista and Andrea Marletta

Book of Short papers
11th International Conference **IES 2023**
Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)

University 'G. d'Annunzio' of Chieti-Pescara

Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services - SVQS

Maurizio Carpita	University of Brescia
Pietro Amenta	University of Sannio
Matilde Bini	European University of Rome
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Luigi D'Ambra	University of Naples 'Federico II'
Paolo Mariani	University of Milan 'Bicocca'
Marica Manisera	University of Brescia
Monica Palma	University of Salento
Pasquale Sarnacchiaro	University of Naples 'Federico II'

Program Committee of the conference IES 2023

Chair: Paolo Mariani	University of Milan 'Bicocca'
Matilde Bini	European University of Rome
Mario Bolzan	University of Padova
Giuseppe Bove	University of Rome 3
Eugenio Brentari	University of Brescia
Ida Camminatiello	University of Campania 'L. Vanvitelli'
Maurizio Carpita	University of Brescia
Carlo Cavicchia	Erasmus University Rotterdam
Roy Cerqueti	University of Rome 'La Sapienza'
Enrico Ciavolino	University of Salento
Corrado Crocetta	University of Bari 'Aldo Moro'
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Pierpaolo D'Urso	University of Rome 'La Sapienza'
Tonio Di Battista	University 'G. d'Annunzio' of Chieti-Pescara
Simone di Zio	University 'G. d'Annunzio' of Chieti-Pescara
Mauro Ferrante	University of Palermo
Francesca Fortuna	University of Rome 3
Michele Gallo	University of Naples 'Orientale'
Claudiu Herteliu	Bucharest University of Economic Studies
Salvatore Ingrassia	University of Catania
Ron Kenett	Samuel Neaman Institute & KPA Group
Michele La Rocca	University of Salerno
Rosaria Lombardo	University of Campania 'L. Vanvitelli'
Antonio Lucadamo	University of Sannio
Marica Manisera	University of Brescia
Andrea Marletta	University of Milan 'Bicocca'
Lucio Masserini	University of Pisa
Fabrizio Maturo	University Mercatorum
Matteo Mazziotta	ISTAT
Rodolfo Metulini	University of Bergamo
Stefania Mignani	University of Bologna
Enrico Ripamonti	University of Brescia

Silvia Salini	University of Milano
Pasquale Sarnacchiaro	University of Naples 'Federico II'
Violetta Simonacci	University of Naples 'Federico II'
Maurizio Vichi	University of Rome 'La Sapienza'
Mariangela Zenga	University of Milan 'Bicocca'

Organizing Committee

Chair: Tonio Di Battista	University 'G. d'Annunzio' of Chieti-Pescara
Andrea Bucci	University 'G. d'Annunzio' of Chieti-Pescara
Alfredo Cartone	University 'G. d'Annunzio' of Chieti-Pescara
Marco Di Marzio	University 'G. d'Annunzio' of Chieti-Pescara
Marta Di Nicola	University 'G. d'Annunzio' of Chieti-Pescara
Simone Di Zio	University 'G. d'Annunzio' of Chieti-Pescara
Stefania Fensore	University 'G. d'Annunzio' of Chieti-Pescara
Lara Fontanella	University 'G. d'Annunzio' of Chieti-Pescara
Stefano Antonio Gattone	University 'G. d'Annunzio' of Chieti-Pescara
Luigi Ippoliti	University 'G. d'Annunzio' of Chieti-Pescara
Eugenia Nissi	University 'G. d'Annunzio' of Chieti-Pescara
Paolo Postiglione	University 'G. d'Annunzio' of Chieti-Pescara
Annalina Sarra	University 'G. d'Annunzio' of Chieti-Pescara
Pasquale Valentini	University 'G. d'Annunzio' of Chieti-Pescara
Adelia Evangelista (Secretariat)	University 'G. d'Annunzio' of Chieti-Pescara

Editors

Andrea Bucci	University of Macerata
Alfredo Cartone	University 'G. d'Annunzio' of Chieti-Pescara
Adelia Evangelista	University 'G. d'Annunzio' of Chieti-Pescara
Andrea Marletta	University of Milan 'Bicocca'

Sede legale: Via Silvino Olivieri, 111 - 66100 Chieti (CH)

ISBN 979-12-803-3369-8
DOI 10.60984/978-88-94593-36-5-IES2023
https://doi.org/10.60984/978-88-94593-36-5-IES2023

Sponsored by

Preface

Statistical thinking, design and analysis play a crucial role in social life and are useful to society at large. Besides, promoting advanced methodological research is useful to facilitate the dissemination of ideas related to various fields of interest. For this purpose, experts in statistics, data analysis, data mining, statistical methods for decision making, machine learning and related methods come together to understand and analyse phenomena through data.

In line with this objective, the Statistics Group for the Evaluation and Quality of Services (SVQS; www.svqs.it) of the Italian Statistical Society (SIS) has been organizing the Innovation and Society (IeS) conference biennially since 2009, focusing on new developments and ideas in statistics applied to the evaluation and quality of public and private services, attracting national and international statisticians and data scientists. The meeting contributes to spot light on the main statistical approaches and methodologies for the evaluation of public services currently in use in different contexts, as well as to facilitate discussion on the impact of innovative statistical evaluation systems for these services, involving various economic and social policy actors.

The conference "Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)" recorded valuable contributions that are reported in this volume. The papers underscore how the growing availability of data has tasked social and economic actors, organizations, and researchers with the management and analysis of large volumes of unstructured and heterogeneous data. In recent years, many tools for both qualitative and quantitative models have been developed to better describe and understand complex systems and their underlying behaviors, and the papers reported in this volume bear witness to this.

Techniques, technologies and trends: the study of data complexity presents the potential to provide analyses with increased frequency and timeliness, accuracy and objectivity, and to define sustainable models. Traditional quantitative methods for capturing socioeconomic data have often shown limitations in their ability to examine underlying systems, and with the three 'T' just mentioned, the outlines of future developments are starting to emerge.

The volume reports 127 contributions in the following areas:

- Advanced statistical methods for pattern recognition
- Advances in statistical learning from high-dimensional data
- Data analysis for web sources
- Distance and depth-based statistical learning methods for robust data analysis

- Economics and environment
- Education and labour
- Inequalities in the labour market
- Innovations and challenges in official statistics
- Labour market: trends, perspectives and new challenges
- Methodological and applicative contributions for evaluating sustainable development
- Methodological developments and applications for the assessment of student competencies
- Networks data analysis: new perspectives and applications
- New advanced statistical methods for data science
- Recent advances in statistical learning and data analysis
- Statistical analysis and modeling of environmental pollution data
- Statistical methods and complexity for evaluation in finance
- Statistical methods and composite indicators for healthcare
- Statistical methods and models for land monitoring with spatio-temporal data
- Statistical methods for environmental monitoring and sustainability
- Statistical methods for the analysis of university student choices and academic performance
- Statistical methods for the assessment of transport services and sustainable emissions
- Statistical methods for education and educational services
- Statistics in sports
- Tourism and territory.

The Conference event attracted many contributions as well as numerous Authors, not just from Italy but also from abroad. Over the three-day meeting, the Community has the opportunity to witness some of the state-of-the arts, new trajectories, and methodological challenges in 24 solicited sessions, 7 sessions of free contributes, two round tables - organized by Maurizio Vichi and Matilde Bini respectively - and three keynotes sessions with Ron S. Kennet of Samuel Neaman Institute of Israel, Luigi D'Ambra of Federico II University of Naples, and the former Minister Enrico Giovannini from University of Tor Vergata.

Organizers

Paolo Mariani
Chair of the Program Committee
Tonio Di Battista
Chair of Local Organizing Committee
Maurizio Carpita
Coordinator of the SVQS Group

Contents

Plenary Session

1. *Befitting Cross Validation with Three Case Studies* (Kenett R.S.)

Befitting Cross Validation with Three Case Studies

Ron S. Kenett

Abstract Computer age statistics typically involves large amounts of data and application of computer intensive methods. In this talk we focus on cross validation methods that account for the data generation structure. We discuss the use and limitations of cross validation and introduce befitting cross validation (BCV). Specifically, we show how BCV is used to validate predictive models. The talk will focus on the impact of the data structure on the implementation algorithm with three case studies. One case study is an industrial burn procedure designed to ensure quality of electronic products. The second case study is an association rule analysis of reported drug side effects. It is an application of text analytics to patient comments. The third case study is about demand forecasting used to optimize stock levels in consumer products retail.

References

1. Kenett R.S. and Zacks S.: Modern Industrial Statistics: With Applications in R, MINITAB and JMP, 3rd Edition, ISBN: 978-1-119-71490-3 (2021)
2. Kenett R.S., Zacks S. and Gedeck P.: Modern Statistics: A Computer-Based Approach with Python. Switzerland, AG: Springer Nature (2022)
3. Kenett R.S., Gotwalt C., Freeman L. and Deng X.: Self-supervised cross validation using data generation structure. Applied Stochastic Models in Business and Industry (2022) DOI: 10.1002/asmb.2701

Ron S. Kenett
The KPA Group and the Samuel Neaman Institute, Technion, Israel, e-mail: ron@kpa-group.com

Solicited Session SS1 - *Statistical analysis and modeling of environmental pollution data*

Session of the SIS-GRASPA organized by Pasquale Valentini and Natalia Golini
Discussant: Pasquale Valentini

1. *Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution* (Bottazzi Schenone M., Grimaccia E. and Vichi M.)
2. *Evaluating the nonlinear association between PM_{10} and emergency department visits* (Bucci A., Sanmarchi F., Santi L., Giostra F., Tubertini E., Rosa S., Nante N. and Golinelli D.)
3. *Estimating spatially varying Gaussian Graphical Models to unveil relationships among pollutants in the Red River Delta in Vietnam* (Pronello N., Cucco A., Ignaccolo R. and Ippoliti L.)

Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution

Una misura di qualità ambientale ottenuta clusterizzando un indice basato su modelli ad equazioni strutturali: una applicazione alla qualità dell'aria nelle principali città europee

Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Abstract This paper proposes an innovative computational procedure to determine the optimal number of clusters. The aim is to identify the maximum number of significantly distinct clusters, when the centroids are orderable and order is relevant. The insight is that ranking according to this optimal number of clusters allows to better classify units in order to assess their quality with regard to a variable of interest. By means of bootstrap confidence intervals estimated on clusters' centroids, the procedure allows to identify the optimal number of "well-separated" groups. The centroids are obtained applying a unidimensional k-means clustering and they allow to classify and rank the measure of an Index based on a Structural Equation Model. The procedure ranks European cities according to their level of air pollution.

Abstract *Il lavoro propone una procedura computazionale innovativa per determinare il numero ottimale di cluster. Lo scopo è identificare il numero massimo di cluster significativamente distinti, quando i centroidi sono ordinabili e l'ordine è rilevante. L'intuizione è che la classificazione in base a questo numero ottimale di cluster consente di classificare le unità al fine di valutarne la qualità rispetto a una variabile di interesse. La procedura consente di identificare il numero ottimale di gruppi "ben separati", mediante intervalli di confidenza bootstrap. I centroidi sono ottenuti applicando un clustering k-medie unidimensionale e permettono di classificare un Indice stimato con un Modello ad Equazioni Strutturali. La procedura consente di classificare le principali città europee in base al loro livello di inquinamento atmosferico.*

Mariaelena Bottazzi Schenone
Department of Statistical Sciences, Sapienza University, Rome (Italy), e-mail: mariaelena.bottazzischenone@uniroma1.it

Elena Grimaccia
ISTAT - Italian National Institute of Statistics, Rome (Italy), e-mail: elgrimac@istat.it

Maurizio Vichi
Department of Statistical Sciences, Sapienza University, Rome (Italy), e-mail: maurizio.vichi@uniroma1.it

Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Key words: Bootstrap confidence intervals, Simulation study, Environmental quality, Structural Equation Models, Cluster analysis, Air Pollution.

1 Introduction

Most of the commonly employed clustering methods aims at identifying the optimal minimum number k of centroids [16,14,15]. However, when the aim of the study is the ranking of units according to a measure, it would be useful to identify the maximum number of clusters. This paper presents a procedure aimed at finding the optimal maximum number of well separated clusters, classifying an index resulting from a Structural Equation Model (SEM). In the application, the k-means clustering is applied to an index that measures Air Pollution, taking into account simultaneously the six main pollutants usually considered in the literature. The estimation of the SEM accounts also for several meaningful socio economic and climate-related covariates that enhance the significance of the estimates. The clustering of European metropolitan areas with respect to different air pollution levels is presented. The number of centroids has been chosen, considering the maximum number of clusters whose (1-α)% confidence intervals do not overlap by more than α. The ranking of the main 130 European cities provides useful information to design policies, aimed at reducing urban air pollution [5].

2 Data

The six main air pollutants identified by the Environmental Protection Agency (EPA) are: Ground-level ozone (O_3), Particle pollution (also known as Particulate Matter (PM), including PM2.5 and PM10), Carbon monoxide (CO), Sulphur dioxide (SO_2) and Nitrogen dioxide (NO_2). Data on pollutants in 130 metropolitan areas in the European Union are obtained from the Worldwide Air Quality data (https://aqicn.org/), which cover pollutants and atmospheric conditions around the world [1]. In addition, socio-economic features of cities (GDP per capita, population density, elderly and youth dependency ratios, employment, unemployment and participation rates) are included [12,4] together with meteorological and atmospheric covariates: air temperature, humidity, air pressure, wind-gust (m/s) and wind-speed (m/s), according to [10]. Traffic-related air pollutant emissions have become a global environmental problem, most of all in urban areas [3]. Therefore, the motorization rate (Number of passenger cars per thousand inhabitants, available at country level), and the Number of registered cars per 1000 population (at city level) have been included in the study from the Eurostat metropolitan regions (NUTS3) database. Geographical covariates (Latitude and Longitude) have been included in the analysis, in order to take into account the spatial configuration of the phenomenon of air pollution [17].

3 Methods

The Air Pollution Index employed for measuring atmospheric environmental quality is built with a Structural Equation Mode that takes into account both endogenous and exogenous variables [7]. The relationships between observed (manifest) variables and latent factors, and among latent constructs themselves are estimated simultaneously [8]. In order to rank units with respect to the SEM-based index, the centroid-based model of k-means [18] is employed. The k-means method assumes that each observation is equal to one of the k centroids. All the observations assigned to each centroid, perturbed by error in measuring the features, forms a cluster. The clustering goal is to partition the units in a disjoint set of k clusters to maximise the dissimilarity between centroids of the clusters. Because of its deterministic nature, k-means does not yield confidence information about centroids' distribution and estimated cluster memberships, although this could be useful for inferential purposes. It is possible to achive such information by means of a non-parametric bootstrap procedure. This procedure provides centroids' distributions [11] which can be used to derive probabilistic membership information on each object from all bootstrap samples. It also yields confidence information about the centroids in the form of confidence intervals [9]. Given a sample of units and a number of clusters k, this can be done bootstrapping those units a number B of times. The results are B vectors of k centroids. The final estimates of the k clusters' centroids as well as their empirical distributions can be obtained computing the mean and plotting the histograms of the bootstrap replicates. Given the k centroids' point estimates with the corresponding $\alpha/2$ and $(1-\alpha/2)$ percentile estimates, it is possible to build k percentile confidence intervals of the desired confidence level α. If some of these confidence intervals do overlap by more than α, then the clusters are not "well-separated". The optimal number of clusters k* will be the maximum k such that none of the k intervals do overlap by more than α. Given a sample of size n, for a given k, the partitioning algorithm is run. The corresponding k centroids' confidence intervals are built applying bootstrap to that sample of n units. Bootstrap allows to estimate each of the k centroids as the mean of the centroids' values for all the bootstrap replicates. This technique allows also to estimate the corresponding centroid standard error and therefore compute the $(1-\alpha)\%$ percentile confidence interval [13]. The clustering algorithm has been applied for different values of k, starting from k = 2. If the k bootstrap confidence intervals do not overlap, the k clusters can be considered well separated, k is increased by one and the partitioning algorithm is run again. The procedure is iterated until two overlapping confidence intervals are found. A crucial point is the need of ordering the clusters with respect to their centroid value, from the smallest to the largest. This allows to find the consecutive clusters' confidence intervals to be compared. The partitioning algorithm chosen is a centroid-based 1-dimensional k-means and units are classified according to an index built my means of SEM. In this particular unidimensional case, an optimal dynamic programming algorithm has been developed by [6].

4 Results

In this study, a multidimensional index to measure air pollution is built by means of a hierarchical SEM [2]. This model has the advantages of taking simultaneously into account a number of levels in the hierarchy and to exploit the information available in meaningful explanatory variables [7]. We called the resulting index Air Pollution Index (API). Cluster analysis is then applied to find groups of cities homogeneous with respect to the air pollution level. European cities are grouped into clusters, each represented by a centroid that corresponds to an API value. It is important to note that to allow cities' ranking, clusters must be ordered with respect to the corresponding centroids. The clustering technique of 1 dimensional k-means is applied using the R function "Ckmeans.1d.dp" of the homonymous R package [19]. The clustering algorithm is run for k = 2 up to 10. The bootstrap procedure (with a number of bootstrap replicates equal to 10000) is used to compute the corresponding centroids confidence intervals at 90% shown in Fig.1.

When the difference between the upper bound of a cluster and the lower bound of the consecutive one is smaller than $\alpha = 0.05$, then they are considered well separated. In this application, for k = 8 the clusters do overlap by more than α, and therefore the optimal k* is 7. It is worth mentioning that this method of choosing k is very different compared to the Elbow, Silhouette or Gap Statistics methods, whose aim is to find the minimum k, such as the units are optimally allocated in separate clusters. According to all these 3 methods, in fact, k*=2.

Based on the previous results groups are ranked from 1 to 7 considering the centroids' values from the highest to the lowest: rank 1 corresponds to the lowest centroid and therefore to the group of less air polluted cities. Fig. 2 shows groups of European cities with a similar situation in terms of air pollution levels in 2022. It

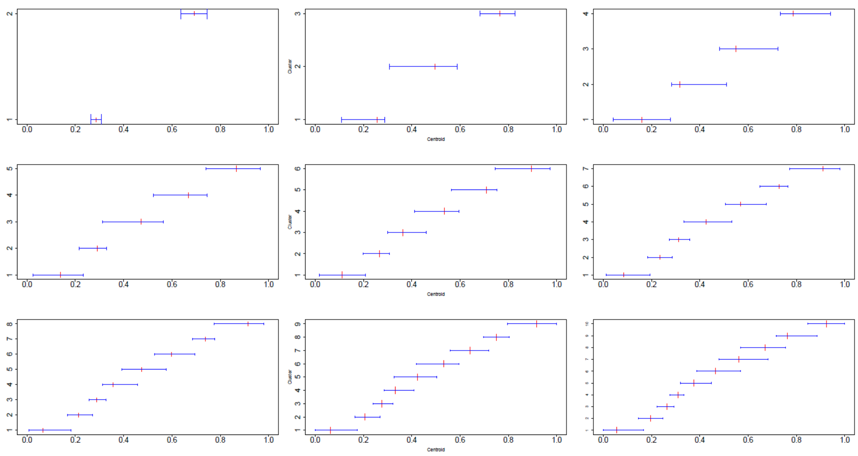


Fig. 1 95% bootstrap confidence intervals for k-means centroids. k ranges in 2-10.

is possible to note that close points tend to have the same colour: cities in the same country mostly have a similar air pollution level.

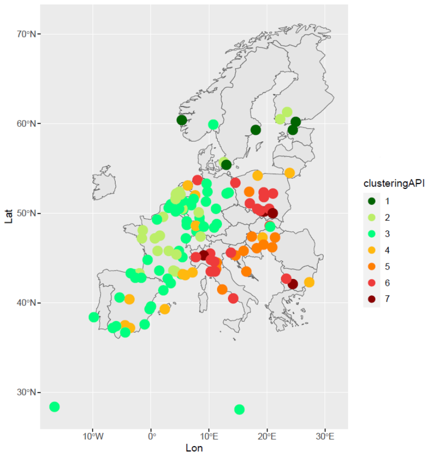


Fig. 2 Clusters of European cities according to API.

5 Concluding remarks

In this paper, European cities are grouped in clusters and ranked with respect to their air pollution level, measured by a structural equation model based index (API). The optimal number of clusters is the maximum number of significantly different centroids, according to centroids' percentile confidence intervals built by means of bootstrap. In this way, an innovative procedure aimed at identified the best maximum number of cluster is exemplified. Moreover, European cities are more granularly classified in seven clusters, with a gain in information, compared to the much smaller number of clusters suggested by other classical methods. The procedure could be improved employing a multidimensional clustering technique, to be compared to this unidimensional approach.

References

1. Boaz R. M., Lawson A. B., Pearce J. L.: Multivariate air pollution prediction modelling with partial missingness. Environmetrics, 30(7): e2592 (2019)
2. Cavicchia C., Vichi M.: Second-order disjoint factor analysis. Psychometrika, 87 (1), 289–309 (2022)
3. Choma E. F., Evansb J. S., Gomez-Ibanezc J. A., Did Q., Schwartzb J. D, Hammitte, J. K., Spenglerb J. D.: Health benefits of decreases in on-road transportation emissions in the United States from 2008 to 2017. PNAS, 118 (51) (2021)
4. Davis M. E.: Recessions and Health: The Impact of Economic Trends on Air Pollution in California. Am J Public Health, 102(10), 1951–1956. (2012)
5. Dominici F., Samet J. M., Zegeral S. L.: Combining Evidence of Air Pollution and Daily Mortality from the 20 Largest US Cities: A Hierarchical Modelling Strategy. Journal of the Royal Statistical Society Series A. (2000)
6. Froese R., Klassen J. W., Leung C. K. and Loewen T. S.: The Border K-Means Clustering Algorithm for One Dimensional Data. IEEE International Conference on Big Data and Smart Computing, pp. 35-42. (2022)
7. Grimaccia E., Bottazzi-Schenone M., Vichi M.: Structural-Equation-Model-based assessment of Pollution in European urban Areas. Conference of European Statistics Stakeholders 2022. Available via https://drive.google.com/file/d/1V3BN-K9SY66Q7mB5UnGuQvJSGYV7f2dC/view Cited 10 March 2023
8. Hair J. F. and Sarstedt M.: Explanation plus prediction – The logical focus of project management research. Project Management Journal, 52(4), 319–322. (2021)
9. Hofmans J.: On the Added Value of Bootstrap Analysis for K-Means Clustering. Journal of Classification (2015)
10. Liu, Y., Zhou, Y., Lu, J.: Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Sci Rep 10, 14518. https://doi.org/10.1038/s41598-020-71338-7 (2020)
11. Martella F., Vichi M.: Clustering microarray data using model-based double K-means. Journal of Applied Statistics (2012)
12. Martori J.C., Lagonigro R., Pascual R.I.: Sustainable Cities and Society Social status and air quality in Barcelona: A socio-ecological approach. Sustainable Cities and Society, 87, 104210 (2022)
13. Rizzo M.: Statistical Computing with R. Computer Science and Data Analysis Series. Chapman and Hall/CRC The R Series. p. 198. (2008)
14. Rousseeuw P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. (1987)
15. Shi C., Wei B., Wei S. Wang W., Liu H., Liu J.: A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, (1), 1-16 (2021)
16. Tibshirani R., Walther, G., Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistic. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63(2), 411– 423. (2001)
17. Urdangarin A., Goicoa T. and Ugarte M.D.: Evaluating recent methods to overcome spatial confounding. Revista Matemática Complutense 10.1007/s13163-022-00449-8. (2022)
18. Vichi M., and Kiers H. A. L.: Factorial k-means analysis for two-way data. Computational Statistics and Data Analysis, 37, 49–64 (2001)
19. Wang H. and Song M.: Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. The R Journal Vol. 3/2. (2011)

Evaluating the nonlinear association between PM_{10} and emergency department visits

Valutazione dell'associazione nonlineare tra PM_{10} ed accessi al pronto soccorso

Andrea Bucci, Francesco Sanmarchi, Luca Santi, Fabrizio Giostra, Eleonora Tubertini, Simona Rosa, Nicola Nante and Davide Golinelli

Abstract Air pollution is one of the most threatening risk factors for human health. In fact, the epidemiological literature has widely proved that the exposure to high levels of PM_{10} is associated with an increase in cardiovascular and respiratory events. In this paper, we investigate the relationship between air pollution and emergency department visits for cardiovascular and respiratory diseases. Since the relationship between air pollution and emergency care accesses is complex and may by subject to regime switches, especially due to the COVID-19 pandemic, we propose to use a threshold autoregressive model where the regime-switching mechanism is driven by the temporal evolution. This allows us to identify one or more structural breaks and estimate the regime-changing relationship between emergency care accesses and PM_{10}. Using daily data from the metropolitan area of Bologna (Italy), we show that the effect of the air pollutant changes through time, mostly due to the outbreak of the COVID-19 pandemic.

Andrea Bucci
University of Macerata, Via Crescimbeni 14 Macerata, e-mail: andrea.bucci@unimc.it

Francesco Sanmarchi
Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, e-mail: francesco.sanmarchi@gmail.com

Luca Santi
IRCCS S. Orsola Malpighi, Bologna, e-mail: lucasanti@hotmail.it

Fabrizio Giostra
Department of Emergency, IRCCS S. Orsola Malpighi, Bologna

Eleonora Tubertini
Department of Emergency, IRCCS S. Orsola Malpighi, Bologna

Simona Rosa
Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna

Nicola Nante
Department of Molecular and developmental medicine, University of Siena, Siena, e-mail: nicola.nante@unisi.it

Davide Golinelli
Department of Molecular and developmental medicine, University of Siena, Siena, e-mail: davide.golinelli@unisi.it

Abstract *L'inquinamento atmosferico è uno dei fattori di rischio più rilevanti per le condizioni di salute nell'essere umano. Infatti, la letteratura epidemiologica ha ampiamente dimostrato che l'esposizione a livelli elevati di PM_{10} è associata ad un aumento degli eventi cardiovascolari e respiratori. In questo articolo, indaghiamo la relazione tra l'inquinamento dell'aria e le visite al pronto soccorso per malattie cardiovascolari e respiratorie. Poiché la relazione tra inquinamento dell'aria e accessi in pronto soccorso è complessa e può essere soggetta a cambiamenti di regime, soprattutto a causa della pandemia di COVID-19, proponiamo di utilizzare un modello auoregressivo a soglia in cui il cambio di regime è guidato dal tempo. Questo permette di identificare uno o più cambiamenti strutturali e stimare la relazione tra accessi al pronto soccorso e PM_{10} nei diversi regimi. Utilizzando i dati giornalieri dell'area metropolitana di Bologna, mostriamo che l'effetto dell'inquinamento sugli accessi cambia nel tempo, principalmente a causa dell'esplosione della pandemia di COVID-19.*

Key words: Air pollution, Emergency department visits, Threshold autoregressive model

1 Introduction

Air pollution is a significant risk factor for human health, in particular for respiratory diseases such as asthma, chronic obstructive pulmonary disease, and pneumonia [11]. Exposure to high levels of air pollution has also been linked to increased mortality and hospitalizations for cardio-respiratory diseases [2].

In recent years, there has been a growing concern about the potential impact of air pollution on emergency care accesses for the exacerbation of cardio-respiratory symptoms [1]. In this context, the COVID-19 pandemic has created unique challenges for understanding the relationship between air pollution and cardio-respiratory health. Lockdowns and other pandemic-related restrictions have led to significant changes in air pollution levels, as well as shifts in healthcare utilization patterns. It follows that the potential effects of the air pollution on an health-related measure, such as the emergency care accesses, may have changed through time in different regimes. To account for these regime changes and provide a more nuanced understanding of the relationship between air pollution and respiratory health, we propose to use a threshold autoregressive (TAR) model [10]. This approach allows for the identification of changes in the relationship between emergency department (ED) visits due to cardio-respiratory diseases and air pollution. We use as a transition variable the temporal trend so that each-regime switch is a structural break.

The remainder of the paper is organized as follows. Section 2 introduces the TAR model, while Section 3 presents the data and the results of the estimated relationship between emergency department visits and air pollution, while Section 4 concludes.

2 Threshold Autoregressive Model

Let y_t be the time series of emergency department visits, for $t = 1,\dots,T$, the 2-regime threshold autoregressive (TAR) model is defined as follows

$$y_t = \left(\phi_{1,0} + \sum_{j=1}^{p} \phi_{1,j} y_{t-j} + \boldsymbol{\beta}_1 \mathbf{X}_t\right) \mathbb{1}(s_t \leq c) + \\ + \left(\phi_{2,0} + \sum_{j=1}^{p} \phi_{2,j} y_{t-j} + \boldsymbol{\beta}_2 \mathbf{X}_t\right) \mathbb{1}(s_t > c) + \varepsilon_t \quad (1)$$

where $\mathbf{X}_t$ is a $k \times 1$ vector of exogenous variables (in our case the lagged values of the PM_{10}), $\boldsymbol{\beta}_i$ is the vector of coefficients associated with the covariates, with $i = 1,2$, p is the order of the autoregressive process, $\phi_{i,j}$ is the autoregressive parameter for the p-th lag, c is the threshold value that separates the two regimes, s_t is a weakly stationary transition variable that drives the regime-switching mechanism, and ε_t is a white noise error term. A typical choice for s_t is between the lags of the dependent variable, although exogenous variables are possible as well [7]. An interesting way to specify s_t is using a temporal trend, therefore $s_t = t/T$. In this way, the identification of the threshold parameter, c, coincides with identifying a structural break and the model can be considered a special case of the time-varying autoregressive model, where the parameters $\phi_{i,j}$ and $\boldsymbol{\beta}$ changes through regimes.

3 Estimating the regime-switching relationship

Daily PM_{10} levels in the municipality of Bologna from January 1st, 2018 to December 31th, 2021 are obtained by averaging the collected data from three monitoring air quality stations. As a dependent variable, we use the number of accesses to the emergency department either for cardiovascular and respiratory diseases. Therefore, we estimate four models: one for each disease and each municipality. To account for short-term persistence in both the series of the dependent variable and the PM levels, we use seven lags, which correspond to a lagged week. Moreover, we include in the model a dummy variable on Saturdays and Sundays, to consider the weekend effect on the access to the emergency department [5].

We set the number of regimes equal to three, since we are implying that the relationship between air pollution and accesses to emergency department has changed due to the COVID-19 pandemic. This means that we are supposing that there is a regime before the COVID-19, a regime during the lockdown period where air pollution dropped and the effects on health changed accordingly [4], and a regime after the stronger waves of COVID (i.e., between March and June 2020 in Italy).

The results of the three-regime TAR model for each disease and for the total accesses to the emergency department for both diseases are reported in Table 1. Not surprisingly, the estimated first thresholds for all the dependent variables coincide with the beginning of the lockdown period in Italy. In fact, it is reasonable to suppose that with lower air pollution levels, the association between the PM_{10} has changed accordingly. Differently, the second regime ends in different periods between the two diseases. It can be noticed that the lagged PM does not affect the visits to the emergency department for cardiovascular diseases in any of the regimes. Conversely, the lag of the PM_{10} levels of the previous day and from 7 days before, positively affect the number of accesses to the ED for respiratory diseases in the first regime, while there is no effect in the second period. Once again, this may happen mainly due to the different air pollution levels during this period which has led to changes in human behavior, including reductions in traffic and industrial activity. Moreover, there was a relevant change in healthcare services utilization during the COVID period which has caused a drop in ED visits in Northern Italy [3, 8]. Interestingly, the effect of the levels of PM_{10} after 7 days is almost three times higher ($\beta = 0.100$) in the period from August to December 2021. The findings on the estimated model on the total ED visits are almost overlapping with those on accesses for respiratory diseases.

4 Conclusion

In conclusion, the results of the TAR model confirm the strong association between PM_{10} and ED visits and the impact of the social distancing measures enforced by the Italian government on both air pollution and healthcare services utilization. However, the study's findings suggest that particulate matter levels contributed to the exacerbation of pulmonary but not cardiovascular diseases. While evidence suggests that both pulmonary and cardiovascular diseases are associated with exposure to PM_{10}, the impact of air pollution on these two systems is not equal. Pulmonary diseases, such as chronic obstructive pulmonary disease (COPD) and asthma, are particularly sensitive to air pollution. This is because inhaled particles and gases can directly affect the lungs and exacerbate existing respiratory conditions, while for cardiovascular diseases, such as heart disease and stroke, the effects are less direct. This can be attributable to the different pathogenesis between pulmonary and cardiovascular diseases. Pulmonary diseases are primarily caused by inflammation and damage to lung tissue, more directly exposed to air particles, while cardiovascular diseases are caused by damage to the heart and blood vessels, therefore the effects may need time to manifest. The different pathophysiological characteristics of pulmonary and cardiovascular diseases are probably also the basis of the different medium-long-term effects of air pollution on ED visits emerging from the analysis. In particular, this study shows a "rebound" effect of PM_{10} on pulmonary diseases at the end of the pandemic waves; this can be explained by the fact that relaxation of COVID-19 restrictions coincided with decreased face covering use and increased social mixing, and may have caused a rebound in lung diseases such as acute respiratory infections and asthma exacerbations. Therefore, the end of COVID-19 pre-

Table 1 Estimates from the TAR model

Variable	Cardiovascular	Respiratory	Total
Regime 1			
const	**10.059(0.446)**	**1.286 (0.572)**	**3.912 (0.805)**
y_{t-1}	0.019 (0.024)	**0.301 (0.031)**	**0.182 (0.029)**
y_{t-2}	**0.067 (0.023)**	0.009 (0.030)	**0.114 (0.030)**
y_{t-3}	0.031 (0.025)	-0.022 (0.028)	**0.051 (0.031)**
y_{t-4}	**0.046 (0.025)**	**0.072 (0.029)**	**0.085 (0.029)**
y_{t-5}	0.011 (0.021)	0.020 (0.032)	**0.057 (0.030)**
y_{t-6}	-0.009 (0.025)	**0.182 (0.030)**	**0.135 (0.029)**
y_{t-7}	**-0.068 (0.024)**	**0.233 (0.031)**	**0.164 (0.028)**
PM_{t-1}	0.009 (0.014)	**0.036 (0.018)**	**0.036 (0.019)**
PM_{t-2}	-0.030 (0.018)	-0.020 (0.023)	-0.034 (0.025)
PM_{t-3}	0.013 (0.018)	0.024 (0.025)	0.023 (0.026)
PM_{t-4}	-0.003 (0.015)	-0.012 (0.021)	-0.011 (0.026)
PM_{t-5}	0.015 (0.019)	0.002 (0.023)	0.027 (0.026)
PM_{t-6}	0.013 (0.018)	0.017 (0.020)	0.020 (0.025)
PM_{t-7}	0.007 (0.014)	**0.044 (0.018)**	**0.034 (0.019)**
Week	**1.715 (0.602)**	**1.369 (0.816)**	**1.797 (0.456)**
Regime 2			
const	**5.866 (0.401)**	0.448 (0.294)	**2.450 (1.215)**
y_{t-1}	**-0.186 (0.085)**	0.060 (0.053)	**0.137 (0.076)**
y_{t-2}	-0.121 (0.086)	**0.160 (0.052)**	**-0.136 (0.077)**
y_{t-3}	0.005 (0.084)	**0.202 (0.053)**	**0.237 (0.073)**
y_{t-4}	-0.003 (0.085)	**0.146 (0.053)**	0.062 (0.075)
y_{t-5}	-0.086 (0.083)	**0.108 (0.053)**	**0.311 (0.073)**
y_{t-6}	**-0.169 (0.081)**	0.075 (0.052)	-0.054 (0.075)
y_{t-7}	0.053 (0.083)	0.079 (0.052)	**0.170 (0.074)**
PM_{t-1}	0.037 (0.047)	-0.026 (0.024)	-0.069 (0.042)
PM_{t-2}	-0.077 (0.061)	0.041 (0.033)	**0.094 (0.056)**
PM_{t-3}	0.049 (0.064)	0.005 (0.034)	0.006 (0.058)
PM_{t-4}	-0.004 (0.063)	-0.008 (0.034)	0.022 (0.058)
PM_{t-5}	-0.006 (0.062)	0.047 (0.034)	0.001 (0.058)
PM_{t-6}	0.036 (0.061)	-0.014 (0.033)	0.007 (0.057)
PM_{t-7}	-0.010 (0.046)	-0.001 (0.024)	-0.01 (0.043)
Week	**5.873 (0.400)**	0.448 (0.294)	**2.450 (1.215)**
Regime 3			
const	**5.207 (0.242)**	0.074 (0.626)	1.029 (0.879)
y_{t-1}	0.062 (0.038)	**0.284 (0.091)**	**0.243 (0.081)**
y_{t-2}	-0.005 (0.039)	**0.282 (0.094)**	**0.361 (0.084)**
y_{t-3}	0.052 (0.039)	0.012 (0.096)	**0.188 (0.088)**
y_{t-4}	0.022 (0.039)	**0.203 (0.095)**	0.067 (0.090)
y_{t-5}	-0.011 (0.039)	0.004 (0.098)	-0.002 (0.089)
y_{t-6}	-0.049 (0.039)	0.008 (0.097)	-0.075 (0.085)
y_{t-7}	-0.036 (0.038)	0.086 (0.089)	0.023 (0.081)
PM_{t-1}	-0.019 (0.017)	0.070 (0.046)	0.090 (0.056)
PM_{t-2}	0.015 (0.024)	0.010 (0.060)	-0.008 (0.073)
PM_{t-3}	-0.019 (0.025)	-0.079 (0.062)	-0.113 (0.075)
PM_{t-4}	0.008 (0.025)	0.038 (0.062)	0.057 (0.075)
PM_{t-5}	-0.016 (0.025)	-0.077 (0.061)	-0.037 (0.074)
PM_{t-6}	0.014 (0.024)	0.017 (0.060)	0.039 (0.072)
PM_{t-7}	-0.017 (0.017)	**0.100 (0.048)**	**0.102 (0.058)**
Week	**5.207 (0.242)**	0.074 (0.626)	1.029 (0.879)
$\hat{c}_1$	2020/03/21	2020/03/25	2020/03/24
$\hat{c}_2$	2020/08/14	2021/08/16	2021/08/15

Note: Standard errors are reported in parentheses. Underlined values identify a significant effect at a 5% confidence level. `Week` denotes the weekend dummy. $\hat{c}_1$ and $\hat{c}_2$ denotes the dates at which the regime changes from the first to the second regime and from the second to the third regime, respectively.

cautions may have had the unexpected effect of magnifying asthma activity, COPD, and other chronic lung diseases.

Acknowledgements This study has been approved by the Emilia Romagna AVEC research ethics committee board with identifier n° 726/2020/Oss/AOUBo on 03.08.2020. This retrospective study was carried out in conformity with the regulations on data management with the Italian law on privacy (Legislation Decree 196/2003 amended by Legislation Decree 101/2018).

References

1. Birnbaum, H. G., Carley, C. D., Desai, U., Ou, S. and Zuckerman, P.R.: Measuring The Impact Of Air Pollution On Health Care Costs. Health Affairs, 39 (2020)
2. Chen, C., Liu, X., Wang, X., Qu, W., Li, W., and Dong, L.: Effect of air pollution on hospitalization for acute exacerbation of chronic obstructive pulmonary disease, stroke, and myocardial infarction. Environmental Science and Pollution Research, 27: 3384-3400 (2020)
3. Golinelli, D., Campinoti, F., Sanmarchi, F., Rosa, S., Beleffi, M., Farina, G., Tampieri, A., Fantini, M.P. Giostra, F., and Santi, L.: Patterns of Emergency Department visits for acute and chronic diseases during the two pandemic waves in Italy. American Journal of the Emergency Medicine. 50: 22-26 (2021)
4. Granella, F., Reis, L.A., Bosetti, V. and Tavoni, M.: COVID-19 lockdown only partially alleviates health impacts of air pollution in Northern Italy. Environmental Research Letters, 16: 035012 (2021)
5. Guttmann, A., Schull, M. J., Vermeulen, M. J., Stukel, T. A.: Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ. 341, c5340 (2010)
6. He, C., Teräsvirta, T., and González, A.: Testing Parameter Constancy in Stationary Vector Autoregressive Models Against Continuous Change. Econometric Reviews. 28: 225-245 (2008)
7. Kheifets, I.L. and Saikkonen, P.J.: Stationary and ergodicity of vector STAR models. Econometric Reviews. 39: 407-414 (2020)
8. Santi, L., Golinelli, D., Tampieri, A., Farina, G., Greco, M., Rosa, S., Beleffi, M., Biavati, B., Campinoti, F., Guerrini, S., Ferrari, R., Rucci, P., Fantini, M.P. and Giostra, F.: Non-COVID-19 patients in times of pandemic: Emergency department visits, hospitalizations and cause-specific mortality in Northern Italy. Plos One. 16: e0248995 (2021)
9. Solimini, A.G., and Renzi, M.: Association between Air Pollution and Emergency Room Visits for Atrial Fibrillation. International Journal of Environmental Research and Public Health. 14: 661 (2017)
10. Tsay, R.S.: Testing and Modeling Threshold Autoregressive Processes. Journal of the American Statistical Association. 84, 231-240 (2012)
11. Zhou, H., Wang, T., Zhou, F., Liu, Y., Zhao, W., Wang, X., Chen, H., and Cui, Y.: Ambient air pollution and daily hospital admissions for respiratory disease in children in Guiyang, China. Frontiers in Pediatrics. 7: 400 (2019)

Estimating spatially varying Gaussian Graphical Models to unveil relationships among pollutants in the Red River Delta in Vietnam

Stima di modelli grafici gaussiani variabili nello spazio per svelare le relazioni tra inquinanti nel delta del fiume Rosso in Vietnam.

Nicola Pronello, Alex Cucco, Rosaria Ignaccolo and Luigi Ippoliti

Abstract The objective of this study is to tackle the challenge of estimating Gaussian Graphical Models on a non-uniform grid when dealing with a non-stationary multivariate process. Our proposed methodology offers a means to estimate and visually depict the connections among variables within a multivariate process that exhibits spatial variations over a specific grid-based domain structure. In environmental sciences, it is both practical and advantageous to represent variable relationships through estimated graphs. By employing graph analytics, we can unveil intriguing characteristics among the variables of interest and observe how they evolve across the domain. Indeed, in the case study of the Red River Delta, the proposed methodology is useful to map the evolution of the water chemistry along the stream of the river.
Abstract *Lo scopo di questo studio è affrontare la sfida di stimare modelli grafici gaussiani su una griglia non uniforme nel caso di un processo multivariato non stazionario. La proposta metodologica offre uno strumento per stimare e rappresentare graficamente le connessioni tra variabili componenti di un processo multivariato che presenta variazioni spaziali su un dominio costituito da una griglia. Nelle scienze ambientali, è pratico e vantaggioso rappresentare le relazioni tra variabili attraverso grafi stimati. Utilizzando l'analisi dei grafi, possiamo scoprire*

Nicola Pronello
Università degli Studi 'G. d'Annunzio' Chieti - Pescara, Via dei Vestini, 31, Chieti IT, e-mail: nicola.pronello@unich.it

Alex Cucco
Imperial College London, Exhibition Rd, South Kensington, London SW7 2BX UK e-mail: a.cucco20@imperial.ac.uk

Rosaria Ignaccolo
Università degli Studi di Torino, Via Giuseppe Verdi, 8, Torino IT e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
Università degli Studi 'G. d'Annunzio' Chieti - Pescara, Via dei Vestini, 31, Chieti IT e-mail: luigi.ippoliti@unich.it

caratteristiche interessanti tra le variabili di interesse e osservare come si evolvono nel dominio. Infatti, nel caso studio del delta del fiume Rosso, la metodologia proposta risulta utile per mappare l'evoluzione della chimica dell'acqua lungo il corso del fiume.

Key words: Gaussian Graphical Models, Lattice domain, Kernel Smoothing

1 Introduction

Gaussian Graphical Models (GGMs) have gained extensive usage across various research domains for capturing intricate dependencies among variables [10]. In this study, we investigate the potential of employing GGMs on irregular grids, which enables effective modeling of conditional dependence relationships among variables. This approach proves especially advantageous when dealing with a large number of variables relative to the sample size. In comparison to alternative spatial modeling techniques like spatial autoregressive models, kriging models, and Bayesian hierarchical models, GGMs on irregular grids excel in capturing the conditional dependence relationships among variables. They present a robust tool for examining environmental issues characterized by intricate spatial patterns, particularly at the local scale. In summary, GGMs on irregular grids exhibit significant promise as a valuable technique in environmental research. They offer particular benefits in investigating environmental phenomena such as water quality in rivers and lakes, biodiversity in ecosystems, and air pollution in urban areas.

The pollution levels in the Red River Delta in Vietnam serve as our primary example of interest. Specifically, we examine the scenario where various sampling points along the river can be viewed as locations within a stream network, where multiple pollutant levels are observed.

2 Gaussian Graphical Models on irregular grid

We consider a p-dimensional process $\mathbf{Y}^{\mathbf{c}} = \{Y_1^{\mathbf{c}}, \dots, Y_p^{\mathbf{c}}\}$ observed on an irregular grid $\mathscr{L}$ with n (ordered) sites each having coordinates $\mathbf{c} \in \mathbb{R}^k$. We deal with the case of Gaussian variables, that is

$$\mathbf{Y}^{\mathbf{c}} \sim \mathscr{N}(\mu^{\mathbf{c}}, \boldsymbol{\Sigma}^{\mathbf{c}}).$$

In the framework of the graphical models, to evaluate relationships among multiple variables one works with the precision matrix $\boldsymbol{\Theta}^{\mathbf{c}} = (\boldsymbol{\Sigma}^{\mathbf{c}})^{-1} = \{\theta_{j,r}^{\mathbf{c}}\}_{(j,r)=1,\dots,p}$, which provides a measure of the partial correlations.

For each coordinate $\mathbf{c}$, let $G^{\mathbf{c}} = \{V, E^{\mathbf{c}}\}$ be a undirected graph where $V = \{1, \dots, p\}$ represents the set of vertices corresponding to the p random variables and

$E^{\mathbf{c}} \subseteq \{(j,r) \in V \times V,\ j \neq r\}$ represents the set of edges specified by means of

$$(j,r) \notin E^{\mathbf{c}} \text{ if } \theta^{\mathbf{c}}_{j,r} = 0.$$

In the Gaussian case, the condition above is equivalent to conditional independence between the j-th and r-th variables, and by considering the p random variables as vertices of the graph $G^{\mathbf{c}}$ we can say that $\mathbf{Y}^{\mathbf{c}}$ follows a *Gaussian Graphical Model (GGM)*.

To retrieve a network, for every $\mathbf{c}$, among p random variables in a *GGM* we need to estimate $\Theta^{\mathbf{c}}$. Moreover, since we are interested in underlying just the most important connections among variables, we consider a sparse estimator for $\Theta^{\mathbf{c}}$. To this goal, we consider the classical graphical lasso criterion introduced by [1]. The sparsity in the precision matrix is achieved by imposing a lasso penalty, so that the estimated $\widehat{\boldsymbol{\Theta}}^{\mathbf{c}}$ at each $\mathbf{c}$ is obtained by solving an optimization problem:

$$\widehat{\boldsymbol{\Theta}}^{\mathbf{c}} = \underset{\boldsymbol{\Theta}^{\mathbf{c}}}{\operatorname{argmax}} \left(\log\det \boldsymbol{\Theta}^{\mathbf{c}} - \operatorname{trace}(\hat{\boldsymbol{\Sigma}}^{\mathbf{c}} \boldsymbol{\Theta}^{\mathbf{c}}) - \lambda \sum_{j \neq r} |\theta^{\mathbf{c}}_{j,r}| \right), \tag{1}$$

where $\hat{\boldsymbol{\Sigma}}^{\mathbf{c}}$ is an estimate of $\boldsymbol{\Sigma}^{\mathbf{c}}$ (that needs to be obtained) and λ is a nonnegative tuning parameter. The magnitude of the tuning parameter λ controls the sparsity degree of the estimates graph $G^{\mathbf{c}}$, to assess the problem of the choice of this parameter we refer to the information criteria proposed by [1] and [10].

3 Nonparametric estimator of $\boldsymbol{\Sigma}^{\mathbf{c}}$

The reconstruction of sparse network structures by means of Equation (1) is possible only after defining a suitable estimator of $\boldsymbol{\Sigma}^{\mathbf{c}}$ varying on the grid $\mathscr{L}$, that is one for each coordinate $\mathbf{c}$. However, constructing an estimator of $\boldsymbol{\Sigma}^{\mathbf{c}}$ in a nostationarity framework represents a statistical challenge. By considering mean corrected data for the sake of simplicity, a naive estimate of $\boldsymbol{\Sigma}^{\mathbf{c}}$ is given by the raw covariance

$$\boldsymbol{\Sigma}^{*\mathbf{c}} = \left\{\boldsymbol{\Sigma}^{*\mathbf{c}}_{j,r}\right\}_{j,r=1,\ldots,p} = \left\{\mathbf{y}^{\mathbf{c}}{}_{j}\mathbf{y}^{\mathbf{c}T}{}_{r}\right\}_{j,r=1,\ldots,p},$$

that exploits only the information in one datum $\mathbf{y}^{\mathbf{c}}$ and is not full rank since by construction $\operatorname{rank}(\boldsymbol{\Sigma}^{*\mathbf{c}}) = 1$. Then to obtain a reliable estimator of $\boldsymbol{\Sigma}^{\mathbf{c}}$ we propose to exploit information from the neighbours of the unit $\mathbf{c}$ in the grid by means of linear smoothing.

Given n units in the grid $\mathscr{L}$ with coordinates $\mathbf{c}_i$, with $i = 1,\ldots,n$, we consider the class of linear smoother estimators, following [4] and [9], defined by

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}} = \sum_{i=1}^{n} \omega_i(\mathbf{c}) \boldsymbol{\Sigma}^{*\mathbf{c}_i}$$

where the coefficients $\omega_i(\mathbf{c})$ of the linear combination need to be determined. This class includes a wide variety of estimators such as Gaussian process regression estimator, Kernel smoother and local polynomial estimator.

It can be shown that finding $\boldsymbol{\omega}(\mathbf{c}) = (\omega_1(\mathbf{c}), ..., \omega_n(\mathbf{c}))$ is equivalent to solve the following optimization problem:

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{c}} = \mathrm{argmin}_{\boldsymbol{\Sigma}^{\mathbf{c}}} \sum_{i=1}^{n} K(\mathbf{c}_i, \mathbf{c}) d(\boldsymbol{\Sigma}^{\mathbf{c}}, \boldsymbol{\Sigma}^{*\mathbf{c}_i}),$$

where and $K(\cdot,\cdot) : \mathscr{L} \times \mathscr{L} \to \mathbb{R}$ is a kernel function and $d(\cdot,\cdot)$ is a suitable distance between covariance matrices.

3.1 Determining a kernel K from $\mathscr{L}$

We propose to construct kernel weights that are suitable for a grid domain. To properly take into account the nature of the domain where graphs live, we define the kernel from a Laplacian $\mathbf{L}$ resulting from the neighbourhood structure of $\mathscr{L}$. Let $\mathbf{W}$ be an adjacency matrix such that entries are equal to 1 if two units are neighbours, and 0 otherwise, the resulting Laplacian $\mathbf{L}$ is then defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

with $\mathbf{D} = \mathrm{diag}\{\mathbf{W}\mathbf{1}_n\}$.
Finally, as kernel values in the vector $\omega(\mathbf{c})$ for the linear smoother we use the elements of the matrix (following [6])

$$\mathbf{K} = (\mathbf{I} + \gamma\mathbf{L})^{-1},$$

where γ is a smoothing parameter (to be fixed) that controls the Kernel width. In fact, γ tunes the values of the weights and the extension of the neighborhood of each unit in $\mathscr{L}$, such as for $\gamma \to \infty$ one has $\omega_i \to 1/n$.

4 Red river delta dataset

Measurements of physical, chemical, and biological characteristics taken from the waters of 21 river locations spanning the Red River Delta in northern Vietnam are publicly available in the dataset provided by the Environmental Information Data Centre [2]. Monthly data collection occurred between February 2018 and January 2020. The dataset contains observations of a multivariate process since recorded variables include temperature, dissolved oxygen, salinity, pH, turbidity, conductivity, total dissolved solids, dissolved nitrate, dissolved nitrite, dissolved ammonium, total nitrogen, dissolved phosphate, dissolved silicate, alkalinity, dissolved major ions, as well as various chlorophyll and carotenoid pigments derived from riverine seston. Figure 1 shows the region of interest with the 21 sampling locations on the rivers.

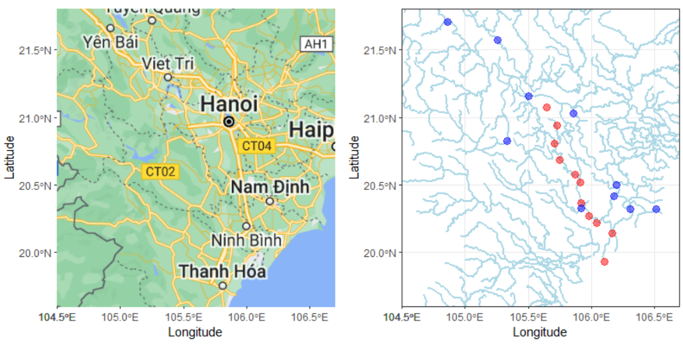


Fig. 1 Left plot: map of the considered region in Northern Vietnam. Right plot: fluvial network with superimposed sampling locations, in red sampling locations along the Day river in blue along the Red river.

The sampling sites' neighboring arrangement can be accurately depicted using a network and this allows to focus on estimating a graph that encompasses multiple pollutants within each node. Estimated graphs show relationships among physical, chemical and biological characteristics that vary with the sampling site, and so along the river. However, considering that a river possesses an inherent directional nature associated with the stream's flow, this characteristic can be effectively integrated into the proposed methodology through the definition of a distinct Laplacian matrix $\mathbf{L}$.

5 Discussion

In this study, we present a non-parametric approach for estimating Gaussian Graphical Models over an irregular grid, denoted as $\mathscr{L}$. Our proposed methodology enables the construction of undirected graphs that capture the relationships between variables of interest, which vary with the coordinates $\mathbf{c}$ of the grid $\mathscr{L}$. An alternative approach involves estimating multiple Gaussian Graphical Models simultaneously, assuming comparable network structures across distinct sub-populations [8]. However, in our study, we relax this assumption and allow for distinct estimated covariances, leading to differences in the resulting networks. This approach aligns with the growing interest in methodologies of this nature within the field of functional data analysis, as evidenced by related works [7] and [3]. Notably, this proposal has already been adapted to handle spatio-temporal functional data in [5]. An interesting avenue for future research includes selecting the optimal smoothing parameter λ and exploring novel strategies to ensure positive definiteness in the estimates of $\boldsymbol{\Sigma}_{\mathbf{c}}$. Furthermore, an extended version of this work will discuss results related to pollution in the Red River Delta in Vietnam, demonstrating the practical relevance of the proposed methodology.

References

1. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. Biostatistics 9(3), 432-441 (2008)
2. McGowan, S., Salgado, J.: Water chemistry from the Red River Delta, Vietnam, 2018 to 2020. NERC EDS Environmental Information Data Centre (Dataset) https://doi.org/10.5285/7e35e760-0ca2-4290-8970-464ead03055d (2022)
3. Moysidis, I., Li, B.: Joint Functional Gaussian Graphical Models. doi: 10.48550/arXiv.2110.06653 (2021)
4. Petersen, A., Deoni, S., Müller H.G.: Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional co-evolution of myelination in the developing brain. The Annals of Applied Statistics 13(1), 393-419 (2019)
5. Pronello, N., del Gobbo, E., Fontanella, L., Ignaccolo, R., Ippoliti, L., Fontanella, S.: Functional Graphical Models to map Brexit debate on Twitter. In: Proceedings of SIS2023, Pearson (in press, 2023)
6. Smola, A.J., Kondor, R.: Kernels and Regularization on Graphs. In: Schölkopf, B., Warmuth, M.K. (eds.) Learning Theory and Kernel Machines, pp. 144-158. Springer Berlin, Heidelberg (2003)
7. Solea, E., Li, B.: Copula Gaussian Graphical Models for Functional Data. Journal of the American Statistical Association 117(538), 781-793 (2022)
8. Tsai, K., Koyejo, O., Kolar, M.: Joint Gaussian graphical model estimation: A survey. WIREs Computational Statistics 14(6), e1582 (2002)
9. Yin, J., Geng, Z., Li, R., Wang, H.: Nonparametric covariance model. Statistica Sinica 20(1), 469-479 (2010)
10. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika 94(1), 19-35 (2007)

Solicited Session SS2 - *Statistics in sports*

Session supported by BDsports, ISI Groups on Sports and Math&Sport organized by Marica Manisera and Rodolfo Metulini
Chair: Rodolfo Metulini

1. *Clustering Athlete Performances in Track and Field Sports* (Argiento R., Colombi A., Modotti L. and Montagna S.)
2. *A Cross-Country Analysis of Engagement in Physical Activity and Sport Practice Learnt from Eurobarometer Survey Data* (Simone R.)
3. *Strong eras in male professional tennis* (Breznik K., Candila V., Milekhina A. and Restaino M.)
4. *NonParametric Combination method for data analytics in basketball matches* (Barzizza E., Biasetton N., Ceccato R., Disegna M. and Vezzosi G.)

Clustering Athlete Performances in Track and Field Sports

Clustering della Performance degli Atleti di Atletica Leggera

Raffaele Argiento, Alessandro Colombi, Lorenzo Modotti and Silvia Montagna

Abstract This study aims to cluster track and field athletes based on their average seasonal performance. Athletes' performance measurements are treated as random perturbations of an underlying individual step function with season-specific random intercepts. A hierarchical Dirichlet process is used as a nonparametric prior to induce clustering of the observations across seasons and athletes. By linking clusters across seasons, similarities and differences in performance are identified. Using a real-world longitudinal shot put data set, the method is illustrated.

Abstract *L'obiettivo di questo lavoro consiste nel raggruppare atleti di atletica leggera in base alla loro performance stagionale. Le misurazioni di performance degli atleti sono trattate come perturbazioni casuali di una funzione a tratti individuale con intercette casuali stagionali. Si usa un processo nonparametrico di Dirichlet gerarchico a priori per raggruppare le osservazioni tra stagioni e atleti. Unendo i cluster inter-stagionali, si identificano somiglianze e differenze di performance. Il metodo è illustrato utilizzando un dataset reale longitudinale di lancio del peso.*

Key words: Hierarchical Dirichlet process, Longitudinal data analysis, Nonparametric Bayesian modelling, Sports analytics

---

Raffaele Argiento
Università degli Studi di Bergamo, Via dei Caniana 2, Bergamo, e-mail: `raffaele.argiento@unibg.it`

Alessandro Colombi
Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milano, e-mail: `a.colombi10@campus.unimib.it`

Lorenzo Modotti
Columbia University, 665 West 130th Street, New York, e-mail: `lmodotti27@gsb.columbia.edu`

Silvia Montagna
Università degli Studi di Torino, C.so Unione Sovietica 218/bis, Torino, e-mail: `silvia.montagna@unito.it`

1 Introduction

Sports analytics employ data and quantitative methods to measure and analyse athletes and teams performance in professional sports, typically with predictive purposes. Specifically, the use of hierarchical Bayesian methods is gaining popularity in sports analytics, as they allow information sharing across time and athletes [1]. The applied motivation for this work comes from a longitudinal dataset of professional shot put athletes, whose performance was measured at each event they competed at, along with their demographic information. This work aims to cluster athletes with a similar average performance within a season and study the evolution in athletes' performance throughout their careers. We rely on a hierarchical Dirichlet process [3] model to generate ties across seasons and athletes. Within each season, athletes sharing the same latent model parameters belong to the same cluster and thus have the same seasonal average performance. The latent parameters are also shared across seasons, thereby also allowing for the global clustering of athletes.

2 Shot put dataset

Shot put is a track and field event in which athletes have to *put* (throw) a heavy spherical ball, the *shot*, as far as they can. Our data set[1] consists of measurements (the throw lengths or marks) recorded during the professional shot put competitions from 1997 to 2016. During these 19 seasons, 41,033 measurements on 653 athletes have been recorded. For each athlete taking part in a competition, a record comprising the mark, the finishing position, information regarding the competition itself and the shot putter is stored. Relevant covariates are the event date, the environment (indoor or outdoor event), sex and age of the athlete.

Since the aim is to model the mean seasonal performance for each shot putter, the season number associated with each observation corresponds to the number of seasons elapsed since the beginning of the athlete's career, also counting seasons in which the athlete did not compete. Thus, season 1 consists of measurements observed during each athlete's first season of activity, regardless of the calendar year in which these have been recorded. In this way, observations result in being grouped in seasons that reflect the athletes' years of experience. Figure 1 depicts the evolution in performance throughout the athlete's career for two randomly selected shot putters from the dataset. Different athletes take part in different competitions, and the length and profile of their performance careers vary. Clearly, performance is expected to vary throughout an athlete's career, but, as evident from this plot, it is rather steady within each season. This suggests that, despite being rough approxima-

[1] Available at `https://github.com/PatricDolmeta/Bayesian-GARCH-Modeling-of-Functional-Sports-Data/blob/main/COMPLETE_DATA.txt`.

tions, the seasonal mean performances capture the essential features of the athletes' careers.

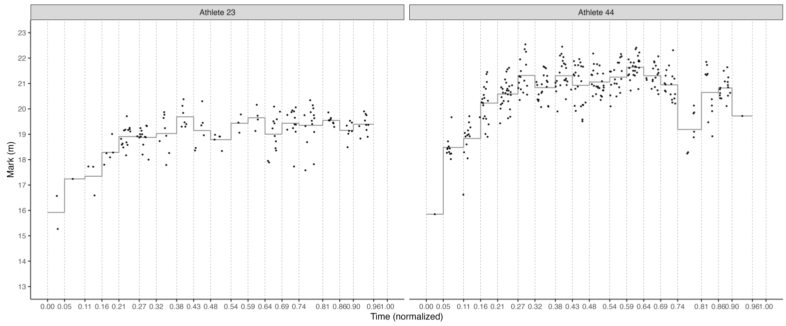


Fig. 1 Shot put measurements collected throughout an athlete's career for two randomly selected athletes with empirical seasonal mean performance (solid grey line). Dotted grey lines delimit seasons.

3 Proposed methodology, prior elicitation and posterior inference

We suppose that n_s athletes compete in season s, with $s \in \{1,\dots,S\}$ and $S = 19$ in our application. Each active athlete i in season s, $i \in \{1,\dots,n_s\}$, takes part in N_{si} events. At each event $j \in \{1,\dots,N_{si}\}$ the athlete's mark Y_{sij} is measured. For simplicity, we rescale time so that measurements are collected at $t_{sij} \in [0,1]$. Moreover, a set of d covariates is available, $\boldsymbol{x}_{sij} := \boldsymbol{x}_i(t_{sij}) = \left[x_{sij}^{(1)},\dots,x_{sij}^{(d)}\right]^\top$. Assuming that observations are noisy measurements of an underlying athlete-specific function, a general model for these data is

$$Y_{sij} = g_i(t_{sij}; x_{sij}) + \varepsilon_{sij}, \tag{1}$$

with $\varepsilon_{sij} \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \frac{1}{\tau_{si}}\right)$, where τ_{si} denotes the precision of the distribution. Suppose that the athlete-specific functions are piecewise constant that is

$$g_i(t_{sij}) = \sum_{s=1}^{S} \mu_{si} \mathbb{1}_{[t^{(s)},t^{(s+1)})}(t_{sij}) + \boldsymbol{x}_i(t_{sij})^\top \boldsymbol{\beta}_s, \tag{2}$$

where μ_{si} is a season-specific random intercept, $t^{(s)} := \min_{i,j} t_{sij}$ is the beginning of each season s, $t^{(S+1)} := \max_{i,j} t_{Sij}$ is the end of season S, and $\boldsymbol{\beta}_s$ is a d-dimensional vector of regression parameters, shared among all the athletes in season s. Therefore, within each season s, the athlete's observations are normally distributed as $Y_{sij} \mid \mu_{si}, \tau_{si}, \boldsymbol{\beta}_s; \boldsymbol{x}_{sij} \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu_{si} + \boldsymbol{x}_{sij}^\top \boldsymbol{\beta}_s, \frac{1}{\tau_{si}}\right)$. Let $\theta_{si} := (\mu_{si}, \tau_{si})$; the sampling model

(1)-(2) is completed with the hierarchical Dirichlet prior [4] for θ_{si}, i.e.,

$$\theta_{si} \mid P_s \overset{\text{iid}}{\sim} P_s; \quad P_s \mid \alpha_0, P_0 \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha_0 P_0); \quad P_0 \mid \alpha, H \sim \mathrm{DP}(\alpha H),$$

where $\mathrm{DP}(\alpha H)$ denotes a Dirichlet process with concentration parameter α and base distribution H. Assuming conjugacy, the baseline distribution is a Normal-Gamma $H \sim \mathrm{NG}\left(\mu_0, p_0, \frac{\nu_0}{2}, \frac{\nu_0}{2}\xi_0^2\right)$. Finally, we assume $\boldsymbol{\beta}_s \overset{\text{iid}}{\sim} \mathrm{N}_d\left(\boldsymbol{\beta}_0, \Sigma_0\right)$, where N_d denotes the d-dimensional Normal distribution and Σ_0 is its variance-covariance matrix, $\alpha \sim \mathrm{Gamma}(a,b)$, and $\alpha_0 \sim \mathrm{Gamma}(a_0, b_0)$. For posterior computation, we exploit the Chinese restaurant franchise representation of the hierarchical Dirichlet process [3], which allows us to design a Markov chain Monte Carlo sampling scheme for the model above. In this metaphor, customers are represented by parameters θ_{si} and seasons are represented as restaurants. Customers are clustered into tables within each restaurant, and these tables are further clustered into dishes. Observations are clustered across restaurants at the second level of the clustering process restaurants when dishes are associated with tables. One can think that the first customer sitting at each table chooses a dish from a common menu, which is then shared by all subsequent customers at that table. As usual in model-based clustering, we say that two observations, say (s,i) and (l,j), belong to the same cluster if $\theta_{si} = \theta_{lj}$. Under the hierarchical Dirichlet process, the values of the parameters are shared within the seasons, e.g $\theta_{si} = \theta_{sj}$, as well as between the seasons, leading to two-levels, model-based, clustering of the athletes.

4 Results

In this Section, we present the results obtained on the shot put dataset. The hyperparameters of the baseline distribution H were chosen setting $\mu_0 := \bar{\boldsymbol{y}} = 0.0$, $p_0 := \frac{1}{\mathrm{range}(\boldsymbol{y})^2} = 0.002250395$, $\nu_0 := 2$, $\xi_0^2 := 0.5$, where $\boldsymbol{y}$ denotes the whole set of observations across athletes and seasons. We included three covariates: sex, age (centered around the global mean) and environment. The hyperparameters of the prior distribution for the multiple regression parameters were set to $\boldsymbol{\beta}_0 := [0,0,0]^\top$, $\Sigma_0 := I$. The hyperparameters of the prior distributions for the concentration parameters α and α_0 were set to $a = a_0 := 1$, $b = b_0 := 8$. After a burn-in of $10{,}000$ iterations, $50{,}000$ samples have been retained. We examined the estimated posterior distributions of the regression parameters. For the age covariate, it appears that at the beginning of their career, athletes who are older than the mean tend to perform better, probably due to the different stage of physical development. This difference in performance vanishes during their mid-career, but it may become relevant again in the last part. Due to page constraints, we omit other results on covariates from this work as less insightful.

Concerning clustering, 12 global clusters have been found. The estimated locations μ are well dispersed across the range of the observations, while most of the precisions τ are concentrated around 2 and 4. Regarding seasonal clustering, the active estimated components within each season are reported in Table 1: the smallest number of clusters per season is 3 (seasons 18 and 19), while the largest is 9 (seasons 7, 8 and 10). Generally, the estimated components are shared across different seasons, as desired. The resulting clustering is illustrated in Figure 2. Despite obser-

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•			•	
3	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•
4	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
5	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
6	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		
7			•				•					•		•					
8				•	•	•	•		•										
9					•		•	•	•	•	•	•	•	•	•				•
10						•		•									•		
11								•		•	•		•		•				
12										•									
Total	6	5	7	7	8	8	9	9	8	9	8	8	8	8	8	4	6	3	3

Table 1 Active estimated components (rows) per season (columns).

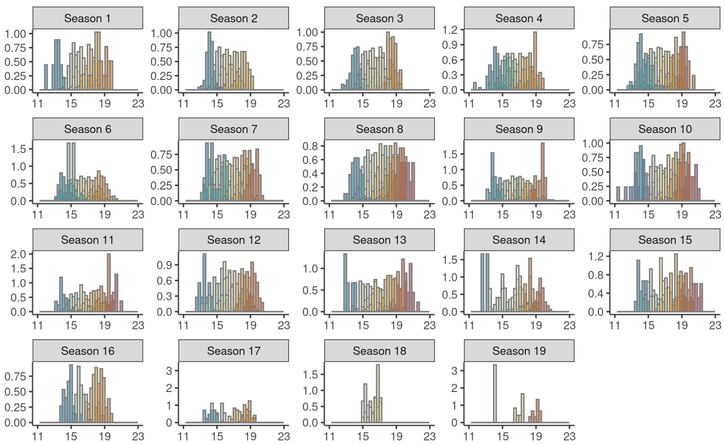


Fig. 2 Histograms of the shot put observations for all athletes split in the 15 seasons and coloured according to the estimated clusters' memberships.

vations are numerous and overlapping within each season, the model has been able to retrieve a reasonable clustering. Figure 3 provides evidence for the goodness of fit at the individual level. For most seasons, the two athletes obtained similar results and indeed they have been clustered together. In seasons 8, 12 and 13 the two shot putters have been assigned to different clusters. By inspecting the plot it seems that athlete 55 performed slightly better than 94 in these seasons, on average. Overall, the model fits the data well. Both local and global clustering of the athletes are reasonable, and the estimated number of clusters allows to model precisely the mean seasonal performance of the shot putters, without overfitting the data.

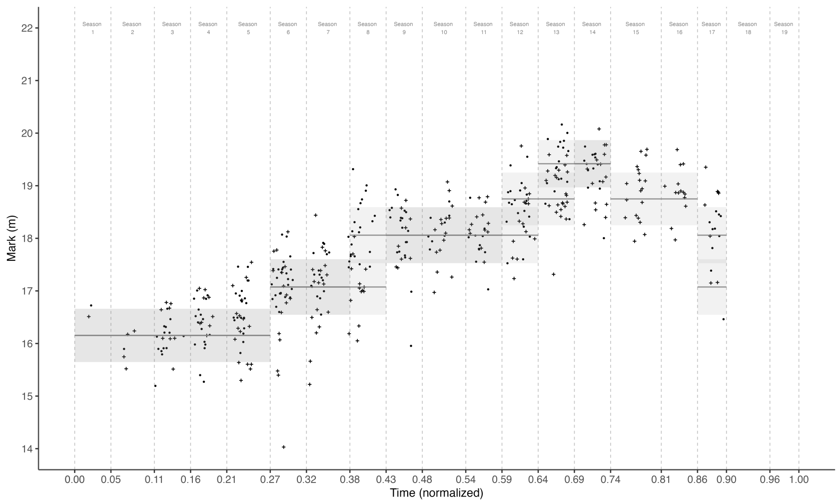


Fig. 3 Shot put measurements for athletes 55 (dot) and 94 (plus) with estimated mean (solid line) of the assigned clusters. The shaded areas comprise values within one estimated standard deviation from the estimated mean.

5 Conclusions

In this work, local and global clustering of longitudinal sports data have been investigated by employing a hierarchical Dirichlet process mixture model. Several extensions of this model may be investigated, for example the hierarchical Dirichlet process could be replaced by more general nonparametric priors. Further, the entire longitudinal curve could be used to induce clustering instead as in [2].

References

1. Baio G., Blangiardo M.: Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics. 37, 253–264 (2010)
2. Page, G.L., Quintana, F.A.: Predictions Based on the Clustering of Heterogeneous Functions via Shape and Subject-Specific Covariates. Bayesian Analysis 10, 379–410 (2015)
3. Teh Y.W., Jordan M.I., Beal M.J., Blei D.M.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association. 101, 1566–1581 (2006)
4. Teh Y.W., Jordan M.I.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N.L., Holmes, C., Müller, P., Walker, S.G. (eds.) Bayesian Nonparametrics, pp. 158–207. Cambridge University Press (2010)

A Cross-Country Analysis of Engagement in Physical Activity and Sport Practice Learnt from Eurobarometer Survey Data

Analisi comparata dell'engagement nell'attività fisica e nella pratica sportiva sulla base dell'indagine Eurobarometro sullo sport

Rosaria Simone

Abstract On the basis of the special Eurobarometer surveys on Sport carried out in 2017 and in 2022, the contribution presents a cross-country analysis of the engagement in physical activity and in sport practice. Specifically, the traits under examine are expected to be heterogeneous within the population, yet being associated to a universal structure foreseeing a cluster of inactive people, a cluster of people with intermediate engagement, and possibly a further group of very active people. For this reason, finite mixtures of discretized Beta distributions provide an insightful framework to parameterize the extent by which inactivity and activity hold and evolve over time due to their flexible shapes.

Abstract *L'obiettivo del presente lavoro è di condurre un confronto tra paesi dell'engagement alla pratica sportiva e della propensione all'inattività fisica sulla base delle indagini Eurobarometro sullo sport condotte nel 2017 e nel 2022. A tal fine, le misture di distribuzioni Beta discretizzate per dati ordinali risultano indicate per parametrizzare in modo flessibile i tratti in esame.*

Key words: Sport Practice, Engagement in Physical Activity, Survey Data, Finite Mixture Modelling, Discretized Beta Distribution

1 Motivating data

From the first 2002 edition, the European Commission carries out periodically the special Eurobarometer survey on sport and physical activity[1] in order to assess participation in sport and engagement in physical activity throughout countries, as well as to understand motivation and barriers. With reference to the 2017 edition, it turns out that the level of participation overall decreased from the previous 2013

Rosaria Simone
University of Naples Federico II, Italy, e-mail: rosaria.simone@unina.it

[1] https://europa.eu/eurobarometer/surveys/detail/2164.

survey, with the exception of few countries (Belgium, Luxembourg, Cyprus, Malta, Finland and Bulgaria) and despite all promotion efforts made by policy makers. The latest results issued in 2022 revealed that sport engagement is remained almost unchanged from the previous wave. These statements are based on the responses to the following two questions, which will be the focus of the forthcoming analysis:

QB1: How often do you exercise or play sport?[2]
QB2: And how often do you engage in other physical activity such as cycling from one place to another, dancing, gardening, etc.?[3]

with answers collected on a (ordered) scale with $m = 6$ categories: 1 = *Never*, 2 = *Less often*, 3 = *1 to 3 times per month*, 4 = *1 to 2 times per week*, 5 = *3 to 4 times a week*, 6 = *5 times a week or more*. As a matter of fact, statements on these data (aggregated per countries) are limited to comparisons of relative frequencies. The goal of the contribution is to boost the potential of such data in supporting policy-makers to assess the degree by which inactivity and engagement holds and how these traits evolve over time with a suitable parameterization within a statistical modelling framework. Response distributions to QB1 and QB2 present indeed some structural features, composed of a cluster of responses anchored to the *Never* category, and an intermediate cluster of responses floating between the endpoints of the response scale. Ideally, a third cluster of responses with modal value at the last category would correspond to people with a strongly active lifestyle and engagement in physical activity. For the sake of illustration, Figure 1 displays the response distributions for QB1 and QB2 for Italy, in 2017 (top) and in 2022 (bottom). Similar patterns characterize the distributions of responses to QB1 and QB2 for all countries. As we see, comparisons made solely on frequencies of given categories can be belittling. Thus, in this contribution we present a parametric setting in which response features can be directly modelled.

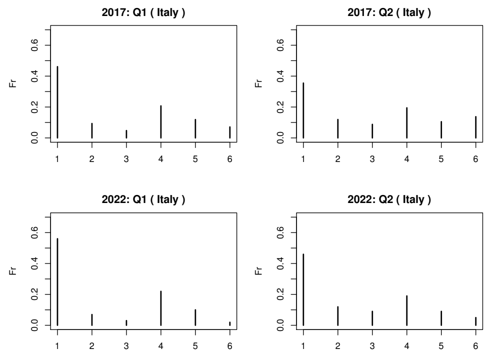
2 Methodological framework

When comparisons of the distributions of an ordered response R over independent groups or times is of interest, statistical models on the discrete support that parameterize relevant features of the distribution provide an appropriate methodological framework. Assuming the Binomial model as a benchmark, the class of CUB mixture models provides specification of latent sentiment, heterogeneity, inflated frequencies and possible over dispersion in a versatile and parsimonious way [8]. However, this class of models cannot account for all possible shapes, and in particular it is not always well-suited in case of U-shaped distributions or for

[2] By "exercise" we mean any form of physical activity which you do in a sport context or sport-related setting, such as swimming, training in a fitness centre or a sport club, running in the park.

[3] By "other physical activity" we mean physical activity for recreational or non-sport-related reasons.

Fig. 1 Frequency distributions of responses to QB1 and QB2 for Italy in 2017 (top) and in 2022 (bottom)



circumstances where three different clusters of response exist, as in case of polarization towards the extremes of the scale and floatation in within. In this setting, a more general and flexible option can be defined on the basis of a discretized Beta model [4, 11]. Then, the OFS finite mixtures of discretized Beta Distributions [9] is defined by[4]:

$$Pr(R=r|\theta) = \delta_1\ db(r;\alpha_1,1)+\delta_2\ db(r;\alpha_2,\beta_2)+\delta_3\,db(r;1,\beta_3), \quad r=1,\ldots,m \tag{1}$$

where:

- a discretized Beta distribution with parameters $\alpha_1 \in (0,1)$ and $\beta_1 = 1$ is specified to account for polarization towards the lowest category (*opponents' pole*);
- a discretized Beta distribution with parameters $\alpha_2, \beta_2 > 1$ is specified to account for floatation between the scale endpoints (with parameters satisfying a constraint to have a modal value at an inner category). Asymmetry and kurtosis adjusted for skewness [2] for this component are informative about the location and the concentration of the unpolarized responses;
- a discretized Beta distribution with parameters $\alpha_3 = 1$ and $\beta_3 \in (0,1)$ is specified to account for polarization towards the top of the scale (*supporters' pole*).

The general model given in (1) is referred to an OFS_{111} mixture, with pedix indicating that all three components are specified. If case the supporter pole is not specified, the corresponding OFS_{110} mixture provides a unified parameterization of the extent by which polarization towards the lowest extreme of the scale holds in terms of parameter α_1 (with lower values indicating a stronger polarization), whereas its mixture weight δ_1 gives the size of this pole. Parameters α_2, β_2 of the floatation component, instead, provide information on unpolarized responses, and depending on α_2 being equal, greater or lower than β_2, on the location of this cluster (floating towards either one of the pole, or corresponding to *neutral* responses if case of symmetry $\alpha_2 = \beta_2$) and its concentration. An OFS_{101} model, instead, is adequate to fit U-shaped distributions. Maximum likelihood estimation

[4] The acronym OFS stands for Opponent-Floatation-Support.

can be performed with an EM algorithm. Then, bootstrap or numerical methods allow to derive standard errors of the parameters' estimates. Notice that the chosen terminology for OFS clusters is inherited from the political context, as these three clusters are structurally present in case the probability to vote a given party is surveyed. With respect to the chosen questions within the Eurobarometer Sport survey, α_1 is the parameter indicating the strength of inactivity, whereas δ_1 measures the size of the cluster.
Specifying a CUB with shelter at the first category is also a possibility that entails good fitting performance for the data at hand, beyond being more parsimonious in terms of model complexity [6]. However, this model would be tantamount to saying that the cluster of inactive people reduces only to those responding 'Never'. With OFS_{110}, instead, the cluster corresponding to modal value in this category shapes response attitude towards the inactivity pole: the *shelter* effect is contemplated as a limit case if $\alpha_1 \to 0$, so that the discretized Beta model with parameters $(\alpha_1, 1)$ tends to a degenerate distribution with mass at the first category. Similar comments hold for zero-inflated models usually used to analyse sport participation [1, 3]: in addition, classical zero-inflated modelling requires the specification of covariates, which could undermine the effectiveness of comparisons across countries and time.

2.1 Surrogate residuals' approach

Usually, model selection is driven on the basis of information criteria and other indicators of both fitting and predictive performance. This stage should be supplemented by some investigation about the correct specification of the model. For instance, a necessary condition for a model to be correctly specified is that the surrogate residuals for the estimated model are uniformly distributed [7]. This check has been sucessfully run for the present analysis, with the same logics adopted in [10] to perform local diagnostics of binomial trees for rating responses.

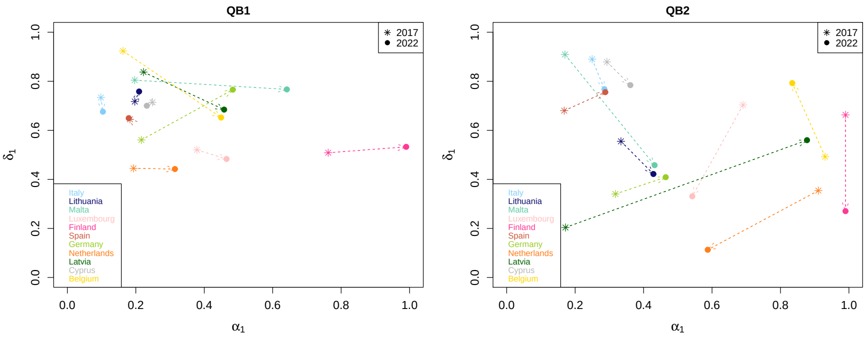
3 Some Results on Eurobarometer Survey on Sport Practice and Physical Activity

A visualization of OFS_{110} polarization estimates is provided in Figure 2 for selected countries. As an instance of the interpretative extent of the model, it follows that:

- For Italy, sport inactivity and lack in engagement in physical activity are basically stable between 2017 and 2022: it seems however that the size of the cluster of disengagement (δ_1) tends to reduce.
- For Belgium, the size of the inactive cluster reduced for sport practice but increased for physical activity. The strength of inactivity, instead, decreased from

2017 to 2022 for sport practice as α_1 increases, whereas the intensity of physical inactivity stayed about the same.

Fig. 2 Visualization of estimates of polarization parameters (δ_1, α_1) for OFS_{110} model fitted to responses to QB1 and QB2 in 2017 and 2022 for selected countries

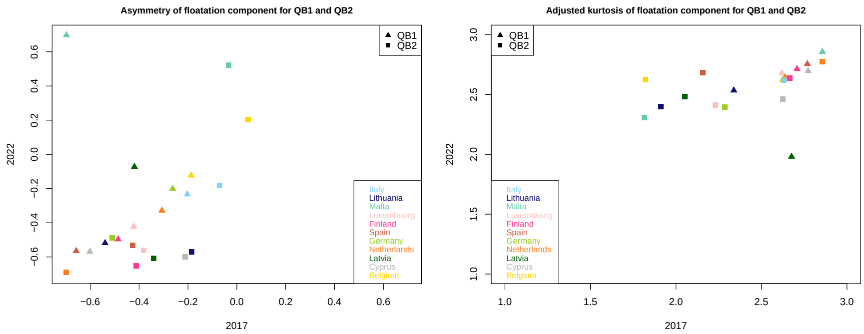


For floatation parameters (α_2, β_2), instead, Figure 3 presents a possible visualization in terms of asymmetry and kurtosis adjusted for skewness for the estimated component. In this regard, remarkably asymmetry turns to be positive from 2017 to 2022 for Malta for both sport exercise and physical activity, and for Belgium for physical activity only, meaning that there has been a shift towards the activity pole. Overall, negative asymmetry seems to reduce slightly for all countries and for both sport practice and physical activity. With respect to kurtosis, instead, it can be stated that it remains overall stable from 2017 to 2022 for the intermediate cluster floating between the two poles of sport practice, whereas it slightly increases for what concerns physical activity. The joint analysis of asymmetry and kurtosis adjusted for skewness allows to state that engagement in physical activity is slowly growing, yet faster than participation in sports practice.

4 Concluding remarks and outline for future works

The contribution aims at showing the outreach of a finite mixture of discretized Beta distributions for survey data concerning sport participation and engagement in physical activity. As all statistical models on the discrete support that parameterize relevant features of the distribution, the chosen setting can be successfully applied to compare the extent by which polarization towards inactivity and engagement hold, as well as floatation between the response extremes for independent groups of respondents (to investigate gender and age differences, for instance).

Fig. 3 Visualization of asymmetry and kurtosis adjusted for skewness over time for the estimated floatation component of OFS_{110} model fitted to responses to QB1 and QB2 for selected countries



References

1. Bauman, A., Sallis, J., Dzewaltowskki, D., Owen, N.: Toward a better understanding of the influences on physical activity. American Journal of Preventive Medicine, 23(2 Suppl.), 5-14 (2002)
2. Blest, D.C.: A new measure of kurtosis adjusted for skewness. Australian and New Zealand Journal of Statistics, 45(2), 175-179 (2003)
3. Downward, P. Lera-Lopèz, F., Rasciute, S.: The Zero-Inflated ordered probit approach to modelling sports participation. Economic Modelling, 28, 2469-2477 (2011)
4. Fasola S., Sciandra M.: New Flexible Probability Distributions for Ranking Data. In: Morlini I., Minerva T, Vichi M (eds.), *Advances in Statistical Models for Data Analysis*. Springer, Springer-Verlag, pp.117-124 (2015)
5. Hovemann, G., Wicker, P.: Determinants of sport participation in the European Union. European Journal for Sport and Society, 6(1), 51–59 (2009)
6. Iannario, M., Simone, R.: Zero-inflated ordinal data models with application to sport (in)activity. In: A. Abbruzzo, E. Brentari, M. Chiodi and D. Piacentino (Eds), Book of Short Papers SIS 2018, pp. 88-96, Pearson (2018)
7. Liu, D., Zhang, H.: Residuals and diagnostics for ordinal regression models: A surrogate approach. Journal of the American Statistical Association, 113(522), 845-854 (2018)
8. Piccolo, D., Simone, R: The class of CUB models: statistical foundations, inferential issues and empirical evidence. Statistical Methods & Applications, 28, 389-435 (2019)
9. Simone, R.: On Finite Mixtures of Discretized Beta Model for Ordered Responses. TEST. 31, 828-855 (2022)
10. Simone, R.: Uncertainty diagnostics of Binomial Regression Trees for Ordered Rating Data. Journal of Classification, doi:10.1007/s00357-022-09429-5 (2023)
11. Ursino M., Gasparini M.: A new parsimonious model for ordinal longitudinal data with application to subjective evaluation of a gastrointestinal disease. Statistical Methods in Medical Research, 27(5), 1376–1393 (2018)

Strong eras in male professional tennis

Ere forti nel tennis professionistico maschile

Kristijan Breznik, Vincenzo Candila, Antonina Milekhina and Marialuisa Restaino

Abstract This paper aims to define the "strong" and "weak" eras in male professional tennis competitions. A strong era is defined as the subsequent periods where at least two top players, ranked at the first two positions of the official ranking, are co-moving jointly. On the contrary, during a weak era, there is a single player dominating all the others. From a statistical point of view, the strong era consists of subsequent periods when the Association of Tennis Professionals (ATP) points of at least two top players are cointegrated. According to a sample of over thirty years of matches, the cointegration analysis here employed detected several strong eras, which include, for instance, the Big Four (Djoković, Federer, Murray, and Nadal).

Abstract *Questo lavoro mira a definite le ere "deboli" e "forti" nelle competizioni del tennis professionistico maschile. Un'era forte è definita come un insieme di periodi susseguenti dove almeno due top players, classificati nelle prime due posizioni del ranking, si muovono congiuntamente (ovvero, so co-integrati). Al contrario, durante un'era debole, c'è un solo giocatore che domina su tutti gli altri. Dal punto di vista statistico, un'era forte è costituita da periodi consecutivi i cui i punti dell'Association of Tennis Professionals (ATP) di almeno due giocatori sono cointegrati. Considerando un sample di oltre 30 anni, l'analisi di cointegrazione ha trovato diverse ere forti che includono, per esempio, i cosiddetti Big Four (Djoković, Federer, Murray e Nadal).*

Key words: Tennis, cointegration, weak and strong eras, male competition

Kristijan Breznik
International School for Social and Business Studies, Celje, Slovenia, e-mail: `kristijan.breznik@mfdps.si`

Vincenzo Candila
University of Salerno, Fisciano, Italy, e-mail: `vcandila@unisa.it`

Antonina Milekhina
Higher School of Economics, Moscow, Russia, e-mail: `antonina.milekhina@yandex.ru`

Marialuisa Restaino
University of Salerno, Fisciano, Italy, e-mail: `mlrestaino@unisa.it`

1 Introduction

In the last few decades, tennis has become increasingly popular, particularly with the epic matches among the big four men's players - Federer, Nadal, Murray, and Djoković (intentionally in order of their births). The female competition also provided some big rivalries at the beginning of the century, with the Williams sisters probably the most prominent example among them. However, in recent years, women's tennis has been marked by much more variability, with a lot of players alternating in and out of the top ten. It also seems that female competition has been in the shadow of male players.

Tennis research undertakes analysis from various angles, even if we do not consider research the game of tennis itself, such as strokes, court movement, and psycho-physical condition. There are several methods developed for measuring the strength of tennis players. First, the official Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) ranking procedure incorporates the results of the past 52 weeks (with some minor exceptions) [3, 12]. This official ranking is used to seed players in the tournaments, which can be extremely important since most tennis tournaments are played to elimination, meaning that the player who loses the match is eliminated (even in the early phase of the tournament). However, academic research has spurred alternative methods to measure the strength of tennis players, such as PageRank for individual competition [11] or doubles [6], standard and weighted ELO systems [1], paired comparison models [2, 9] and a survey on all of them [10].

The best tennis players sometimes prove their superiority during so-called dominant seasons, such as Djoković in 2011 and a decade later in 2021, or Federer in 2006. However, in this sense, they probably cannot compare to the achievement of Steffi Graf, who won the Golden Slam in 1988, when she won all four Grand Slams and the Olympic games.

If there are different contributions to defining the greatest player of all time [4, 5], the definition of a restricted set of players which dominate all the others jointly for a delineated period has not received adequate attention. This paper aims to fill this gap. In particular, we aim at identifying "strong" and "weak" periods/eras, in male tennis, from a statistical point of view. A "strong" period of tennis should include the dominance of top players related to stability/quality of competition (strong players winning Grand Slams and top ten players do not vary too much), records, and achievements (if players are breaking records and achieving milestones at a higher rate than usual), the longevity of top players (how long the top players can maintain their dominance).

To aim our research question of identifying strong and weak periods in tennis, we adopt the following definitions:

Definition 1. A strong period is when at least two top players, ranked at the first two positions of the official tennis ranking, are co-moving jointly. A strong era is a time interval consisting of at least two subsequent strong periods.

Definition 2. A weak period is when no top players, ranked at the first positions of the official tennis ranking, are co-moving jointly. A weak era is a time interval consisting of at least two subsequent weak periods.

Therefore, during a strong era, a restricted set of players (composed of at least two top players) dominates all the others for a delineated period of time. On the contrary, during a weak era, there is much more uncertainty in the outcomes of matches or, alternatively, a single player dominates all the others. The previous definitions are based on the cointegration analysis of the ATP points. The main reason to employ the cointegration analysis is that the ATP points are non-stationary, and in particular, they are integrated in the same order. Moreover, if Definition 1 holds, a given linear combination of the ATP points for a restricted set of players produces a stationary series.

The rest of the paper is organized as follows. Section 2 describes the data and illustrates the method applied. Sections 3 is devoted to the empirical application. Conclusions follow.

2 Data and Methods

The data used for current research includes information on the ATP tour. More precisely, the dataset provides information on the week-by-week occupier of each ranking place on the ATP tour and the related points. The data were freely available online at the GitHub project by Jeff Sackmann. In this study, data from the start of 1990 until March 2023 were used. The raw data information for the top ten players for each week was used for further analysis and identification of strong and weak periods, as described below. As we build our research around ranking points, it is worth mentioning that the ATP ranking system changed in 2009, resulting in a big difference in points for the periods before and after 2009. We took the change into account in our research by running the cointegration analysis as described below, where T_{in} represents the length of the rolling window used:

1. Regress ATP points of the first player on the ATP points of the second and third players from the period $t = 1 + j$ to $t = T_{in} + j$. Store the residuals of the previous regression.
2. Run the Augmented Dickey-Fuller (ADF) [7] test (using the critical values provided by [8]) on the residuals of the previous step.
3. Iterate steps 1 and 2, with $j = \{0, 1, 2, \cdots\}$, until the end of the sample.

The cointegration analysis described above identifies a strong period if the null of no-stationarity of the ADF test on the residuals (in step 2) is rejected. Otherwise, the period under investigation is a weak period. In the empirical application, we set $T_{in} = 60$.

3 Results and Discussion

In Figure 1, the distributions of ATP points for the top ten male players between the beginning of 1990 to March 2023 are displayed. Some interesting observations can be made based on these distributions. The change in the ATP ranking system in 2009 is visible as more points are awarded afterward (not only to top players). There is also a dominance of one player (Roger Federer) followed not so closely by another player (Rafael Nadal) between 2004 and 2008. Some shorter dominant periods of one player are observed in 1994/95, 1997, and 2014/16. In addition, a very stable period in 2020 belongs to the Covid-conditioned period when ATP points were frozen for some time.

By the previously described method, we were able to identify strong periods and strong eras. The first strong period was identified between 13 May and 6 June in the year 1991 and involved six tennis players: Stefan Edberg, Boris Becker, Jim Courier, Ivan Lendl, Michael Stich, and Pete Sampras. These five players were cointegrated for the whole time between May 1991 and November 1992 only Lendl dropped out in some strong periods. Therefore, we can identify the first strong era between May 1991 and November 1992.

In a similar way, we identified the next eight strong eras. All of them are listed below (in the bracket are players that were not cointegrated during the whole era but only partially):

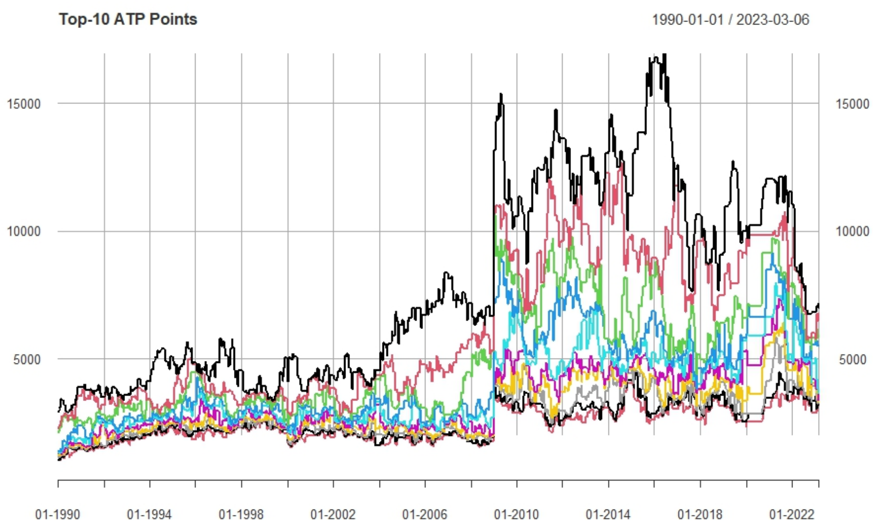


Fig. 1 Distribution of points for top 10 players on the ATP ranking lists between the start of 1990 and March 2023

1. *May 1991 - November 1992 (9 periods)*: Stefan Edberg, Boris Becker, Jim Courier, Michael Stich, Pete Sampras, (Ivan Lendl);
2. *April 1994 - August 1995 (5 periods)*: Pete Sampras, Andre Agassi, Michael Stich, Goran Ivanisević, Stefan Edberg, Sergi Bruguera, Boris Becker, (Thomas Muster);
3. *February 1996 - August 1998 (10 periods)*: Pete Sampras, Thomas Muster, Michael Chang, Goran Ivanisević, Yevgeny Kafelnikov, (Andre Agassi), (Boris Becker), (Patrick Rafter), (Petr Korda), (Marcelo Rios);
4. *March 1999 - November 2000 (14 periods)*: Pete Sampras, Yevgeny Kafelnikov, Andre Agassi, Gustavo Kuerten, (Patrick Rafter), (Alex Corretja), (Magnus Norman), (Marat Safin), (Carlos Moya);
5. *October 2002 - March 2004 (11 periods)*: Lleyton Hewitt, Andre Agassi, Juan Carlos Ferrero, Andy Roddick, Roger Federer, Marat Safin;
6. *March 2008 - June 2009 (2 periods)*: Roger Federer, Rafael Nadal, Novak Djoković, Andy Murray;
7. *March 2010 - May 2011 (2 periods)*: Roger Federer, Rafael Nadal, Novak Djoković, Andy Murray;
8. *June 2013 - December 2014 (5 periods)*: Novak Djoković, Rafael Nadal, Andy Murray, Roger Federer, David Ferrer, Stan Wawrinka;
9. *September 2018 - October 2020 (13 periods)*: Rafael Nadal, Roger Federer, Novak Djoković, Alexander Zverev, (Andy Murray), (Grigor Dimitrov), (Marin Cilić), (Juan Martin del Potro), (Dominic Thiem);

4 Conclusion

In this paper, we introduce the notion of the strong period as the basis for a strong era in the sport of tennis. To the best of our knowledge, this is the first attempt to evaluate the intensity of competition in the sport of tennis over time and distinguish more and less intensive periods of competition. This is a more statistically meaningful approach to evaluating competition on tour as it takes into consideration several top players at once as opposed to searching for the best player (great of all time).

The idea of strong periods/eras provides multiple ways for further research. One of them is to study strong periods/eras for different surfaces (grass, clay, hard). Another way of thought is to compare strong eras on male and female tours. This direction of research on strong periods and eras will help to better explore the nature of competition and from the applied perspective, it can help tour managers to adjust and promote the tours in a more appropriate way.

From the player and trainer perspective, such research can help to understand the characteristics of players who thrive in different intensities of competition. Additionally, from an applied standpoint of view, further research on the topic will help to provide winning strategies for both players and trainers during eras of different intensities.

We also expect this research to be useful for betting purposes as well as it can help to provide betting strategies for different periods of competition.

References

1. Angelini, G., Candila, V., De Angelis, L.: Weighted Elo rating for tennis match predictions. European Journal of Operational Research. 297, 120–132 (2022)
2. Arcagni, A., Candila, V., Grassi, R.: A new model for predicting the winner in tennis based on the eigenvector centrality. Annals of Operations Research. (2022)
3. ATP World Tour 2017 Rulebook: ATP World Tour. `https://www.atptour.com/rankings/rankings-faq.aspx`
4. Baker, R. D., McHale, I.: A dynamic paired comparison model: Who is the greatest tennis player?. European Journal of Operational Research. 236(2), 677–684 (2014)
5. Baker, R. D., McHale, I.: An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest womens tennis player? European Journal of Operational Research. 258(1), 328–333 (2017)
6. Breznik, K.: Revealing the best doubles teams and players in tennis history. International journal of performance analysis in sport. 15(3), 1213–1226 (2015)
7. Dickey, D. A., Fuller, W. A.: Distribution of the Estimators for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 74(366a), 427–431 (1979)
8. Engle, R. F., Yoo, B. S.: Forecasting and testing in co-integrated systems. Journal of Econometrics. 35(1), 143–159 (1987)
9. McHale, I., Morton, A.: A Bradley-Terry type model for forecasting tennis match results. International Journal of Forecasting. 27(2), 619–630 (2011)
10. Kovalchik A.: Searching for the GOAT of tennis win prediction. Journal Quantitative Analysis in Sport. 12(3), 127–138 (2016)
11. Radicchi F.: Who is the best player ever? A complex network analysis of the history of professional tennis. PLoS ONE. 6(2), e17249 (2011)
12. WTA rankings: Women's Tennis Association. `https://en.wikipedia.org/wiki/WTA_rankings`

NonParametric Combination method for data analytics in basketball matches

Metodo di combinazione non parametrica per l'analisi dei dati nelle partite di basket

Elena Barzizza, Nicolò Biasetton, Riccardo Ceccato, Marta Disegna and Giacomo Vezzosi

Abstract This study examines the usefulness of the NonParametric Combination technique in comparing basketball players' performance. After an extensive review of the literature on indices that can be computed to evaluate individual and team performance in basketball matches, the NonParametric combination technique is used to compare two players in the same role both on each performance index and on overall. Data collected during the 14 matches played by the basketball womens team at the University Sport Centre Padova in the 2022 winter season have been analysed in this study. Three couples of players have been analysed and the results show that only in one couple, one player was overall superior to the other.
Abstract *In questo studio viene analizzata l'utilità della tecnica di NonParametric Combination nel confrontare le prestazioni dei giocatori di basket. Dopo aver condotto un'estesa revisione della letteratura sugli indicatori che possono essere calcolati per valutare le prestazioni individuali dei giocatori e della squadra, la tecnica di NonParametric Combination è stata utilizzata per confrontare due giocatori nello stesso ruolo sia sui singoli indicatori che a livello complessivo. I dati*

Elena Barzizza
Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza, 36100, Italy, e-mail: elena.barzizza@phd.unipd.it

Nicolò Biasetton
Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza, 36100, Italy, e-mail: nicolo.biasetton@phd.unipd.it

Riccardo Ceccato
Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza, 36100, Italy, e-mail: riccardo.ceccato.1@unipd.it

Marta Disegna
Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza, 36100, Italy, e-mail: marta.disegna@unipd.it

Giacomo Vezzosi
Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza, 36100, Italy, e-mail: giacomo.vezzosi.00@gmail.com

raccolti durante le 14 partite giocate dalla squadra femminile del Centro Universitario Sportivo di Padova nella stagione invernale del 2022 sono stati analizzati in questo lavoro. Tre coppie di giocatrici sono state analizzate e i risultati dimostrano che solo in una coppia una giocatrice è globalmente superiore rispetto all'altra.

Key words: NonParametric Combination, Permutation test, Basketball

1 Introduction

In order to perform comparisons between pairs of players, the NonParametric Combination (NPC) technique [1] can be applied. NPC is a highly flexible permutation-based methodology which allows us to deal with complex problems, and, in the context of basketball data, it can be used to identify significant differences, in terms of multiple performance indicators, between players in the same role.

Let us assume that we want to compare the performance of two players, named P_1 and P_2, to verify whether one player outperforms the other. For each player we have information on V variables (i.e. performance indicators). The system of hypotheses of interest can be written as follows:

$$\begin{cases} H_0 : \mathbf{M}_1 = \mathbf{M}_2 \\ H_1 : \mathbf{M}_1 > \mathbf{M}_2 \end{cases} \tag{1}$$

where $\mathbf{M}_k$, $k = 1, 2$, is a vector of V statistics (such as median or mean) computed for the k-th player. The NPC methodology initially decomposes this system of hypotheses into V sub-systems (i.e., one for each performance indicator). The generic v-th sub-system of hypotheses ($v = 1, \ldots, V$) is as follows:

$$\begin{cases} H_{0v} : M_{1v} = M_{2v} \\ H_{1v} : M_{1v} > M_{2v} \end{cases} \tag{2}$$

After choosing an appropriate test statistic, NonParametric Combination addresses each sub-system individually and provides us with V partial p-values achieved through permutation, so that we can detect differences between players in terms of each performance indicator. It should be noted that the same permutation mechanism is applied for the computation of each partial p-value [1], so that we implicitly take into account the existing dependence between variables. Then, a combination step in the NPC algorithm allows us to retrieve a combined p-value, which merges the insights provided by the partial p-values, to be used to address the multivariate problem of interest expressed in the system of hypotheses (1).

This methodology only requires us to make two fundamental choices: the choice of the test statistic and the choice of the combining function to be adopted in the final combination step. The first choice is driven by the nature of the problem at hand. Since in this study we deal with numerical variables (see Section 2) characterised by outliers, the multivariate medians have been used for comparison between players. Therefore, the test statistic related to the v-th performance indicator is simply computed as a difference between medians as follows: $\text{median}_{1v} - \text{median}_{2v}$

The second choice is mainly driven by the number of available sub-problems in which the null-hypothesis is expected to be rejected and the correlation between variables [9]. In this study, we rely on Fisher's combining function, as it has been shown to be the best solution in many scenarios.

2 Data description

The case study analysed in this research refers to 14 matches played by women's basketball team of the University Sport Centre (CUS) Padova during the 2022 winter season. The official roster is made up of 16 players. Given a specific game, for each player who took part in the game, several information have been collected, including: shooting statistics (points scored; field goals attempted and made, separating also 2 and 3-point shots; free throws attempted and made; throws attempted and made from each part of the pitch), assists, rebounds, blocks, steals, fouls, turnovers and minutes played.

2.1 Basketball performance indexes

As described in [4] and [3], several indicators have been computed to evaluate the performance of players. Based on the data collected during the matches, the following indeces have been computed:

- Performance Index Rating (PIR) [6]. This metric allows us to assess the efficiency of players in a match, summing up all the "positive" actions (i.e. points, total rebounds, assists, steals, blocks, received fouls) and subtracting from this quantity all the "negative" actions (i.e. missed field goals, missed free throws, fouls committed, turnovers). The disadvantage of this index is that it does not consider neither the importance of the individual statistic among all the statistics nor the role of the player in the match.
- Player Impact Estimate (PIE) [7]. This index is computed for each player, for each team, and for each match. It allows us to obtain a measure of the overall contribution of the player and the team to each match. This index is useful for comparing players and teams.
- Floor Impact Counter (FIC) [5]. This index allows us to obtain a player's evaluation rating system in which greater importance is given to assists, offensive rebounds, and to the construction of offensive actions in general.
- Offensive Efficiency (OE) [8]. This index is used to measure the quality of player's offensive production. The index is computed as the rate between the

number of profitable offensive possessions in which the player has been involved and the total number of potential end-of-possession situations for the player.

- Efficient Offensive Production (EOP) [8]. This index is derived from the previous one and is computed as: (0.75 × assists + points) × OE. The tuning parameter of the assists (equals to 0.75) represents the estimated contribution of an assist to the final points scored.
- Adjusted Field Goal (AFG) [6]. This index allows us to obtain a measure of the shooting ability of each player.

2.2 Data analysis

The performance indicators computed as described in Subsection 2.1 have been divided by the minutes played by each player in each match before to perform the descriptive and inferential analysis. The analysis have been conducted on 6 players of the same team. The descriptive analysis allowed us to identify potential differences between pairs of players in the same role, comparable in terms of played matches and minutes played (see Figure 1). In particular, the pair of players is formed as follows: group 1 is made up of player 1 and player 2; group 2 is made up of player 3 and player 4; group 3 is made up of player 5 and player 6. Looking at Figure 1 it is possible to notice that in the first group player 2 outperforms, in terms of the median value for the EOP and the OE indices, the teammate. In the second group, player 3 outperforms (in terms of median value) player 4 with respect to FIC, PIE and PIR indices. Finally, in the third group, player 5 appears to perform better looking at the median value of the PIE index with respect to player 6.

3 Findings and conclusion

The results of the permutation-based testing procedure are reported in Table 1. Please note that the problem of multiplicity control arises when the number of sub-hypotheses to be tested is greater than one. For this reason, we applied the Bonferroni-Holm method to adjust partial *p*-values as indicated in [1].

As anticipated from the descriptive analysis, player 1 significantly outperforms player 2 in terms of EOP, with an adjusted partial *p*-value equal to 0.026. Regarding group 2, player 3 outperforms player 4 with respect to the PIR, PIE, and FIC indices (*p*-values are all lower than 0.05). Finally, in group 3, player 5 outperforms player 6 in terms of PIE only (*p*-value 0.044).

The combined *p*-values give us further useful indications regarding the players' overall performance. Using this combined test, it is possible to affirm that overall player 3 outperforms player 4 (combined *p*-value equals 0.006) while in group 1 and group 2 it is not possible to identify a player who overall outperforms the other

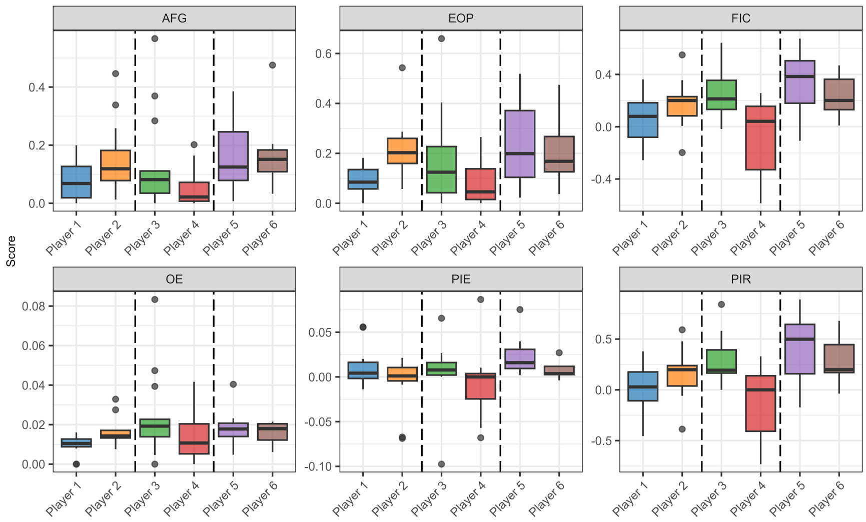


Fig. 1 Distribution of the different performance indicators per player.

Comparison	PIR	PIE	FIC	OE	EOP	AFG
$P_1 < P_2$	0.318	0.791	0.277	0.318	0.026	0.430
$P_3 > P_4$	0.014	0.014	0.014	0.134	0.544	0.669
$P_5 > P_6$	0.582	0.044	0.546	0.844	0.844	0.776

Table 1 Partial p-values.

one (the combined p-value are equal to 0.051 and 0.197 respectively for group 1 and group 2).

The proposed permutation-based test can be easily extended to the comparison of more than two players, integrating it with the multivariate ranking procedure by [2].

Acknowledgements This study was carried out within the MICS (Made in Italy Circular and Sustainable) Extended Partnership and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors views and opinions, neither the European Union nor the European Commission can be considered responsible for them. This study has also been supported by the following projects: BIRD 2022 project titled Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis; PNRR MOST - Centro Nazionale per la MObilità soSTenibile, Spoke8: Maas & Innovative services.

References

1. Pesarin, F., Salmaso, L.: Permutation tests for complex data: theory, applications and software. John Wiley & Sons (2010)
2. Arboretti, R., Bonnini, S., Corain, L., Salmaso, L.: A permutation approach for ranking of multivariate populations. Journal of Multivariate Analysis. 132, 39–57 (2014)
3. Metulini, R., Gnecco, G.: Measuring players importance in basketball using the generalized Shapley value. Annals of Operations Research. 1–25 (2022). https://doi.org/10.1007/s10479-022-04653-z
4. Corain, L., Arboretti, R., Ceccato, R., Ronchi, F., Salmaso, L.: Testing and ranking on round-robin design for data sport analytics with application to basketball. Statistical Modelling. 19(1), 5–27 (2019)
5. Ferrario, A.: Basketball Analytics: the use of data science to describe and predict the performance of an NBA team. *Master's thesis, Università degli studi di Milano Bicocca*, (2021)
6. Cene, E., Parim, C., Özkan, B.: Comparing the performance of basketball players with decision trees and TOPSIS. Data Science and Applications. 1(1), 21–28 (2018)
7. Senatore, J.V., Fellingham, G., Lamas, L.: Efficiency and productivity evaluation of basketball players' performance. Motriz: Revista de Educação Física. 28 (2022)
8. Lee, D.-J., Page, G.L.: Big Data in Sports: Predictive Models for Basketball Player's Performance. Mathematics in Industry Reports. (2021)
9. Langthaler, P.B., Ceccato, R., Salmaso, L., Arboretti, R., Bathke, A.C.: Permutation testing for thick data when the number of variables is much greater than the sample size: recent developments and some recommendations. Computational Statistics. 38, 101–132 (2022)

Solicited Session SS3 - *Statistical methods for the analysis of university student choices and academic performance*

Organizers and Chairs: Isabella Sulis and Marialuisa Restaino

1. *The influence of labor market conditions on students' career disruption: first insights from Italy* (Usala C., Sulis I. and Porcu M.)
2. *Socio-economic aspects that may affect South-North students' mobility in Italy* (Genova V.G. and Boscaino G.)
3. *An analysis of student's performance in bachelor's degree* (La Rocca M., Niglio M. and Restaino M.)
4. *An exploratory strategy for analyzing students' mobility data* (Primerano I. and Giordano G.)

The influence of labor market conditions on students' career disruption: first insights from Italy

L'influenza del mercato del lavoro sulle carriere universitarie: prime evidenze dall'Italia

Cristian Usala, Isabella Sulis and Mariano Porcu

Abstract This contribution aims to disentangle the effect of the labor market conditions and inequalities at the entrance in determining the outcomes of students' university careers at the end of their first year. For this sake, multiple data sources have been merged with the MOBYSU.IT, which contains the Italian National Student Archive (ANS) microdata related to the cohorts of students enrolled for the first time in one Italian university between 2013 and 2018. The core of the analysis is devoted to assessing the determinants of students' career outcomes in terms of career disruption (dropout and risk of dropout) or success. A two steps procedure, which relies on multilevel models, has been adopted to disentangle the effects of job market conditions on students' career outcomes from the ones related to the attended university, students' characteristics and high school background.

Abstract *Lo scopo di questo contributo è quello di misurare l'effetto del mercato del lavoro e delle disuguaglianze scolastiche nel determinare il rischio di abbandonare gli studi universitari in Italia. Per questo motivo sono state utilizzate più fonti di dati: MOBYSU.IT, che contiene i microdati relativi all'Anagrafe Nazionale Studenti (ANS) sulle coorti di immatricolati negli atenei italiani dal 2013 al 2018, le statistiche territoriali sul mercato del lavoro e i dati AlmaLaurea sull'inserimento dei laureati nel mercato del lavoro. L'analisi si focalizza sulla valutazione delle determinanti degli esiti della carriera degli studenti al termine del primo anno, definendo tre categorie di studenti (abbandoni, a rischio di abbandono, regolari). È stata adottata una procedura a due stadi, che si basa su modelli multilivello, per*

Cristian Usala
Department of Political and Social Sciences
University of Cagliari, e-mail: cristian.usala@unica.it

Isabella Sulis
Department of Political and Social Sciences
University of Cagliari, e-mail: isulis@unica.it

Mariano Porcu
Department of Political and Social Sciences
University of Cagliari, e-mail: mrporcu@unica.it.

valutare l'interazione tra cause multiple nella determinazione degli outcome osservati.

Key words: MOBYSU.it, dropout, multilevel models, school effect, labour market conditions

1 Introduction

The Italian university system suffers from chronic issues related to students' career disruption. For example, the 2021 OECD statistics on full-time students' completion rates show that only 18.5% of men and 22.3% of women who entered at bachelor's programme graduate on time. Moreover, the students' careers completion rates stay low even after 3 years from the end of the theoretical duration, with about 50% of men and 45% of women who did not complete their studies [3, 12, 13, 4, 11]. This dramatic scenario, together with the high dispersion rates at high schools, explains the low rate of young people in the age group 30-34 with a tertiary education (about 27%), which locates Italy in the last positions of the ranking of the European countries. The information provided by national assessment surveys on students' competencies for Italy shows a high variability in students' competencies between schools and classes at all grades, especially in the South regions of the country, and a stratification of students in type of secondary schools which in average reflects the socioeconomic status of the families [9]. According to INVALSI surveys for 2019 [9] (before the pandemic), students attending scientific or classic high schools scored on average 50 points more than students attending professional schools and about 30 points more than those attending technical schools when considering reading performances. A similar stratification of students in schools is also registered regarding math competencies. Moreover, survey results provide evidence of a clear association between students' competences and families' socioeconomic and cultural status (ESCS index), with students in the highest quartile of the ESCS distribution who, on average, score 27 points more in reading and 24 in math than those coming from the bottom quartile of the distribution. Indeed, marked divergences between the average ESCS index is detected across schools curricula (e.g. liceo vs technical and vocational schools) [9, 2], indicating that the combination of family of origin and the secondary school attended generate clear inequalities in students competencies at the entrance, which in turn affect their university career [16, 5, 15, 1]. Barbieri et al. [1] analyze students' career disruption in order to disentangle the high school's effect from the university effect with the double aim of building up a set of indicators to compare university inefficiencies and assess the role played by poor university services in enhancing students' pathways disruption. Their results show that the school effect (i.e, the specific institution attended) is the main predictor of risk to experience adverse events during the 1st year of the university career, over and above the kind of high school attended in terms of curricula.

We move from this first result to focus on the role that factors external to the universities, such as inequality at the entrance in students' background and the socioeconomic conditions of the territories, play in determining students' status at the end of the 1st year. In doing this, this contribution exploits the role of (i) job market conditions of the territories where students reside (origin) and where the university is located (destination) and (ii) future expected economic rewards of having a degree in a specific disciplinary field (in terms of salary and probability to get a job) in influencing students' choices in terms of continuing the university studies in the second year, by accounting for the inequalities at entrance in students' competencies due to the secondary schools attended.

Previous research on the topic shows that the expected return of education affects students' and family's decisions to invest in higher education [8, 14, 6, 4], but there is no clear evidence on how labor market conditions may indeed affect the decision of dropout. Indeed, better job market opportunities may increase the expectations of higher return of education for individuals with a university degree, and thus they could push students to invest more in higher education; at the same time, especially for disadvantaged students, the increase in the cost opportunity of the investment in education due to the earning lost during the university studies may determine an increase of the dropout rates [6]. In this framework, the risk of dropping out during students' first year of career may be higher for those who experienced difficulties at entrance and, thus, doubt about their effective possibilities of getting a degree. Moreover, individuals' reaction to adverse labor market conditions is not straightforward. In economic crisis periods, individuals may decide to invest more in education to have higher chances of getting a good job, but they could also be forced to dropout because they cannot afford to support their university studies financially [8, 14, 6].

The main findings from the Italian framework highlight a negative relationship between adverse job market conditions and dropout [6, 4]. However, the results indicate that the sign and the magnitude of the effects related to labor market conditions vary depending on the set of predictors at the territorial level that is selected [6] and the considered disciplinary areas [4].

Moving from this framework, and the results of previous studies [6, 5, 4], the aim of this contribution is to disentangle the effect of labor market conditions and inequalities at entrance in determining students' university career at the end of their fist year of career in Italy.

2 Data

In order to assess the role of labor market conditions and the inequalities at the entrance in terms of students' characteristics and high school background, multiple sources of data have been combined. The information on students' careers and high school backgrounds has been gathered from the database MOBYSU.IT, which contains the microdata from the Italian National Student Archive (ANS). In particular, the analysis focuses on the cohorts of students enrolled for the first time in one Italian university between 2013 and 2018.

Labor market conditions have been measured using two different perspectives. First, the data from the AlmaLaurea surveys related to graduates' employment conditions allows us to account for divergences in job market opportunities among universities and degree programs. In particular, we have obtained data on the employment rate and average salary one year after graduation for each university that participated in the AlmaLaurea survey and each degree program. Second, the assessment of the impact of factors external to the university system such as the labor market conditions of students' areas of residence and universities' hosting areas, have been operationalized using the data from ISTAT on the provincial unemployment rate, total taxable income per capita at the municipal level, and the number of local firms or branches of firms in the municipality.

These data sources are combined to analyze students' career status at the end of the first year of their careers. In particular, students are classified into three categories: dropouts, at risk of dropout and regular. Dropouts are students who did not enroll in any university during their second year. Students at risk of dropout are those that earned less than 25 formative credits (CFU) at the end of their first year, while regular students are the residual category.

Therefore, in this framework, we aim to identify the determinants of students' career disruption related to their high school background, the labour market conditions in both origin and destination areas and the job market opportunities related to the attended university and degree program.

3 Methods

This analysis relies on the results of a two steps procedure. In particular, the first step is used to disentangle the effects of the cross-classification of students in high schools, universities and degree programs on their career outcomes. At this aim, we estimate several multilevel models to identify the role of these elements on students' performances in terms of CFU at the end of their first year. The posterior estimates obtained in the first step are then used in the second as predictors to account for the role of students' educational backgrounds on the probability of observing one of the possible statuses of students' careers (regular, at risk of dropout, dropout). At this aim, the second step estimates a multinomial logit model to assess the role of job market conditions and opportunities on students' career outcomes that do not depend on the heterogeneity in students' characteristics or educational background. The approach has been carried out two times using as predictors of the labor market conditions the indicators related to students' areas of residence and universities' hosting areas.

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian NorthSouth divide.", n. 2017HBTK5P - CUP B78D1900018000.

References

1. Barbieri B., Porcu M., Salaris L., Sulis I., Tedesco N., Usala C.: University dropout and churn in Italy: an analysis over time. In Balzanella A., Bini M., Cavicchia C., Verde R. Editors. Book of Short Papers 51st Scientific Meeting of the Italian Statistical Society, Caserta, 22-24 June, p. 1734-1739, Pearson, Milano (2022)
2. Busetta, A., Fabrizi, E., Sulis, I., Ragozini, G.: Mobilità sociale delle famiglie. In Rapporto sulla popolazione. Le famiglie in Italia. Forme, ostacoli, sfide (pp. 207-234). il Mulino (2023)
3. Belloc, F., Maruotti, A., Petrella, L.: How individual characteristics affect university students dropout: a semiparametric mixed-effects model for an Italian case study, Journal of Applied Statistics, 38:10, 2225-2239 (2011)
4. Contini, D., Zotti, R.: Do Financial Conditions Play a Role in University Dropout ? New Evidence from Administrative Data. In: Checchi, D., Jappelli, T., Uricchio, A. (eds) Teaching, Research and Academic Careers. Springer, Cham (2022)
5. Contini, D., Cugnata, F., Scagni, A.: Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy. High Educ 75, 785–808 (2018)
6. Di Pietro G.: Regional labour market conditions and university dropout rates: Evidence from Italy, Regional Studies, 40:6, 617-630 (2006)
7. Di Pietro, G., Cutillo, A.: Degree flexibility and university drop-out: The Italian experience. Economics of Ed Di Pietro, G., Cutillo, A. Degree flexibility and university drop-out: The Italian experience. Economics of Education Review, 27, 546-555.ucation Review, 27, 546-555 (2008)
8. Duncan, B.: Dropouts and the unemployed. Journal of Political Economy 73, 121-134 (1965)
9. INVALSI. Rapporto prove INVALSI 2019. INVALSI
10. Goldstein, H. Multilevel Statistical Models. Wiley Series in Probability and Statistics, 4 Edn. Wiley & Sons (2019)
11. OECD, Education at a Glance 2022: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/3197152b-en (2022)
12. Meggiolaro S., Giraldo A., Clerici R.: A multilevel competing risks model for analysis of university students' careers in Italy. Studies in higher education, vol. 42, 1259-1274 (2017)
13. Perchinunno, P., Bilancia, M., Vitale, D.: A statistical analysis of factors affecting higher education dropouts. Social Indicators Research, 156(2), 341-362 (2021)
14. Rees, D. I., Mocan, H.N.: Labor market conditions and the high school dropout rate: Evidence from New York State, 16 (2), 103-109 (1997)
15. Tedesco, N., Salaris, L.: University dropout and mobility in Italy. First evidence on first level degrees, SIS 2020 - Book of Short Papers. PEARSON, 1601-1606 (2020)
16. Trivellato, P., Triventi, M.: Differentiated trends in student access and performance during the Bologna Process. The case of universities in Milan. Italian Journal of Sociology of Education, 3(2) (2011)

Socio-economic aspects that may affect South-North students' mobility in Italy

Aspetti socio-economici che possono influenzare la mobilità studentesca in Italia

Vincenzo Giuseppe Genova and Giovanni Boscaino

Abstract Students' mobility phenomenon affects several aspects, such as demographic and economic ones, and when the balance of the movement is negative for a particular region, socioeconomic problems can arise. Driven by the better socioeconomic conditions of some areas, in Italy, thousands of students every year leave the South to study in the Centre-North. Such a process worsens the historical and socioeconomic gap between the North and South. This paper investigates how some social, economic and educational aspects can encourage students to move. We considered different official data and statistics merged together. Results highlight the impact of job-market and scholarship support and hint at a chain migration effect on student mobility.

Abstract *La mobilità degli studenti è un fenomeno che riguarda diversi aspetti, per esempio demografici ed economici, e quando il saldo del movimento tra diverse area non è bilanciato possono insorgere diseguaglianze. In Italia, ogni anno migliaia di studenti lasciano il Sud per studiare al Centro-Nord, e non vice-versa. Questo processo aggrava il divario storico e socioeconomico tra Nord e Sud. Il presente lavoro analizza come alcuni aspetti sociali, economici ed educativi possano incoraggiare gli studenti a spostarsi. L'analisi si basa su diversi dati e statistiche ufficiali integrati insieme. I risultati evidenziano l'impatto del mercato del lavoro e del sostegno alle borse di studio e accennano a un effetto di catena migratoria sulla mobilità degli studenti.*

Key words: Students' mobility, Quality of life, Universities' services, Gravity model, Chain Migration

---

Vincenzo Giuseppe Genova
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo
e-mail: vincenzogiuseppe.genova@unipa.it

Giovanni Boscaino
Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo
e-mail: giovanni.boscaino@unipa.it

1 Introduction

This paper concerns students' domestic mobility within Italy. This phenomenon needs close attention because it does not affect Italian regions equally. Typically, the students' mobility origins from the southern areas towards the Centre-Northern ones, increasing in recent years. For example, in the academic year 2016-2017, 25.6% of students from the south who went to university studied in the Centre-North. The share of residents in the Centre-North studying in the South is just 1.9%. The net migration balance is about 157,000 units, and it is continuously increasing [15]. Therefore, students' mobility from an opportunity has become a socioeconomic problem. These flows have worsened regional disparities and the existing dualism between Italy's southern and northern regions [1]. From the origin region perspective, consistent capital is transferred to support students' life and studies; often, the best students are those who move; in most cases, moving students never return to their region of origin. Economic problems are strictly connected to social ones. Perry & Wiewel [12] introduced the notion of the university as an "urban developer and regenerator" as the attraction of students increases in both population and demand for consumer goods. Besides, universities also play a core role in the level of education in the area in which they are. This importance is amplified when the economy is knowledge-based, for which human capital is of primary importance, and the labour market requires highly qualified people [6]. The empirical evidence shows how, when we consider the entire range of working life, the rate of Return On Investments in human capital in the southern regions is about half of that observed in the "richest" areas of the Centre-North (typically Piedmont, Lombardy and Emilia-Romagna) [7].

Students mobility is mainly due to obtaining better education from the best universities and increasing the prospects of getting a job and being well paid. In addition, the study carried out by Dotti et al. [6] highlights that, although the quality of a university and its characteristics play a central role in the decision process of a student, the features of the labour market in a specific region must be taken into account as well. They point out that in the southern areas, no universities attract students living more than 200 kilometres of distance, and the number of registrations at a university is linked to students' job prospects.

Finally, students' choice is probably facilitated by the presence of friends or relatives who live in the areas of the universities or, more simply, their advice/opinions on the studies they are following there. That is known in the literature as chain migration [10, 8]. Beine *et al.* [2] argued that students' mobility could be affected by migration costs and the size of the network effect at the destination. They define the network effect as migrants from the origin city that live at the destination, and this network can facilitate migration flows. Former migrants – such as family members, friends or high-schoolmates – are likely to provide assistance and information to those students who decide to migrate, reducing migration costs [2, 3]. Therefore, according to Haug [8], when a student moves to a destination area because family members, high-schoolmates or friends advise, we are in the presence of chain migration.

The study presented in this paper aims to explore the determinants of students' domestic mobility in Italy. We have considered information from official statistics about universities and city areas' characteristics, seeking aspects that can motivate students to abandon the South. In particular, the focus is on the pull factors of the destination universities and areas. We aim to find some guidance as answers to the questions i) "What institutional aspects, if any, are most attractive?"; ii) "What regional aspects, if any, are most attractive?"; iii) "Does exist a chain migration effect?".

2 Data

Data - drawn from the "Italian Anagrafe Nazionale della Formazione Superiore"- has been processed according to the research project "From high school to the job market: analysis of the university careers and the University North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero Università e Ricerca", and INVALSI. In particular, the database used here is the so-called MOBYSU.IT database [11] that is a longitudinal database that collects individual information about the population of Italian freshmen followed up to graduation, such as Gender, Degree Course, Topic Area, High School Diploma, and High School Grade.

We considered just the freshmen from Sicily. It represents a particular region of the South of Italy because it is economically deprived and has no territorial bounds that ease mobility: very motivated students should move towards Centre-North.

To answer the three main research questions, we considered the most attractive university of the Centre-North for Sicilian students. We integrated the dataset with some additional context variables. In particular, concerning question i), we referred to the annual survey *La classifica CENSIS delle università italiane* [4, 5] to use some indicators on university characteristics (e.g. Scholarship, Services for the students, Structures quality, and Web Communication). To answer question ii), we referred to the annual survey *Qualità della vita de il Sole 24 ore* [13, 14], considering some basic indicators about several aspects like the Unemployment Rate, House Price, Free Time (considering the presence of bars and restaurants, bookshops, cinemas, theatres, sports and cultural events), and Public Order. These two surveys are the most relevant reports on the Quality of Life and Universities in Italy, collecting several data and official statistics.

For our analysis, we referred to two cohorts of freshmen: 2014 and 2017. The choice is due to the availability of full information from the three data sources used in this analysis. Specifically, we selected the 2017 cohort, the most recent available year. For comparison, we have considered the 2014 cohort, as two too-close cohorts could have a similar pattern.

3 Methodology

In the students' mobility context, following the idea of Beine *et al.* [2], we applied a Gravity Model with mixed effects on the provinces of "out of Sicily" universities (provinces of destination). The aim is to capture the variability not explained by the model's covariates for the attractiveness of non-Sicilian provinces. Following Huang's suggestions [9], we assume that the response variable is distributed as Conway-Maxwell-Poisson $N_{ij|k} \sim CMP(\lambda(\mu, \nu))$, where $N_{ij|k}$ is the number of outgoing students from cluster i to university's province j conditioned to the $k-th$ covariate profile; λ is the rate parameter of the CMP that is a function of $\mu = E(Y)$ and ν the dispersion parameter. Under these assumptions, the model is formulated as follows:

$$\ln(\mu_{ij|k}) = \alpha_{ij} + V_i\phi + X_j\beta + Z_u\gamma + \sum_{t=t_s}^{T-1} \theta_{ij}^t \ln(M_{iju}^t) + \omega_j, \tag{1}$$

where:

- α_{ij} is the intercept;
- V_i is the vector of student's characteristics in the cluster i;
- X_j are socioeconomic covariates of the provinces of destination;
- Z_u are covariates related to the characteristics of university u;
- T is the year of the cohort under analysis;
- t_s is a given starting year of previous cohorts, with $t_s < T$;
- M_{iju}^t are students of previous years from cluster i that study at university u in the province j at time t;
- ϕ, β, and γ are vectors of unknown regression parameters;
- θ_{ij}^t is the network effect at time t;
- $\omega_j \sim N(0, \sigma^2)$ is the province of destination level random effect.

On the one hand, we included in Eq. 1 some socioeconomic and university covariates that can act as pull factors. On the other hand, to study the possible network at the destination effect [3] and how far in time this effect is significant, we included the θ parameter. If a chain migration effect exists, the parameter θ should be significant until a give t in the past and, at the same time, it can suggest how many years back this effect is present for a given cohort. Then θ should be considered as a proxy of chain migration, according to the definition of Haug [8].

4 Results

As mentioned in Section 2, we used a database coming from merging information from MOBYSU.IT [11], CENSIS [4, 5], and "Il Sole 24 Ore" Quality of Life of the Italian cities survey [13, 14] datasets. The baseline profile for the model in Eq. 1 is a Female student with High School Grade $[60-70]$, a Scientific Diploma, that enrol on a Degree Course in the Humanities Area, along with the quantitative covariates mentioned above.

Results come from several model comparisons and selections, using AIC for model selection. For the 2014 cohort, Males seem to move more than Females, similar to the best-graded students and those who gained a Scientific Diploma. It is possible to consider a detractor effect due to the Unemployment Rate – as expected, the higher it is, the lower the expected number of outgoing students. Instead, the availability of Scholarships at the university of destination positively affects the expected number of outgoing students. Finally, paying attention to the effect of the network at the destination, it seems to be a significant and positive effect for the two years that precede the movers of the cohort 2014.

The results concerning the 2017 cohort are quite similar. Still, the expected number of outgoing students is greater for those students that chose a scientific area, and it's lower for those who chose a non-humanistic area. Such a result is maybe due to an evolution of the labour market between the two periods, which could have determined a more rational choice of degree courses. Focusing on the network effect at the destination, also for this cohort, there is a significant effect on the previous three years.

Furthermore, covariates concerning Life Quality, such as Free Time and Public Order, do not seem to play a role in the students' mobility for both cohorts. Reading this result together with the High School Grade, the Unemployment Rate, and the availability of Scholarships, it seems reasonable the idea that students' families plan the future of their talented and motivated children taking into account mainly the work perspectives and socioeconomic aspects, such as the migration costs that became less expensive due to both scholarship availability and – according to Haug [8] – the network at the destination. In fact, that can provide assistance and crucial information. That could partially explain the non-effect of Public Order and Free Time covariates. The presence of friends and relatives may make families more confident in sending their children off. Similarly, students may place more importance on the network of friends and family at the destination than the city's recreational prospects.

Funding

This paper has been supported by the national grant of the Italian Ministry of Education University and Research (MUR) for the PRIN-2017 program (Progetti di ricerca di Rilevante Interesse Nazionale), n. 2017HBTK5P - CUP B78D19000180001- Research Project Title "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide".

References

1. Attanasio, M., Enea, M.: La mobilità degli studenti universitari nell'ultimo decennio in Italia. il Mulino, Bologna (2019)
2. Beine, M., Docquier, F., Ozden, C.: Diasporas. J. Dev. Econ. (2011)
3. Beine, M., Noel, R., Ragot, L.: Determinants of the international mobility of students. Econ. Educ. Rev. (2014). DOI 10.1016/j.econedurev.2014.03.003
4. CENSIS: La classifica CENSIS delle università italiane (edizione 2015). CENSIS, Roma (2015)
5. CENSIS: La classifica CENSIS delle università italiane (edizione 2018). CENSIS, Roma (2018)
6. Dotti, N.F., Fratesi, U., Lenzi, C., Percoco, M.: Local labour markets and interregional mobility of italian university students. Spat 8(4), 443–468 (2013)
7. Guagnini, M., Mussida, C.: Il rendimento dell'istruzione nelle regioni italiane. In: AIEL. Associazione Italiana Economisti del Lavoro, Atti XXIV National Conference of Labour Economics, University of Sassari (2009)
8. Haug, S.: Migration networks and migration decision-making. J. Ethn. Migr. 34(4), 585–605 (2008)
9. Huang, A.: Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. Stat. Model 17(6), 1–22 (2017)
10. MacDonald, J.S., MacDonald, L.D.: Chain migration ethnic neighborhood formation and social networks. Milbank Q 42(1), 82–97 (1964)
11. MOBYSU.IT: Database MOBYSU.IT (2016), Mobilità degli studi universitari italiani, Protocollo di ricerca MIUR - Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II, Fonte dei dati ANS-MIUR/CINECA (2016)
12. Perry, D.C., Wiewel, W.: The University as Urban Developer: Case Studies and Analysis: Case Studies and Analysis. Routledge, London (2015)
13. Sole 24 Ore: Qualit'a della vita 2014 (2014). URL `https://lab24.ilsole24ore.com/`
14. Sole 24 Ore: Qualit'a della vita 2017 (2017). URL `https://lab24.ilsole24ore.com/`
15. Svimez: Rapporto Svimez. Il Mulino, Roma (2018)

An analysis of student's performance in bachelor's degree

Un'analisi della performance degli studenti nelle lauree di primo livello

Michele La Rocca, Marcella Niglio and Marialuisa Restaino

Abstract This manuscript aims at exploring the differences in students' performances among bachelor's degrees by using regression models. The analysis concerns students enrolled at 3-year degrees in an Italian university (located in the South of Italy) during the last 5 years. To measure students' success, we focus on the number of ECTS credits earned during the first year. Hence, the main purposes are to i) estimate the probability of getting at least a certain number of credits at the end of the first year, and ii) identify which students' features might affect it.
Abstract *Il presente lavoro intende studiare le principali difference nella performance degli studenti iscritti alle lauree di primo livello, attraverso un modello di regressione. L'analisi riguarda gli studenti iscritti alle lauree triennali di una università italiana dell'Italia meridionale durante gli ultimi 5 anni. Per misurare il successo degli studenti, ci si concentra sul numero di crediti maturati durante il primo anno. Quindi, gli obiettivi sono due: i) stimare la probabilità di raggiungere almeno un certo numero di crediti alla fine del primo ano, ii) individuare quali caratteristiche degli studenti potrebbero influenzarla.*

Key words: students performance, ECTS credit, regression analysis

1 Introduction

Educational Data Mining (EDM) is an emerging research field that focuses on the application of techniques and methods of data mining in educational environments. In particular, it is concerned with developing, researching, and applying machine learning, data mining, and statistical methods to explore and detect patterns in large collections of educational data that would be difficult to analyze [10]. So, EDM has become an effective tool used to identify hidden patterns in educational data, predict academic achievement, and improve the learning/teaching environment [5, 7].

Michele La Rocca · Marcella Niglio · Marialuisa Restaino
Dept of Economics and Statistics (DiSES) & Laboratory of Statistics (StatLab)
e-mail: [larocca, mniglio, mlrestaino]@unisa.it

The educational data and services that might be analyzed by data mining methods are i) students'/staff/professors' mobility; ii) students' information and their success; iii) degree and time for graduation; iv) students' dropouts; v) Student Evaluation of Teaching (SET) and their participation in class; vi) quality of research.

In this manuscript, we will focus on student success, which is a crucial component in higher educational institutions and is often used as an essential criterion for assessing the quality and performance of educational institutions. Early detection of students at risk and the adoption of some preventive measures can help decision-makers to provide and plan the needed actions for improving students' success, and eventually revise the course of study.

Among several definitions and different levels of student success, available in the literature (for a comprehensive review of the definitions see [3]), we will aim attention to students' performance, measured as the number of European Credit Transfer System (ECTS) credits, earned during their carrier.

Several papers have investigated students' performance, by identifying the main determinants, even if the results are not always in the same direction. An interesting classification of papers could be between

- papers accounting for students' social and demographic characteristics (among others, [6], [8], [9]); and
- papers accounting for their previous performance and/or psychological and subjective features (among others, [2], [4]).

The literature generally suggests that the impact of the determinants varies (in terms of extent and direction) according to the context (economic, social, political, demographic, etc.) and results should hold just in that context.

Within this framework, we will analyze the differences in students' performance between bachelor degrees, focusing on the threshold of 40 ECTS (Annual Review Report by ANVUR - Italian National Agency for the Evaluation of Universities and Research Institutes) earned by students during the first academic year, and figuring out which background students' characteristics might influence their performance. The main research questions we wish for looking into are:

a) Does the pre-enrollment test score affect the probability of reaching the 40 ECTS during the first academic year?
b) Is a gender gap detectable?
c) Are there any differences among the courses of study?

The manuscript is structured as follows. In Section 2, the education data used and the merging process are described. The statistical model implemented for analyzing students' performance is also briefly outlined. Finally, Section 3 concludes by showing the main results of the data analysis.

2 Data, merging, and methods

The data refers to students who enroll at the University of Salerno, and among all departments, we collect information for the Department of Economics and Statistics

(DiSES). The data was downloaded from the Student Information System (ESSE3), which manages the entire career of students from enrollment to graduation and contains information about students' high school diplomas, personal characteristics, exams, abroad experience, internships, and degrees. Given the aims of the manuscript, we extract and merge information on students' enrollment, exams, and pre-enrollment tests. In particular, the datasets, with different sizes, downloaded are

1) "*Student enrolled*" (Iscritti), which contains a record for each student, with all background information (date of birth, place of residence, high school diploma, diploma mark, and so on).
2) "*Student exams*" (Esami sost/non sost), which contains more records for a student since each record is related to each exam that should be taken by the student in the academic year (date of exam, exam mark, ECTS, and so on).
3) *pre-enrollment test*, which contains the test score of all students who took it. It consists of 36 questions (13 questions on Logic; 10 questions on Reading; 13 questions on Mathematics).

A merging process is shown in Figure 1. The keys used are the students' ID (*matricola*) and the fiscal code.

In more detail, it consists of four major stages. In the first stage, the "student exams" dataset is transformed into a new dataset "exams", where there would be a record for each student. The main variables considered are the total number of ECTS credits earned during the first year, the total number of exams passed during the first year, and the mean of exam marks. In the second stage, the two datasets "student enrolled" and "exams" are merged by the students' ID (*matricola*), leading to "students & exam". In the third stage, students who took the test and enrolled in the courses of study under investigation are selected, getting the "test" dataset. In the last step, "students & exam" and 'test" datasets are merged by the students' fiscal code.

The analysis covers four academic years (2018–2019 to 2021–2022) and concerns the students who have enrolled in one of the bachelor's degrees offered by the Department of Economics & Statistics at the University of Salerno. Given the

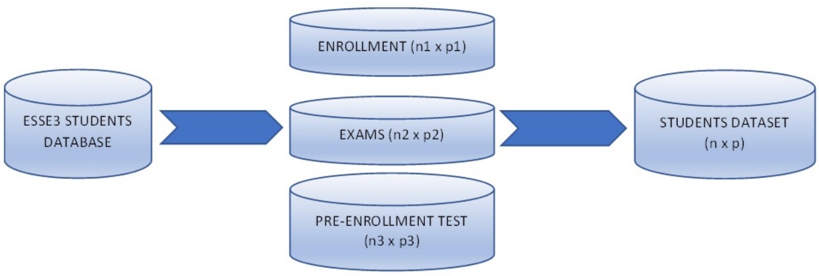


Fig. 1 The merging process between the datasets

heterogeneity of pre-enrollment tests, we only focus on the courses that require the TOLC-E (Test OnLine CISIA for Economics): Business Administration (BA - L18 - code 02121), Economics (E - L33 - code 02124), Statistics for Big Data (SBD - L41 - 02128).

Among all variables available in ESSE3 and TOLC-E, the variables considered as potential features for studying and predicting the students' performance are described in Table 1.

Table 1 List of covariates with a short description and type

Variable	Short Description	Type
Academic Year	Academic Year of enrollment	Nominal
Course of study	Course of study where students have enrolled (BA, E, SBD)	Nominal
Gender	Gender of the graduated student	Nominal
Type of diploma	High school type (Classical studies, Technical, Scientific, ...)	Nominal
Diploma mark	Total marks of high school diploma (from 60 to 100)	Integer
Logic score	Pre-enrollment test - Logic	Integer
Reading score	Pre-enrollment test - Reading	Integer
Mathematics score	Pre-enrollment test - Mathematics	Integer
Total Test score	Pre-enrollment test - Total mark	Integer
AER (OFA)	Additional Educational Requirements (1=Yes vs 0=No)	Nominal

As highlighted in Section 1, the aim is to capture the differences in students' performance between bachelor degrees, looking at the threshold of 40 ECTS and assessing which background students' characteristics might affect their success in reaching it. Hence, the interest is in studying an ordinal response variable Y_i ($i = 1, \dots, n$, where n is the total number of students), that takes $C = 4$ categories, according to the number of credits earned by students.

$$Y_i = \begin{cases} 1 & \text{ECTS} = 0 \\ 2 & 0 < \text{ECTS} \leq 39 \\ 3 & 40 \leq \text{ECTS} < 60 \\ 4 & \text{ECTS} = 60 \end{cases} \tag{1}$$

Let $\mathbf{x}_i$ be a p-dimensional vector for i-th student, that summarizes his/her features as displayed in Table 1. Given the nature of the response variable, an ordinal regression model [1] is employed to assess the existence of some differences in students' performance across the bachelor's degrees, by estimating the cumulative probabilities of the response variable:

$$\Pr(Y_i \leq j|\mathbf{x}_i) = \frac{\exp(\beta_{0j} + \mathbf{x_i}'\beta)}{1 + \exp(\beta_{0j} + \mathbf{x}_i'\beta)} \quad \text{for} \quad j = 1, 2, 3, 4. \tag{2}$$

where the parameters β_{0j} are the intercepts, called *thresholds* or *cutpoints*, that vary across the categories and are in increasing order, while β is not indexed by the category index j, thus the effects of the covariates are constant across response categories.

3 Some results of the data analysis

In this section, we will discuss the main results of the exploratory data analysis and of the ordinal logistic regression.

Table 2 shows the distribution of the number of credits for gender and bachelor degrees. Three-quarters of students earn at least 40 credits and the percentage of females who reach the threshold is higher than that of males. Between bachelor's degrees, students in Economics register higher percentage than the other two degrees.

Table 2 The distribution of the response variable Y, for gender (F, M) and bachelor degrees (BA, E, SBD)

Y	Category	%	F	M	BA	E	SBD
1	0 ECTS	17.23	10.87	20.67	18.56	15.65	17.54
2	from 1 to 39	36.86	34.94	37.90	37.13	36.03	38.25
3	from 40 to 59	26.94	29.04	25.80	22.15	31.44	28.77
4	60 ECTS	18.97	25.16	15.63	22.15	16.87	15.44
			35.11	64.89	44.06	40.40	15.54

The forest plot reveals the odds ratios of students' features from ordinal logistic regression (Figure 2), with the corresponding 95% confidence interval and p-values. This analysis gives us some interesting patterns between the cumulative probability of reaching the threshold and students' characteristics. Looking at our research questions as presented in Section 1, it is possible to sketch that the pre-enrolment test score (and also the diploma mark) might affect the cumulative probability, given that for every one unit increase in test score (or in diploma mark), the odds of having from 40 to 59 credits (versus having less 39 credits) slightly increases, holding constant all other variables. Then, a gender gap is confirmed. In fact, males have a lower cumulative probability of reaching the threshold than females, and the odds of being in group 40–59 credits is 0.73 times that of females. The difference between bachelor's degrees is not statistically significant.

Other interesting effects emerge from this analysis. Having additional education requirements (OFA) is associated with a lower probability of reaching 40 ECTS. An academic year trend is observed. The probability of having more than 40 ECTS decreases during the observation period. This should be caused by the Covid-19 pandemic. The type of high school affects the possibility of having at least 40 ECTS. In

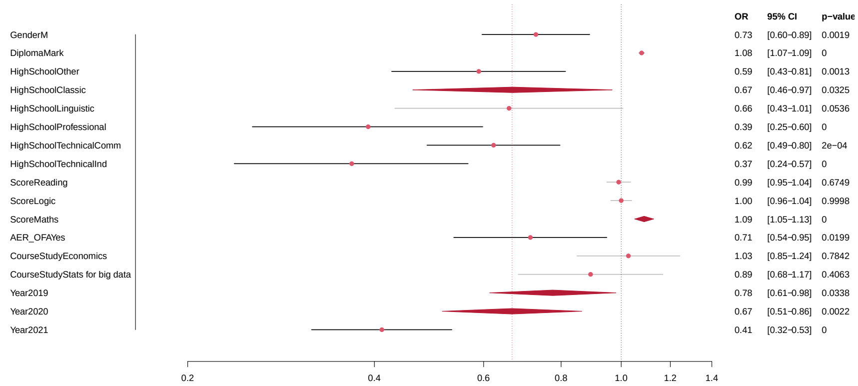


Fig. 2 Forest plot for ordinal logistic regression

particular, students from the scientific lyceum have a higher probability of reaching 40 ECTS, with respect to all other students.

References

1. Agresti, A.: Analysis of Ordinal Categorical Data. John Wiley & Sons, Inc. (2010)
2. Adelfio, G., Boscaino, G., Capursi, V.: A new indicator for higher education student performance. High. Educ. 68(5), 653–668 (2014)
3. Alyahyan, E., Düştegör, D.: Predicting academic success in higher education: literature review and best practices. Int J Educ Technol High Educ 17(3) (2020)
4. Attanasio, M., Boscaino, G., Capursi. V., Plaia, A.: May the students' career performance helpful in predicting an increase in universities income? In: Statistical Models for Data Analysis. Series in Studies in Classification, Data Analysis, and Knowledge Organization, P. Giudici, S. Ingrassia, and M. Vichi, eds., Springer International Publishing, Switzerland (2013)
5. Baker, R. S., Yacef, K.: The state of educational data mining in 2009: A review and future visions. J Edu Data Mining, 1(1), 3–17 (2009)
6. Boscaino, G., Capursi, V., Giambona, F.: The careers' performance of a University students' cohort. DSSM Working paper, n. 2007.1 (2007)
7. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Van Erven, G.: Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. J Bus Res, 94, 335–343 (2019)
8. Grilli, L., Rampichini, C., Varriale, R.: Predicting students academic performance: A challenging issue in statistical modelling. In: Cladag 2013 Book of abstracts, Monerva, Morlini and Palumbo, eds., CLEUP, Padova, pp. 249–254 (2013)
9. Grilli, L., Rampichini, C., Varriale, R.: Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts. Stat Model 16(1): 47–66 (2015)
10. Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.: Handbook of educational data mining, CRC Press, (2010)

An exploratory strategy for analyzing students' mobility data

Una strategia esplorativa per analizzare dati di mobilità studentesca

Ilaria Primerano and Giuseppe Giordano

Abstract This contribution deals with Italian students' decision to churn in the transition from the first to the second year of a bachelor's degree program. Based on these choices, three different churn scenarios are identified according to the decision to move towards a different university, to change degree programs, or both. Exploratory data analysis and Social Network methods are used to trace and visualize the links among universities defined considering these students' flows. The analysis is conducted considering students enrolled in universities located in the Campania region (Italy). The aim is to define an index of students' mobility flows considering their retention ability and students churn.
Abstract *Questo contributo analizza il tema della mobilità studentesca considerando la decisione degli studenti di trasferirsi nel passaggio dal primo al secondo anno di un corso di laurea triennale (churn risk). Da queste decisioni emergono diversi scenari di churn che dipendono dalla scelta di iscriversi in un'altra università, cambiare corso di laurea o entrambi. La Social Network Analysis permette di tracciare e visualizzare i legami tra le università definiti in base a questo specifico tipo di mobilità. L'analisi è stata condotta considerando gli studenti iscritti alle università situate nella regione Campania (Italia). L'obiettivo è definire un indice dei flussi degli studenti basato sulle decisioni di trasferimento e la capacità di ritenzione delle università.*

Key words: Churn risk, Correspondence Analysis, Directed Graphs, Origin Destination Matrix, Weighted Network

---

Ilaria Primerano
Dept. of Political and Social Studies, University of Salerno, e-mail: iprimerano@unisa.it

Giuseppe Giordano
Dept. of Political and Social Studies, University of Salerno, e-mail: ggiordano@unisa.it

1 Introduction

Tracing the main patterns of students mobility in higher education is one of the key issues in territorial policy. In fact, both the entire university and territorial governance systems are interested in understanding the factors behind students' decisions and outcomes with respect to university and degree program choice, and also inherent in post-enrollment choices such as inter-universities and/or degree program re-locations, and dropping out.

There are many factors that influence students' individual choices referred to the transition either from high school to bachelor's degree (first level mobility) or from bachelor's degree to master's degree (second level mobility). Sometime, these choices could lead to drop or churn out from the university system also in year-to-year transitions.

In literature, several studies have analyzed different aspects related to student mobility. Some scholars have dealt with outlining the typical patterns of student mobility in Italy, identifying the main trajectory from North to South [1, 2]. Others have focused on the definition of university attractiveness indices based on students flows among universities and territories [3, 4], also trying to highlight the push and pull factors behind the Italian students choices [5, 6]. Despite Universities implemented several measures to improve students' retention, some studies have identified important risk factors in university dropout [7, 8] and churn decisions [9, 10].

Moving from this framework, in this contribution we focus on students' churn decisions in the transition from the first to the second year of the Bachelor's degree program, defined on the bases of their choice to change disciplinary field and/or to move from the university of first enrolment. Looking at this churn decisions, three different settings could be observed: i) change Bachelor's degree program (i.e., moving from a Bachelor's degree program of one university to another Bachelor's degree program of the same university); ii) change university (i.e., moving from a Bachelor's degree program of one university to the same Bachelor's degree program in another university); and iii) change Bachelor's degree program and university (i.e., moving from a Bachelor's degree program of one university to another Bachelor's degree program in another university). Moreover, all these kinds of churn could be observed in Bachelor's, Master's and single-cycle degree programs, for all year-to-year transitions.

This contribution aims to define a quantitative measure that characterizes the flows between pairs of universities arising from students' choices. The use of Factorial Methods and Social Network Analysis allow to visualize the connections among universities according to students' flows. Based on a previous study [11], the proposed index weights a network characterizing student churn activity accounting for both the size of the university considered as a proxy of retention, and flows direction based on churning choices.

2 The churn propensity index

In this contribution we introduce the definition of a churn propensity index based on students' mobility flows among universities. We rely upon administrative data on the Italian University students collected from MOBYSU.IT database[1].

The analysis is conducted considering only students enrolled in a Bachelor's degree program. We consider only freshmen in the cohort 2018-2019 enrolled for the first time in a university who decide to churn in the transition from the first to the second year of their Bachelor's degree program. At this stage, all the disciplinary fields of the Bachelor's program are jointly considered.

Data are organized into a contingency table that considers on the rows the universities of enrollment (origin) and on the columns the universities (destination) chosen by students who changed university after the first year. In the scope of Social Network Analysis [12], this matrix can be seen as an origin-destination adjacency matrix. Its entries are interpreted as flows of students among universities. Consequently, when reading flows by rows, we refer to churning out students; while reading flows by columns, we refer to incoming students.

Let us notice that the sum of all the off-diagonal elements accounts for the total amount of students who decide to churn; while the sum of the elements on the main diagonal accounts for the total amount of stayers. Specifically, for each university, the corresponding element on the main diagonal is a proxy of its retention capability, the off-diagonal elements summed-up by each row gives the number of churning students as a degree of university disenrollment risk; while the off-diagonal elements summed-up by each column gives the number of incoming students as a degree of university attractiveness. In the language of social network analysis, these two quantities correspond to the nodes' out-degree and in-degree.

From this adjacency matrix, we obtain a one-mode, directed and weighted network representing the flows of churning students, where the nodes are the universities, and the edges weight depend on the occurrences of students that decide to churn. Thus, the origin nodes are the universities selected by students as first choice at the time of enrollment (i.e., in 2018), while the destination nodes are the universities choice after the churn decision (i.e, in 2019).

Thus, we propose a global churn propensity index as the ratio between the total number of churning students and the total number of stayers. This index can be specialized for each university as the ratio between the sum of the off-diagonal elements in a row and the corresponding diagonal element. Furthermore, we can specify this index also for each pair of universities as the ratio between each cell and the corresponding diagonal element in a row.

[1] Data drawn from the Italian 'Anagrafe Nazionale della Formazione Superiore' has been processed according to the research project 'From high school to the job market: analysis of the university careers and the university North-South mobility' carried out by the University of Palermo (head of the research program), the Italian 'Ministero Università e Ricerca', and INVALSI

2.1 Some evidences from the Campania Region

In this contribution, we focus on students enrolled in a Bachelor's degree program at the universities located in the Campania region, in Italy. The seven universities in Campania, according to their main geographical province of reference, are: University of Naples Federico II, University of Naples "L'Orientale", Suor Orsola Benincasa University of Naples, and Parthenope University in provincce of Naples, University of Campania Luigi Vanvitelli in province of Caserta, University of Sannio in province of Benevento; and University of Salerno in province of Salerno. Moreover, we do not consider students enrolled in telematic universities, single-cycle and health professions degree programs.

The adjacency matrix holding the seven universities in the region counts the flows of 18,894 students enrolled in the academic year 2018-2019, of whom the 96% are stayers in the next year (a.y. 2019-2020). Only 4% of the students in this cohort churned after the first year, moving from the university of origin in 2018 toward a different university in 2019, regardless the disciplinary field.

We compute the churn propensity indices introduced in section 2 for all the university in the Campania region, for each single university, and for all possible pairs of universities. In the period under analysis, the global churn propensity index is equal to 4.18%, and it confirms a fairly low global propensity to churn. However, some differences emerge among the seven universities. Looking at the churn propensity indices computed for each university, the lower value is observed for the University of Salerno (1.8%), while the University of Sannio (12.5%) shows the higher one. Instead, by considering the indices computed for each pair of universities, relevant values are observed for the University of Sannio towards the universities of Salerno (4.3%), Naples Federico II (4.1%), and Vanvitelli (4.1%). Moreover, the highest churn values are observed from all universities in Campania towards the University of Naples Federico II; and from this latter towards Vanvitelli (2.2%) and Parthenope (1.3%) universities.

In the graph in Figure 1 (left side), both the churn flows and the role of each university are showed as a network; the size of the node-university is proportional to its churn propensity index, while the width of the edges depends on the occurrences of students flows linking pairs of universities

Furthermore, on the right side of Figure 1 there is the Biplot of the Correspondence Analysis computed for the matrix holding the churn propensity indices for all pairs of universities. Here, the proximity among origin universities (blue) shows similar churning out profiles and are displayed as barycenters of the universities of destination (red).

In conclusion, the graph representation reveals the central role of the University of Sannio in terms of churning students towards other universities (disenrollment risk); and the central position of the University of Naples Federico II who receives students from all the other universities (attractiveness). The important role of the University of Naples Federico II appears evident also in the biplot of the Correspondence Analysis, for both its role of origin and destination of the flows. In fact, in the first quadrant, the central position of the University of Naples Federico II (blue) can be seen among the universities chosen by its churning students (Parthenope, Vanvitelli, Salerno and Orientale universities). While, in the third quadrant, we note the proximity between all the other universities in Campania whose students move towards the University of Naples Federico II (red) to continue their academic studies.

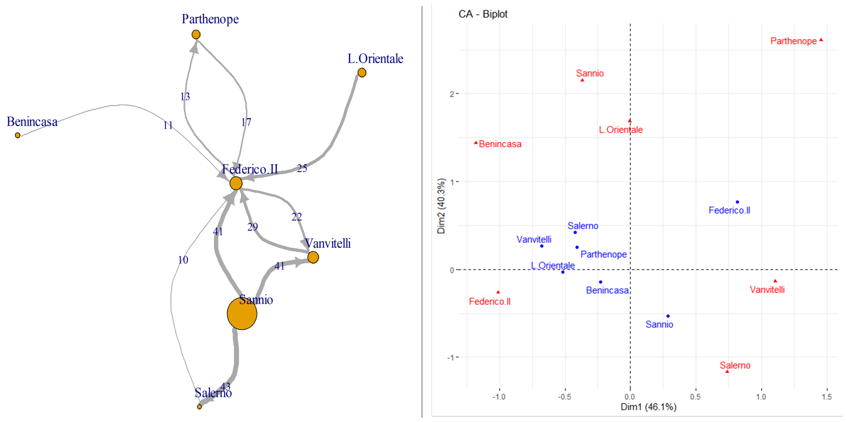


Fig. 1 Graphical representation of the churn-retention propensity index among the seven universities in the Campania region. Left side: network of churning students flows. Right side: Biplot of Correspondence Analysis of the churn propensity index.

3 Concluding remarks

This contribution proposes an exploratory strategy of analysis based on a numerical measure, a visual tool and a graph reading to discover university roles and students attitude to churn out from university of enrollment in the transition from the first to the second year of the Bachelor’s degree program. Specifically, we propose a quantitative measure of university churn propensity based on students individual choices, defined at different levels of analysis, which allows us to identify the most relevant paths among universities according to students outgoing flows. The proposed approach considers a relative measure of churn propensity that take into account the retention capability of each university. It is assumed that churning flows are more politically relevant in those situation where the retention capability is low with respect to the whole amount of enrolled students. We presented a case study on the propensity to churn of students enrolled in universities in the Campania region in order to show the applicability of the proposed indices. Our analysis highlights some differences between universities located in the same territory, made even more evident by the proposed graphical representations. On the one hand, the network highlights the central position of the universities that receive the largest number of students who decided to churn, while the presence of many students leaving the origin universities is rendered by the size of the node-university. On the other hand, the biplot of the Correspondence Analysis, allows us to identify the universities with the most similar profiles, as well as showing the central position of the origin universities compared to the destination ones.

The churn propensity index can be further specified by considering, for example, students flows occurring within the same university, or among universities located into different territorial aggregations, and/or the flows within and between different degree programs, grouped according to the ISCED-F classification. It can also be taken into account in analyzing the attractiveness of the universities by considering the enrollments of churning students. Therefore, the attractiveness of the degree programs offered and the areas in which they are geographically located could also be analyzed. In general, it is best suited in all situation where the retention capability are relatively different with respect to the size (number of students enrolled) of the universities.

Acknowledgements This contribution has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide", n. 2017HBTK5P—CUP B78D19000180001.

References

1. Attanasio, M., Enea, M., Albano, A.: Dalla triennale alla magistrale: continua la fuga dei cervelli dal Mezzogiorno d'Italia. Neodemos, ISSN: 2421–3209 (2019)
2. Genova, V. G., Tumminello, M., Enea, M., Aiello, F., Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electron. J. App. Stat. Anal., 12(4), 774–800 (2019)
3. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M. P.: Correction to: Analysing the determinants of Italian university student mobility pathways. Genus,78(1) (2022)
4. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Geography of Italian student mobility: A network analysis approach. Socio-Economic Planning Sciences, 73, 100918 (2021)
5. Columbu, S., Porcu, M., Sulis, I.: University choice and the attractiveness of the study area: insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models. Socio-Economic Planning Sciences, 74, 100926 (2021)
6. Santelli, F., Ragozini, G., Vitale, M.P.: Assessing the effects of local contexts on the mobility choices of university students in Campania region in Italy. Genus 78(5) (2022)
7. Contini, D., Salza, G.: Too few university graduates. Inclusiveness and effectiveness of the Italian higher education system. Socio-Economic Planning Sciences, 71, 100803 (2020)
8. Sarra, A., Fontanella, L., Di Zio, S. Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. Social Indicators Research 146, 41-60 (2019)
9. La Rocca, M., Niglio, M., Restaino, M. Predicting university students' churn risk, In: Lombardo, R., Camminatello, I., Simonacci, V. (eds.) Book of short Papers IES 2022 Innovation & Society 5.0: statistical and economic methodologies for quality assessment, PKE srl (2022)
10. Primerano I., Santelli F., Usala C., Ragozini, G.: Exploiting students' inter-degree relocations to assess Italian universities' attractiveness. In Book of Abstracts 9th International Conference on Risk Analysis, ISBN: 978-972-674-919-6, pp. 60–62 (2022)
11. Giordano, G., Primerano, I., Vitale, P. A Network-Based Indicator of Travelers Performativity on Instagram. Social Indicators Research, 156(2), 631-649 (2020)
12. Wasserman, S., Faust, K. Social network analysis: Methods and applications, Cambridge University Press (1994)

Solicited Session SS4 - *Statistics in football*

Session supported by BDsports, ISI Groups on Sports and Math&Sport organized by Rodolfo Metulini and Marica Manisera
Chair: Marialuisa Restaino

1. *Community Detection in Sport Market Networks: The Case of Italian Professional Football* (Rondinelli R. and Ievoli R.)
2. *An Original Application to Football of PLS-SEM for the xG Model* (Cefis M. and Carpita M.)
3. *Performance Assessment of Football Players and Combined Permutation Tests with application to Home-Field Advantage* (Bonnini S.)
4. *A First Proposal of the Triad Census for Weighted Networks: an Application to Football* (Rondinelli R. and Palazzo L.)

Community Detection in Sport Market Networks: The Case of Italian Professional Football

Community Detection per il mercato sportivo: il caso del calcio professionistico italiano

Roberto Rondinelli and Riccardo Ievoli

Abstract Network Analysis may be a useful tool to measure the key elements of the football transfer markets and community detection algorithms can be proposed to unveil unobservable patterns. This work explores the applicability of the "walktrap" algorithm considering the flows of player transfers generated by the main 40 Italian professional football teams. We consider the summer market session at the beginning of the season 2022-2023. Results help to understand some peculiarities of the Italian football transfer market and the different approaches of professional teams.
Abstract *L'analisi delle reti può essere un utile strumento per analizzare i principali elementi del mercato dei calciatori e gli algoritmi di community detection possono essere proposti per identificarne le principali relazioni. Questo lavoro intende esplorare l'applicabilità dell'algoritmo "walktrap" prendendo in considerazione i flussi generati dalle principali 40 squadre del calcio professionistico italiano. La sessione di mercato considerata è quella dell'inizio della stagione 2022-2023. I risultati mostrano alcune particolari caratteristiche del mercato italiano e alcune differenze nei comportamenti delle squadre professionistiche.*

Key words: Football Market, Transfer Network, Walktrap Algorithm

1 Introduction

In recent years, Network Analysis (NA) became a well-known and useful tool to visualize and analyse the transfer flows of players in professional football [1, 5]. NA helps to visualize and identify key features of interactions classifying them based on the relationships within a complex system. A recent study applies a set of network indices to 400 football clubs worldwide, finding that clubs acting as hubs or brokers usually achieved better football outcomes or performance in their competitions

---

Roberto Rondinelli
University of Naples Federico II, e-mail: roberto.rondinelli@unina.it

Riccardo Ievoli
University of Ferrara, e-mail: riccardo.ievoli@unife.it

[4]. In this sense, the football market exhibits a small-world structure where each club can easily reach another club in a few steps moving towards its network neighbours [5]. Although the football market can appear as a cohesive network [3], clubs' strategies and interactions can lead to close or weak collaborations that can occur more often than expected [2, 9], highlighting the presence of groups and clusters.

Starting from these considerations, the detection of possible emerging communities from the football transfer market can be useful to identify recurrent patterns. Thus, the aim of this work is to study the applicability of a suitable community detection method, denoted as *walktrap* [7], proposed for unveiling the structures market relationships of the main Italian professional football teams with their commercial partners both in Italy and abroad. To the best of our knowledge, the application of community detection methods has been not fully explored in football transfer networks. A first proposal can be found in [6]. To this end, we use the network of players' flows generated by the 40 Italian professional football teams (who compete in first and second divisions), involving also their connections from all over the world. Data used in the empirical application refer to the summer market session of 2022, at the beginning of the season 2022-2023. The rest of the paper is organised as follows. Section 2 depicts NA in players' transfer market and community detection method, i.e., the *walktrap* algorithm, Section 3 is dedicated to the illustration and the discussion of the results concerning the empirical application to Italian professional football, while Section 4 contains some remarks and possible advances.

2 Materials and Methods

Transfers between football teams can be formally defined by a weighted, directed, and attributed network $\mathscr{G}$ which consists of a 5-tuple $\mathscr{G}(\mathscr{V},\mathscr{E},\mathscr{W},\mathscr{A}_v,\mathscr{A}_e)$, where $\mathscr{V}=\{v_1,v_2,\dots,v_N\}$ is the set of the N teams involved in the transfer network; $\mathscr{E}\subseteq\mathscr{V}\times\mathscr{V}$ is the set of directed edges between the teams, i.e., each edge $e_{ij}\neq e_{ji}$ is an ordered couple indicating a transfer of a player from v_i (the selling club) to v_j (the buying club); $\mathscr{W}:\mathscr{E}\to\mathbb{N}$, with $\mathscr{W}(e_{ij})=w_{ij}\in\mathbb{N}$, is the number of directed transfers between v_i and v_j; $\mathscr{A}_v$ is the attribute matrix of the N teams, such as the belonging league, the nationality, etc.; finally, $\mathscr{A}_e$ indicates the attribute matrix associated with a transfer e_{ij}, which includes information like the type of transfer, the fee, characteristics of players (age, height, weight, etc.) and their market value.

The transfers market, as defined above, can be divided into groups of teams that share structural similarities. Among the available community detection algorithms, the *walktrap* exploits the properties of random walks defining a distance between nodes, thus it is well suited for the case of directed and weighted networks [7]. The algorithm considers each node as a starting community and performs a hierarchical aggregative clustering in the following points: i) compute the distance between communities; ii) aggregate two communities using Ward's method; iii) update the distances between communities. The number of communities, their composition, and the quality of the final partitions depend on the number of random walker steps.

The position and characteristics of each community inside the transfer network can be summarized using descriptive indices at the node level [8]. Here, the following indices are considered and are intended in their *in* and *out* formulations:

- **degree**: number of market partners of team v_i
- **weighted degree**: number of market transfers of team v_i
- **weighted degree and degree ratio (WDDr)**: number of market transfers per partner of team v_i.

The degree centrality and its derivations summarise the main features of a node and, as described above, assume an important meaning in our case. For this reason, these indices, together with the attributes contained in the matrix $\mathscr{A}_e$, are useful to describe the behaviour of the teams included in each community.

3 Empirical Application

Data for the application refer to the 2022 Italian football transfer market session of season 2022-2023, between July 1st and September 1st. The data concerns 40 Italian teams of the first two main divisions (Serie A and Serie B) and have been retrieved from the *Transfermarkt*[1] website. For each transfer, some attributes are considered: a) the age of the transferred player (in years), b) the height of the transferred player (in meters), c) the market value of the player at the end of the market session (in euros), and d) the type of transfer (direct acquisition with a fee, free transfer, loan transfer, end of the loan, the loan with a fee). Loans can also present a redemption right (buy-back clause). In this case, when the loan expires, two mutual links are created: the first is generated when the player comes back to the lender team and the second when he is definitely sold to the borrower team.

Figure 1 summarizes the results of the application of *walktrap* algorithm with respect to the number of steps for the random walker. The most satisfactory combinations maximizing the modularity index (red line) and also keeping reasonably low the number of communities (green line) are those between 9 and 12 steps. We compare the results, in terms of communities, obtained through these four combinations using the adjusted Rand index, ranging from a minimum of 0.59 (comparison between 9 and 12 steps) to a maximum of 0.76 (comparison between 11 and 12 steps). Then, for the sake of brevity, we discuss the results generated by the *walktrap* algorithm performing 9 steps of the random walker.

The network generated by Italian professional football consists of 472 teams and their market interactions (2136 flows) and detected communities are depicted in Figure 2. Analyzing the communities, six of them include only one team competing in the first division (Serie A) together with its own commercial partners out of professional Italian football. These teams are Bologna, Lazio, Roma, Sassuolo, Udinese and Verona and are labelled as COM6, COM9, COM4, COM11, COM1,

[1] https://www.transfermarkt.com

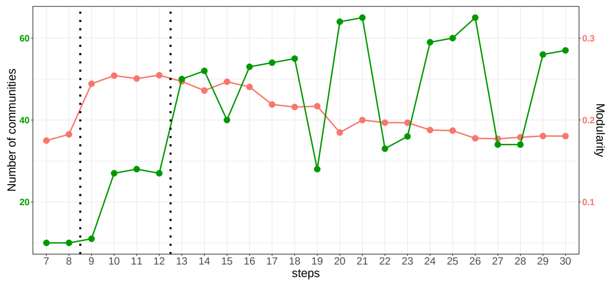


Fig. 1 Comparison between Modularity index and the number of communities through the number of steps of the *walktrap* algorithm.

and COM10. Instead, another community involve one single team of the second Division (Venezia, labelled as COM8) and related teams. In communities guided by one team, the overall number of teams involved ranges between 15 (Verona) and 23 (Bologna). A further community include two teams, one per division (Napoli and Parma - COM7), and their commercial partners for a total of 31 teams. Conversely, the last three communities are more heterogeneous. The first of them (labelled as COM2) includes seven teams of professional Italian football (Fiorentina, Salernitana, and Sampdoria from Serie A, and Ascoli, Benevento, Brescia, and Pisa from Serie B) and their commercial partners for a total of 85 teams. Then, the second (denoted as COM3) involves 47 teams of which 3 are in the First Division (Juventus, Lecce, and Milan) and 3 are in the Second Division (Bari, Cittadella, and Palermo). The third community (COM5) is the most numerous, containing a total of 169 teams, with 18 of them belonging to the First and Second Divisions (7 and 11, respectively) of Italian professional football.

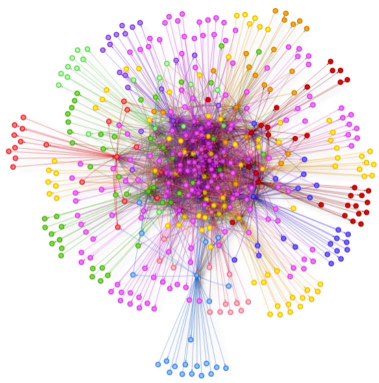


Fig. 2 Network generated by transfer flows of the main 40 Italian professional football teams (Summer Market session season 2022-2023).

With the purpose of describing the behaviour of the 40 main Italian professional teams, we discuss the results of the previously introduced network indices referred to them. Firstly, professional Italian teams involved in COM3 exhibit, on average, the lowest number of relationships (mean of the degrees equal to 37, median equal to 40) compared to other communities. Within teams of COM3, the Second Division team of Cittadella reaches the minimum degree (16) while the maximum is obtained by Lecce and Milan (45). This kind of "moderation" in the flows also emerges from the observed weighted degree (average of six teams equal to 50.5, median equal to 56). The behaviour of COM2 and COM5 emerges from the analysis of the WDDr, i.e., the average number of players exchanged per team. In particular, the 18 Italian professional teams of COM5 show, on average, the highest WDDr (mean equal to 1.60, median equal to 1.62). Considering the inbound and outbound transfers also emerges that Italian professional teams of this community receive and give away 1.66 and 1.56 players per team, respectively. Within these teams, the highest WDDr is reached by the Serie A team Atalanta which trades almost 2 players per team (1.83). This team also presents the highest in-WDDr, receiving more than 2 players per team (2.05), followed by Frosinone and Reggina (1.92 and 1.91, respectively) also involved in COM5. Indeed, the highest out-WDDr are reached by Cremonese and Monza (2.18 and 1.95, respectively) who compete in Serie A and are classified in COM5. Similar considerations can be carried out considering the Italian professional teams of COM2 because teams such as Brescia and Benevento (Second Division) share a high out-WDDr (1.93 and 1.74, respectively).

Peculiarities of such communities also arise from the analysis of the auxiliary variables referred to the transfers of teams included in each community. Regarding the age of players, COM9 and COM7 feature older players both inbound and outbound (median between 25 and 26 years), while younger players are transferred in COM6 and COM11. On average, transfers involving taller players occur in COM8 and COM1 (median height equal to 1.85 meters). For what concerns the market value, COM2 and COM5 present, on average, inbound transfers with lower value compared to other communities (average equal to 1.4 and 1.5 million euros, respectively, median equal to 0.35 and 0.4). The highest market values are observed for COM4 (mean equal to 5.7 million euros, median equal to 3.5). Considering the type of transfer, it is interesting to note that COM2 and COM5 share the highest quantity (on average) of end of the loans for what concerns the inbound transfers. Regarding outbound transfers, loans are more often used in COM5 with respect to other communities. Finally, COM2 COM5 and COM3 present fewer inbound acquisitions with a fee, while for COM2 and COM5 this evidence is confirmed also considering outbound acquisitions with a fee.

4 Discussion and Concluding Remarks

The identification of the market behaviours of the professional football team represents an interesting challenge, and NA can help to unveil some unknown structures through the use of proper community detection methods. Applying the *walktrap* algorithm on the network generated by the main professional football teams in Italy, we tried to profile the market behaviours of such teams.

Three main market behaviours/strategies emerge from the analysis of communities. Firstly, a non-negligible part of the teams competing in the First division results quite isolated from the flows occurring between Italian professional teams, preferring to relate to its own sub-network often composed of youth teams, foreign clubs or teams competing in the minor leagues of Italy. Their behaviour is heterogeneous and derives from the market strategy that each of them carries out. Secondly, competition and/or cooperation arise from the analysis of COM2 and COM5, especially because these communities involve teams of both the First and Second divisions (this also explains the low market value of the transfers, on average) and the number of flows occurring within teams is high with respect to other communities. This is more evident considering the inbound transfers for COM5 and outbound transfers for COM2. In addition, loans are usually more frequent than in other communities. Thirdly, teams involved in COM3, including three teams from the South of Italy and two relevant clubs such as Juventus and Milan, reveal a sort of selection of their interlocutors on the market, making less use of inbound acquisitions with a fee.

Further extension of this work may regard the use of temporal networks [5] to verify the presence of these behaviours/strategies among several market sessions.

References

1. Batagelj, V., Doreian, P., Ferligoj, A., Kejzar, N.: Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution (Vol. 2). John Wiley & Sons (2014)
2. Bond, A. J., Widdop, P., Chadwick, S.: Football's emerging market trade network: Ego network approach to world systems theory. Manag. Sport Leis., 23(1-2), 70–91 (2018)
3. Friedkin, N. E.: Structural cohesion and equivalence explanations of social homogeneity. Sociological Methods & Research 12(3), 235-261 (1984)
4. Liu, X. F., Liu, Y. L., Lu, X. H., Wang, Q. X., Wang, T. X.: The anatomy of the global football player transfer network: Club functionalities versus network properties. PloS one, 11(6), e0156504 (2016)
5. Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S. L., Ortega, G. J.: Transfer market activities and sportive performance in European first football leagues: A dynamic network approach. PloS one, 13(12), e0209362 (2018)
6. Palazzo, L., Rondinelli, R. Clemente, F.M., Ievoli, R., Ragozini, G.: Community structure of the football transfer market network: the case of Italian Serie A. *Working Paper*
7. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algorithms Appl., 10(2), 191–218 (2006)
8. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press (1994)
9. Xu, Y.: The formation mechanism of the player transfer network among football clubs. Soccer Soc., 22(7), 704–715 (2021)

An Original Application to Football of PLS-SEM for the xG Model

Un'Applicazione Originale al Calcio del PLS-SEM per il Modello xG

Mattia Cefis and Maurizio Carpita

Abstract In the field of football analytics, we want to improve (in terms of prediction performance) one of the emerging tools: the expected goal (xG) model. With this final aim, we merged match event data with some players' performance composite indicators obtained using a Partial Least Squares - Structural Equation Model (PLS-SEM). Using a sample of match tracking data (season 2019/2020) of the Italian Serie A, a logistic regression model was applied on different scenarios for sample balanced techniques. Results seem to be interesting in terms of sensitivity, and others metric indices, compared with a benchmark (Understat). In addition, some performance composite indicators obtained by the PLS-SEM and some tracking variables are significant for the xG model.
Abstract *Tramite questo lavoro si vuole migliorare (in termini di accuratezza nella predizione) il modello expected goal (xG), un modello emergente utilizzato per "misurare" la pericolosità di una squadra. Partendo da questo obiettivo, sono stati messi insieme i dati "evento" delle partite con degli indicatori di performance ottenuti attraverso il Partial Least Squares - Structural Equation Model (PLS-SEM). Infine, sincronizzando tali dati con un campione di tracking-data relativo al campionato di Serie A 2019/2020, un modello di regressione logistica è stato applicato a diversi scenari con campioni bilanciati. I risultati preliminari sembrano essere interessanti e significativi in termini di performance del modello confrontati con un riferimento (Understat).*

Key words: PLS-SEM, Expected Goal, Logistic Regression, Imbalanced Sample

---

Mattia Cefis
University of Brescia, Department of Economics and Management, e-mail: mattia.cefis@unibs.it

Maurizio Carpita
University of Brescia, Department of Economics and Management, e-mail: maurizio.carpita@unibs.it

1 Introduction

Football (or soccer, in North America, Australia and New Zealand) is one of the most popular sports in the world. Due to its appeal, football teams are treated more and more as firms: as a consequence, decisions about their management (coach, technical staff and scouting) are becoming strategic, and, for this reason, in the last few years football has been moving towards a data-driven revolution [4]. By this work, the idea is to refine and improve, in terms of prediction accuracy, the well-known expected goal (xG, [1]) model that surpasses the most basic and frequently used metric in football to summarise the team performance: the shot; indeed, it can be a misleading metric since it does not consider the quality of the goal-scoring opportunity from which it arises.

2 Literature overview and data employed

The main idea of the xG model is to assign a value between zero and one to each shot; this value represents the probability of a shot resulting in a goal, using a machine learning probabilistic classifier [15]. During the last years, the xG model has become increasingly popular, and it is used more and more in the football world as a proxy for measuring players' finalisation performance and teams' offensive strength during a match [9]. For this reason, some studies and websites have treated this topic: for example, someone [14, 17] examined shots, taking in consideration only distance and angle to goal, whereas another one [9] made a spatial analysis of shots of the Mayor League Soccer, using a logistic regression. Another recent work [16] tried to quantify the effectiveness of defensive playing styles in the Chinese Football Super League by using xG. The main deficiency is that xG models are based just on event data and do not take into account players' features. As an innovation, we want to improve the xG model by adding some composite indicators related the players' performance and obtained by a Partial Least Squares Structural Equation Model (PLS-SEM, [10]), in order to take into account shooters' and goalkeepers' features [5, 6] (Fig. 1). In order to reach the aim, a merging between data coming from different sources (e.g. Understat[1] for event data, Math&Sport for tracking data and *Sofifa*[2] for building the players' performance indicators) was done. The final dataset was composed of a sample of 600 shots, 23 variables (e.g. 1 binary outcome and 22 regressors) coming from 50 official football matches coming from the season 2019/2020 of the Italian Serie A league.

[1] www.understat.com

[2] www.sofifa.com

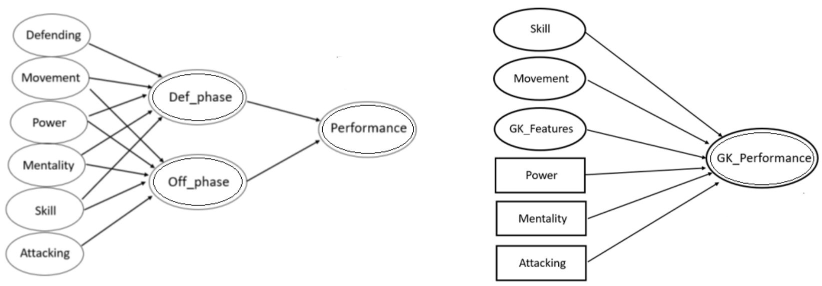


Fig. 1 Players' movement and Goalkeepers models used for creating the composite indicators with the PLS-SEM approach.

2.1 The frameworks developed

From a methodological point of view, since the target, namely the Goal, is a rare event [14], a logit model (LM) with parameters estimated by maximum likelihood [12] was applied on different samples for three situations: the basic Imbalance sample and two machine learning sample-balanced techniques, such as random oversampling examples (ROSE, [13]) and the synthetic minority oversampling technique (SMOTE, [7]). The benchmark adopted was the xG value provided by Understat. From a practical point of view, it was developed a routine in *R* by using a stratified 3-Fold cross validation for evaluating the model fit and computing the performance measures, with 5,000 replications (*R* packages *ROSE* and *RSBID*).

The LM model was preferred because of its easy interpretation concerning the regressors effects and since the real focus was to introduce new predictors in the xG model, in order to improve the goal probability estimation. In the context of the xG, this model lets to estimate the conditional probability of goal for any shots and their set of features values $\mathbf{X}$, and estimate parameters $\hat{\beta}$ in (1). Note that the regression coefficients are estimated by maximum likelihood [12]. Typical classification metrics have been used to assess the models performance [11].

$$xG = P(Goal|\mathbf{X}) = \frac{e^{\mathbf{X}\hat{\beta}}}{1+e^{\mathbf{X}\hat{\beta}}} \tag{1}$$

In addition, when we balance a dataset thanks some techniques like ROSE or SMOTE, we must take into account the effects of our modifications to the training data [2, 8]. For the Bayes' Theorem, posterior probabilities are proportional to the prior ones, which can be estimated as the relative frequency in each category. Therefore, the estimated posterior probabilities (expected goal) obtained using artificially balanced data set can be corrected (calibrated, xG^*) using the following formula (2):

$$xG^* = \frac{\frac{0.1}{0.5}xG}{\frac{0.1}{0.5}xG + \frac{(1-0.1)}{(1-0.5)}(1-xG)} \tag{2}$$

Take into account that in our case 10% and 50% are respectively the real and the artificial (balanced) sample proportion of the rare class.

3 Results and discussion

As preliminary results, the main significant estimated regressors are the shooter position, the angle of shot, and some new variables introduced: the number of opponents' players between the shooter and the goal, the goalkeeper position and some performance features like the shooter movement ability, the goalkeeper mentality and the goalkeeper skill.

Then, for binary classification problems usually the probabilistic classifier (xG) is transformed in the categorical one (Goal, NoGoal) using the threshold 0.5. In Table 1 are proposed all the classification performance metrics and their average scores (5,000 replications), comparing them with the benchmark (that is a punctual value, directly provided from Understat). Take in consideration that asterisks in Table 1 must be interpreted as a value statistically significant different (e.g. * = 5%, ** = 2.5%, *** = 1%) from the benchmark; the metrics that outperform the benchmark are emphasized in bold. Both ROSE and Imbalance significantly outperform in terms of specificity and precision the benchmark (Understat) whereas SMOTE seems able to better detect the goals (sensitivity equals to 0.36 vs 0.16 of the benchmark) and to improve Understat in terms of F1 (the armonic mean between precision and sensitivity) and precision. The Area Under the Receiver Operating Characteristic Curve (AUC) metric is similar for all the situations, except for the Imbalance, that significantly outperforms Understat (0.74 vs 0.72).

Table 1 The performance classification metrics averaged after 5,000 replications compared with the benchmark (classification threshold=0.5).

Metric	ROSE	SMOTE	Imbalance	Understat
Accuracy	0.89*	0.86***	0.90	0.91
Sensitivity	0.14*	**0.36*** **	0.15	0.16
Specificity	**0.98***	0.91***	**0.98***	0.96
Precision	**0.35***	**0.32****	**0.51*** **	0.22
F1	0.16	**0.33*** **	0.23	0.19
AUC	0.72	0.73	**0.74***	0.72

Now, in order to emphasize the importance of the new regressors introduced in the model, we propose one real goal, comparing its expected goal for each framework and introducing some variation, in order to better understand how the xG changes.

In the first real case (Fig. 2) we propose a goal scored from a high distance, in a situation with a high number of opponents in front of the shooter. Then we propose the expected goal for each situation (xG^* for ROSE and SMOTE) and others two scenarios (Table 2): the first one putting a top player as shooter (Ronaldo), whereas the second one leaving the same top player as shooter (Ronaldo, Juventus) and moving a normal goalkeeper. We can see how in the real scenario our balanced frameworks increase the goal prediction accuracy (higher xG than the benchmark); the xG for the imbalance approach is very similar. It's interesting to note that if we introduce a top player as shooter (Scenario 1), then a normal goalkeeper (Scenario 2) the expected goal increase in both three situations, emphasizing the importance of introducing players' performance indices in the model, as innovation of this work.

Fig. 2 Real case: goal from the distance

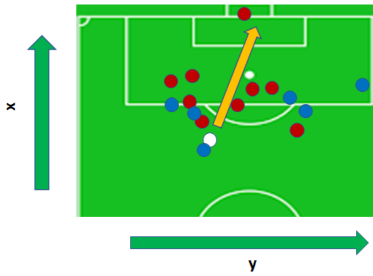


Table 2 The expected goal for each situation and different cases

Case	Shooter	GK	xG^* ROSE	xG^* SMOTE	xG Imbalance	xG Understat
Real	Soriano (Bol)	Handanovic (Int)	4.1%	2.3%	2.0%	2.1%
Scenario 1	Ronaldo (Juv)	Handanovic (Int)	5.4%	3.3%	3.1%	2.1%
Scenario 2	Ronaldo (Juv)	Skorupski (Bol)	7.1%	4.4%	4.5%	2.1%

As summary, the original approach presented in this work seems to suggest that some performance composite indicators (developed by PLS-SEM) and some tracking variables are helpful to detect the goals, improving the benchmark model. As future works, it should be interesting to examine in-depth this topic by a larger sample size, and maybe comparing other classification model (for example, Gompit [3]) performances.

Acknowledgements

We want to thank Math&Sport s.r.l. for providing us the tracking data for the Italian Serie A matches used in the statistical analysis. This research was supported by the project Big Data Analytics in Sports of the University of Brescia (https://bodai.unibs.it/bdsports).

References

1. Anzer, G., Bauer, P.: A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Frontiers in Sports and Active Living 3, 53 (2021)
2. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
3. Cameron, A.C., Trivedi, P.K., et al.: Microeconometrics using stata, vol. 2. Stata press College Station, TX (2010)
4. Cefis, M.: Football analytics: a bibliometric study about the last decade contributions. Electronic Journal of Applied Statistical Analysis 15, 232–248 (2022)
5. Cefis, M., Brentari, E.: Formative vs reflective constructs: a CTA-PLS approach on a goalkeepers' performance model. Book of the Short Papers, 51st Scientific Meeting of the Italian Statistical Society , 323–328 (2022)
6. Cefis, M., Carpita, M.: The higher-order PLS-SEM confirmatory approach for composite indicators of football performance quality. Computational Statistics , 1–24 (2022)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. Journal of artificial intelligence research 16, 321–357 (2002)
8. Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G.: Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE symposium series on computational intelligence, pp. 159–166. IEEE (2015)
9. Fairchild, A., Pelechrinis, K., Kokkodis, M.: Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. Journal of Sports Analytics 4, 165–174 (2018)
10. Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S.: Partial least squares structural equation modeling (PLS-SEM) using r: A workbook (2021)
11. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5, 1–12 (2015)
12. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer (2013)
13. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery 28, 92–122 (2014)
14. Rathke, A.: An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise 12, 514–529 (2017)
15. Robberechts, P., Davis, J.: How data availability affects the ability to learn good xG models. In: U. Brefeld, J. Davis, J. Van Haaren, A. Zimmermann (eds.) Machine Learning and Data Mining for Sports Analytics, pp. 17–27. Springer International Publishing, Cham (2020)
16. Ruan, L., Ge, H., Shen, Y., Pu, Z., Zong, S., Cui, Y.: Quantifying the effectiveness of defensive playing styles in the chinese football super league. Frontiers in Psychology , 1–10 (2022)
17. Umami, I., Gautama, D.H., Hatta, H.R.: implementing the expected goal (xG) model to predict scores in soccer matches. International Journal of Informatics and Information Systems 4, 38–54 (2021)

Performance Assessment of Football Players and Combined Permutation Tests with application to Home-Field Advantage

Valutazione della prestazione dei calciatori e test di permutazione combinati con applicazione al fattore campo

Stefano Bonnini

Abstract The utility of combined permutation tests is shown through the application to multivariate sample data concerning the performance of players (defenders) of the Italian football championship 2021/2022. The goal is to test the Home-Field effect on the performance of the players, by comparing the data from the home matches with the data from the away matches.
Abstract *L'utilità dei test di permutazione combinati è illustrata tramite l'applicazione a dati campionari multivariati riguardanti la prestazione di calciatori (difensori) del campionato italiano di calcio 2021/2022. L'obiettivo è di verificare se il fattore campo influisca sulla prestazione dei calciatori, confrontando i dati delle partite in casa con quelli delle partite fuori casa.*

Key words: Performance Assessment, Home-Field Advantage, Permutation Test, Soccer

1 Introduction

Rating and ranking methods in sport have been the subject of several scientific publications [21]. A considerable literature dedicated to the sport of football (or soccer) is also available. Several works focus on team performance evaluation. Some of them apply a dynamic perspective, in order to carry out a team performance assessment over time [5, 14, 13, 15]. In such a framework, a suitable rating approach should jointly consider the performance of the team and the strength of the opponents [9, 10].

A certain stream of literature is dedicated to studying team performance as a function of the interactions (passes) between players [29, 6, 27, 18, 17]. In these

Stefano Bonnini
University of Ferrara, Via Voltapaletto, 11 - 35020 Ferrara (FE), Italy, e-mail: stefano.bonnini@unife.it

studies, to investigate the effectiveness of different playing styles, the tactics adopted by the teams are represented by networks with the players as nodes and the passes as edges. If the goal is to group tactics, in order to guarantee comparability, the playing styles are represented by networks whose nodes are the areas of the pitch and edges are the ball trajectories with respect to such areas [7]. To deepen the tactical analysis of the matches, an interesting approach is based on player tracking [28, 30, 22]. Positional data consisting of spatial coordinates (x,y) are used to represent the movements of the players on the pitch, to study the effectiveness of certain tactics adopted by the teams.

As mentioned above, other authors are rather dedicated to evaluating the performance of individual players, investigating factors and conditions which affect the performance of the athletes, from the technical, tactical, physiological and environmental points of view. [16] analyze the key player's performance indicators and highlights the importance of the distinction by role. In fact, each role has specific technical requirements [33] that are different depending on whether the team is in possession of the ball or not [32]. In order to define winning team strategies, tactical and technical decisions should be based on the information concerning such individual performance indicators [11].

The complexity of individual performance assessment is mainly due to the multidimensionality of performance. Hence, multivariate statistical techniques should be adopted to simplify data processing, extraction of useful information and interpretation of results. The construction of composite indicators of players' performance, based on the information provided by the Kaggle european soccer database, was studied by [2]. To reduce the dimensionality of the multivariate performance variables, [3] proposed a method based on *Clustering around Latent Variables*, which was compared to the *Principal Covariates Regression* and to the *Bayesian model-based clustering* by [4].

The focus of this paper is on the individual skills of football players rather than on tactics or team strategies. We propose the application of combined permutation tests [23, 24] to investigate the effect of factors or conditions on the individual players' performance, in the presence of multivariate sample data. In particular, we consider the two-sample test for multivariate paired data to test the home-field effect in the Italian national football championship. The existence of a home-field advantage is an opinion fairly agreed upon by experts. Usually, it is measured as the percentage of points obtained or goals scored at home and might depend on the tiredness for the travel (the effect increases with the distance), the familiarity with the local conditions, the encouragement of supporters (psychological effect) and the referee conditioning [25, 26]. This is considered one of the main factors affecting the outcome of a match [8]. It was studied in baseball [20], basketball [12], volleyball [31], and other sports, including football. A recent interesting work based on the application of a regression analysis to football data at the team level is [19]. To the best of our knowledge, there are no studies proposing multivariate inferential procedures to test the effect of the home-field factor on individual football players. The remaining part of the paper is structured as follows: in Section 2, the method of combined permutation tests is presented; Section 3 is dedicated to the empirical analysis of the data concerning the Italian Football Championship; in Section 4 the main conclusions and indications for possible future directions of the research are reported.

2 Combined permutation tests

The idea of combining a set of permutation tests to solve multivariate or other complex testing problems dates back to 2001, with the book of Pesarin [23]. A systematic presentation of the main properties and practical rules for practitioners is provided by [24]. Several applications and software, in particular R scripts and packages, for the application of the method are also presented in [1]. This approach to multivariate tests is distribution free and robust with respect to the underlying distribution. The condition for its applicability is the exchangeability of data between samples. No family of distributions has to be assumed and, in particular, no dependence structure between the components of the multivariate response variable must be modeled. This assures flexibility and does not affect the good power performance of the test under different conditions, because the dependence between the univariate marginal responses is implicitly taken into account by the method. Let us consider the following testing problem for two dependent samples and a q-variate numeric response.

Let x_{ujv} denote the value of the vth component of the multivariate response variable observed on the uth unit in the jth sample, with $u = 1,\ldots,n,\ j = 1,2$ and $v = 1,\ldots,q$. For instance, it could represent the value of the vth performance indicator related to the uth individual (in a sample of n individuals randomly selected from the population of Serie A players) concerning the home matches ($j = 1$) or the away matches ($j = 2$). Let assume that x_{ujv} is a determination of the random variable X_{ujv}, whose probability distribution is completely unknown, such that

$$X_{ujv} = \mu_{jv} + \delta_{jv} + \varepsilon_{ujv}, \tag{1}$$

where μ_{jv} is the expectation of X_{ujv}, δ_{jv} is a non-negative constant and ε_{ujv} a random variable with null mean and variance σ_v^2. Also, assume that $\delta_{1v} = 0$. We want to test the null hypothesis that $\delta_{2v} = 0$ for all v against the alternative hypothesis that $\delta_{2v} > 0$ for at least one v. Formally:

$$H_0 : [\forall v\ \delta_{2v} = 0] \equiv \bigcap_{v=1}^{q} [\delta_{2v} = 0], \tag{2}$$

against

$$H_1 : [\exists v \mid \delta_{2v} > 0] \equiv \bigcup_{v=1}^{q} [\delta_{2v} > 0]. \tag{3}$$

Under the null hypothesis, the observed values x_{u1v} and x_{u2v} come from the same population and the probability to observe x_{u1v} in sample 1 and x_{u2v} in sample 2 is equal to the probability to observe x_{u1v} in sample 2 and x_{u2v} in sample 1. Hence, exchangeability holds, and the null permutation distribution of the test statistic can be obtained by permuting B times some of the n couples (x_{u1v}, x_{u2v}). It is worth noting that the random permutations of different units are independent and, to take into account the dependence between the q response variables, for each unit, a permutation of the two values for one variable implies the permutation for all the other variables.

A suitable test statistic for the vth variable (vth partial test) is

$$T_v = \bar{X}_{\cdot 1v} - \bar{X}_{\cdot 2v}. \tag{4}$$

For each permutation, the corresponding value of all the test statistics is computed and the p-values can be obtained according to the consequent permutation distribution. If $T^*_{v,b}$ represents the value of the test statistic for the vth partial test corresponding to the bth permutation, then the corresponding p-value is $p_{v,b} = [\sum_{i=1}^{B} I_{[T^*_{v,b},\infty]}(T^*_{v,i}) + 0.5]/(B+1)$, where $I_A(x)$ is the indicator function of the set A. Finally, to obtain one final p-value for the overall testing problem represented by (2) against (3), a suitable combination of the partial p-values is required. For instance, the value of the combined test statistic corresponding to the bth permutation could be:

$$T_{comb,b} = max(1 - p_{1,b}, 1 - p_{2,b}, \ldots, 1 - p_{q,b}). \tag{5}$$

The final p-value of the overall test $p_{comb,b}$ can be obtained similarly to $p_{v,b}$.

3 Application

The method presented in Section 2 was applied to test the hypothesis that the performance of football players with the role of defender in the national championship of Serie A league, in the season 2021/2022, was significantly better in the home matches than in the away matches (Home-Field effect). A random sample of 10 defenders who played at least 20 matches in that season was randomly selected and 17 response variables, representing the players' performance, were considered. Even if the main skills required to a defender concern the defensive phase, nowadays all the players contribute to both the phases: defensive and offensive. Hence the 17 performance indicators considered in the study include both defensive and offensive performance indicators. The variables concerning the defensive performance are: tackles, interceptions, fouls committed, caused offsides, clearances, suffered dribbles, saves. The variables concerning the offensive phase are: scored goals, assists, shots on target, key passes, successful dribbles, suffered fouls, offsides, lost possessions, total passes and pass success ratio.

The data were downloaded from the website *whoscored.com*, one of the most complete and with detailed information at the player level, with the distinction between home and away performance. Given that all the partial test statistics should reject the null hypothesis for large values, for the variables fouls committed, suffered dribbles, offsides and lost possession the opposite value was considered, consistently to the rule "the larger the better". For the application of the presented methodology an original R script, specifically created by the author, was used for data processing. The output provided by the software reported a combined p-value of 0.143 indicating no significance.

4 Final Remarks and Perspectives

Combined permutation tests are suitable for evaluating the Home-Field effect on the performance of football players, mainly for the flexibility and robustness with respect to the assumptions about the underlying distribution. Furthermore, it is applicable with a large number q of variables and very small sample sizes (even less than q). According to the result of the application of a combined permutation test to the data regarding the Serie A league, we have no empirical evidence in favor of the hypothesis of the home-field effect on defenders.

Future directions of the research concern the extension of the analysis to the other roles and the classification of variables into defensive and offensive performance, in order to attribute the possible significance of the overall test to specific roles and variables' domains. This choice is appropriate also considering the non-significance of the test in Section 3, which is quite surprising.

References

1. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric Hypothesis Testing. Rank and Permutation Methods with Applications in R. Wiley, Chichester (2014)
2. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modeling team performances of the kaggle european soccer database. Stat. Model. 19(1), 74–101 (2019)
3. Carpita, M., Ciavolino, E., Pasca, P.: Players' Role Based Performance Composite Indicators of Soccer Teams: A Statistical Perspective. Soc. Indic. Res. 156, 815–830 (2021)
4. Carpita, M., Pasca, P., Arima, S., Ciavolino, E.: Clustering of variables methods and measurement models for soccer players' performances. Ann. Oper. Res. (2023) doi: 10.1007/s10479-023-05185-w
5. Cattelan, M., Varin, C., Firth, D.: Dynamic Bradley-Terry Modelling of Sports Tournaments. J. R. Stat. Soc. C-Appl. Stat. 62, 135–150 (2013)
6. Clemente, F.M., Conceiro, M.S., Martins, F.M.L., Mendes, R.S.: Using network metrics in soccer: A macro-analysis. J. Hum. Kinet. 45, 123–134 (2015)
7. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. Stat. Model. 19(1), 28–54 (2019)
8. Edwards, J.D., Archambault, D.: The Home-Field Advantage. In: Goldstain, J.H. (eds.) Sports, Games, and Play: Social and Psychological viewpoints, pp. 409-438. Psychology Press, New York (1979)
9. Elo, A.E.: The Rating of Chess Players, Past and Present. Arco, New York, NY (1978)
10. Franceschet, M., Bozzo, E., Vidoni, P.: The temporalized Massey's method. J. Quant. Anal. Sports 13(2), 37–48 (2017)

11. Franks, I.M., Goodman, D.: A systematic approach to analysing sports performance. J. Sport Sci. 4, 49–59 (1986)
12. Gayton, W.F., Coombs, R.: The home advantage in high school basketball. Percept. Motor Skill. 81(3), 1344–1346 (1995)
13. Glickman, M.E., Stern, H.S.: A state-Space Model for National Football League Scores. J. Am. Stat. Assoc. 93, 25–35 (1998)
14. Glickman, M.E., Stern, H.S.: Estimating the team strength in the NFL. In: Albert, J., Glickman, M.E., Swarts, T.B., Koning, R.H. (eds.) Handbook of Statistical Methods and Analysis in Sports, pp. 113-136. CRC Press, Boca Raton, FL (2017)
15. Harville, D.: Predictions for National Football League Games via Linear-Model Methodology. J. Am. Stat. Assoc. 75, 516–524 (1980)
16. Hughes, M., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C.: Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position. J. Hum. Sport Exerc. 7(2), 402–412 (2012)
17. Ievoli, R., Gardini, A., Palazzo, L.: The role of passing network indicators in modeling football outcomes: an application using Bayesian hierarchical models. AStA Adv. Stat. Anal. 107(1), 153–175 (2023)
18. Ievoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. Knowledge-Based Systems. 222 (2021)
19. Inan, T.: The effect of Crowd Support on Home-Field Advantage: Evidence from European Football. Ann. Appl. Sport Sci. 8(3), e806 (2020)
20. Irving, P.G., Goldstein, S.R.: Effect of home-field advantage on peak performance of football pitchers. J. Sport Behavior 13(1), 1–23 (1990)
21. Langville, A.N., Meyer, C.D.: Who's #1? The Science of Rating and Ranking. Princeton University Press, Princeton, NJ (2012)
22. Memmert, D., Reabe, D.: Data Analytics in Football: Positional Data Collection, Modeling and Analysis. Routledge (2018)
23. Pesarin, F.: Multivariate Permutation Tests with Appplications in Biostatistics. Wiley, Chichester (2001)
24. Pesarin, F., Salmaso, L.: Permutation tests for complex data. Theory, Applications and Software. Wiley, Chichester (2010)
25. Pollard, R.: Home advantage in soccer: a retrospective analysis. J. Sport Sci. 4(3),237–248 (1986)
26. Pollard, R., Gomez, M.A.: Home advantage in football in South-West Europe: Long-term trends, regional variation, and team differences. European J. Sport Sci. 9(6),341–352 (2009)
27. Pina, T.J., Paulo, A., Araujo, D.: Network characteristics of successful performance in association football. A study on the UEFA Champions League. Front. Psychol. 8, 1173 (2017)
28. Reilly, T., Thomas, V: A motion analysis of work-rate in different positional roles in professional football match play. J. Hum. Movement Stud. 2, 87–97 (1976)
29. Reep, C., Benjamin, B.: Skill and Chance in Association Football. J. R. Stat. Soc. 131(4), 581–585 (1968)
30. Yamanaka, K., Liang, D.Y., Hughes, M.: An analysis of the playing patterns of the Japan national team in the 1994 World Cup qualifying match for Asia. In: Reilly, T., Bangsbo, J., Hyghes, M. (eds.) Science and Football III, pp. 221-228. E&FN Spon, London (1997)
31. Yu, Y., Garcia De Alcaraz, A., Cui, K., Liu, T.: Interactive effects of home advantage and quality of opponent in Chinese Women's Volleyball Association League. Int. J. Perf. Anal. Spor. 20(1), 107–117 (2020)
32. Van Lingen, B.: Coaching Soccer. Reedswain, Spring City (1997)
33. Wiemeyer, J.: Who should play in which position in soccer? Empirical evidence and unconventional modelling. Int. J. Perf. Anal. Spor. 3(1), 1–18 (2003)

A First Proposal of the Triad Census for Weighted Networks: an Application to Football

Una proposta di conteggio delle triadi per reti pesate: una applicazione a dati calcistici

Roberto Rondinelli and Lucio Palazzo

Abstract Triad census is a useful tool in various types of applications involving relational data. The conventional method is well-defined for unweighted networks, while it loses information about the intensity of the phenomenon of interest when applied to weighted and dense networks. To overcome this issue, this work proposes an algorithm denoted as “network peeling” to count the different configurations of triads in weighted networks. To show the applicability of the proposed method, we consider a real dataset involving the passing networks of 32 teams in 96 matches in the Group Stage of the UEFA Champions League 2016-2017.
Abstract *Il conteggio delle triadi è uno strumento utile in diverse applicazioni che coinvolgono dati relazionali. Il metodo convenzionale è ben definito per le reti non pesate, ma in caso di reti pesate e dense porta ad una perdita di informazione circa l'intensità del fenomeno di interesse. Per superare questo problema, il presente lavoro propone un algoritmo chiamato “network peeling” per contare le diverse configurazioni di triadi nelle reti pesate. Per dimostrare l'applicabilità del metodo proposto, si considera un dataset comprendente le reti dei passaggi di 32 squadre in 96 partite nella fase a gironi della UEFA Champions League 2016-2017.*

Key words: Weighted Networks, Triad Census, Peeling, Passing Network

1 Introduction

Considering binary networks, an interaction between three nodes is called “triad”, i.e., the minimal group that can be observed in a complex relational structure. According to the presence and type of the relations between three nodes, different triadic configurations (referred to as isomorphism classes) can be defined [11]. In the Network Analysis (NA) theory the triad census has assumed an increasing rel-

Roberto Rondinelli
University of Naples Federico II, e-mail: roberto.rondinelli@unina.it

Lucio Palazzo
University of Naples Federico II, e-mail: lucio.palazzo@unina.it

evance in the extraction of hidden differences between networks, which sometimes are not highlighted by other descriptive measures, especially in the case of small and dense networks. In the context of weighted networks, the increasing complexity of network structures implies an increasing complexity in the analysis of triads, leading to multiple solutions which are based only on the principle of isomorphism classes. For instance, [1] evaluates the local cohesiveness that takes into account the importance of the interaction intensity found on the local triads, while [9] extends this approach associating measures of "intensity" and "coherence", focusing on the information carried out by the weights. These indices extend the measurement of the local clustering coefficient [12] to weighted networks, see e.g., [3] for the case of directed networks. Moreover, to make use of this method on weighted networks, one needs to dichotomize the network relations by defining a threshold or discarding the weights. As a consequence, useful information about the relations is lost.

Given these premises, the aim of the present paper is to propose a new procedure, called "network peeling", to analyze weighted networks by considering not only the presence of a particular triadic configuration but also the intensity of the relationships. To the best of our knowledge, no methodology has yet been developed to count the triadic configurations (i.e., classes of isomorphism in both the undirected and directed cases) in weighted networks.

To show the applicability of the proposed method, we analyze the case study of the passing networks occurring in football matches. For each football team involved in a match, the propounded algorithm allows the construction of a matrix summarizing the relevant information arising from both the isomorphism classes and the weights associated with each passing intensity level.

Different contributions explore the potential of NA to visualize and analyse football passes. Some contributions [5, 7, 8] show a positive relationship between some network indices and the football outcome. Furthermore, [10] characterizes team strategies by testing the style of play according to a set of reference strategies determined by specific triadic configurations, following the approach of [4].

The paper is organized as follows: Section 2 illustrates the NA framework and the proposed extension of the conventional triad census, i.e, the "peeling" of the network. Section 3 includes the application of the proposed method considering 192 passing networks coming from the Group Stage of the UEFA Champions League 2016-2017. Remarks and possible advances are depicted in Section 4.

2 Extending Triad Census

A weighted and directed network $\mathscr{G}$ consists of a triple $\mathscr{G}(\mathscr{V},\mathscr{E},\mathscr{W})$, where $\mathscr{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of the N nodes; $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of directed edges, each edge $e_{ij} \neq e_{ji}$ is an ordered couple indicating a relation from v_i to v_j; $\mathscr{W}$: $\mathscr{E} \to \mathbb{N}$, with $\mathscr{W}(e_{ij}) = w_{ij} \in \mathbb{N}$, is the number of times the relation from v_i to v_j is observed. In the specific case of our application, the passing network is defined for each team playing a football match as above. In particular, the generic directed edge

e_{ij} exists when (at least) a pass from player v_i to player v_j occurs, while w_{ij} is the count of their total passes.

The minimal structural group in a network involves a subset composed of three nodes. Considering directed networks, the number of these relational structures, i.e., the *isomorphism classes*, is equal to 16. Figure 1 shows these isomorphism classes and their labels, from the so-called "003", presenting no edges between nodes, to the fully connected triad "300", involving six links. Triad census analysis has been widely presented and developed in the literature [11, 4] also presenting a probabilistic structure that could be derived under certain assumptions [6]. In general, with a certain degree of degeneration, different network models produce different distributions of the triadic configurations.

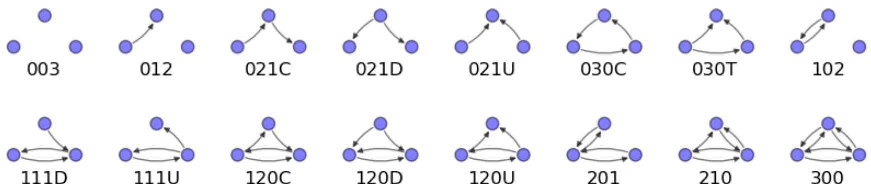


Fig. 1 List of isomorphism classes of the triad census.

As mentioned in Section 1, this is not completely clear when considering weighted networks. To overcome this issue, we propose an algorithm to count the triads based on their isomorphic classes and take into account the information deriving from the weights (w_{ij}).

2.1 Network peeling

The peeling algorithm consists of a nested sequence of (sub)networks in which their edges are "peeled out" each time of a unit value. For each of them, the usual (binary) triad census is computed. Subsequently, the final matrix of all the triad census scores is aggregated to obtain the weighted scores. Let $\mathscr{G}$ be the initial network, the procedure generates a set of weighted (sub)networks $\mathscr{G}_{(t)} \subseteq \mathscr{G}$ at each step t, in which the updated weights $w_{ij,(t)}$ are equal to

$$w_{ij,(t)} = \begin{cases} w_{ij,(t-1)} - 1 & \text{if } w_{ij,(t-1)} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The $\mathscr{G}_{(t)}$ is binarized into $\mathscr{G}^B_{(t)}$ by defining the corresponding edges $e^B_{ij,(t)}$ as

$$e^B_{ij,(t)} = \begin{cases} 1 & \text{if } w_{ij,(t)} > 0 \\ 0 & \text{if } w_{ij,(t)} = 0. \end{cases} \tag{2}$$

Then, the triad census is computed on $\mathscr{G}^B_{(t)}$. Subsequently, the new (sub)network $\mathscr{G}_{(t+1)}$ is updated by removing any isolated nodes from $\mathscr{G}_{(t)}$. The procedure will stop when the resulting (sub)network consists of less than three nodes.

Let $TC_{(t)} = (TC_{1,(t)}, \dots, TC_{16,(t)})$ be the empirical distribution of the triad census at step t, for all the 16 isomorphism classes, then the weighted triad census $wTC = (wTC_1, \dots, wTC_{16})$ is computed as

$$wTC_i = \sum_t TC_{i,(t)}, \qquad \forall i = 1, \dots, 16. \tag{3}$$

In Algorithm 1 the pseudocode of the proposed procedure is summarized.

Algorithm 1 Peeling Algorithm

procedure PEELING(G)
  G ▷ Initial graph
  $G^B \leftarrow binarize(G)$ ▷ Binarized graph
  $triads \leftarrow triad\_census(G^B)$ ▷ TC of the original graph
Loop:
  while $length(G.vertices) \geq 3$ **do**
    $G.weight \leftarrow G.weight - 1$ ▷ update edge weights
    $G.edges \leftarrow G.edges[G.weight > 0]$ ▷ delete null edges
    $G^B \leftarrow binarize(G)$ ▷ Binarized graph
    $triads \leftarrow \{triads, triad\_census(G^B)\}$ ▷ TC of the subgraph
    $G.vertices \leftarrow G.vertices[degree(G.vertices) > 0]$ ▷ delete isolated nodes
  end while
  $wTC \leftarrow rowSums(triads)$.
end procedure

3 Empirical Application

We retrieve the team passing distributions from the Group stage of the UEFA Champions League (UCL) for the Season 2016-2017. The source is the official UEFA website (www.uefa.com) and the data include 192 passing networks, grouped in 6 networks per team (32), each of them referring to a match (for a total of 96 matches).

We apply the peeling over the considered networks, obtaining a matrix M of the weighted triad census scores for each team involved in a match. In Figure 2 the main results of the Correspondence Analysis (CA) applied on M are depicted. Specifically, at the team level, we plot the team centroids computed on the (weighted) triadic results considering 6 matches. Then, the k-means algorithm (with $k = 3$) is applied to the first two components of the CA ("tandem" approach). The two selected dimensions can explain 68% of the total variability (top panel of Figure 2). The first dimension (50%) denotes the level of reciprocity of passes between players (higher in the left part), while the second dimension (17.7%) expresses the level

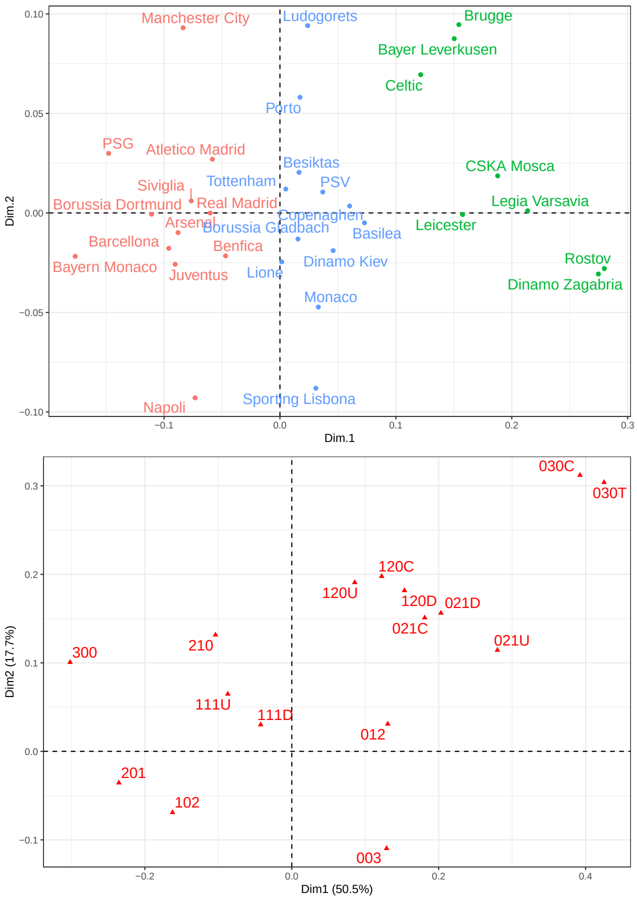


Fig. 2 Output of CA by row and column. The top panel depicts the rows considering the centroids of the matches for each football team. Colours describe the groups found by using k-means. The bottom panel depicts the columns, i.e., isomorphism classes of the triads.

of interaction between players in terms of their capability to close the triads in a more cyclical and transitive perspective. The importance of the first dimension is confirmed by the clustering division (bottom panel of Figure 2). The best partition organizes teams in three clusters arranged horizontally on the axes. From the left to the right, we can notice a clear division between top-tier teams (red group) and the low-tier ones (green group), whereas the majority of teams belonging to the blue group (at the centre of the plot) were relegated in the Europe League competition.

Relevant differences arise comparing these results with those obtained via CA and k-means on the conventional triad census (here not presented for the sake of brevity). Indeed, a low value of the Adjusted Rand Index (0.17) suggests that the new approach produces a diverse kind of information.

4 Conclusions

Triad census is a useful tool of NA to describe the structure of a binary network taking into account triplets of nodes. However, the weighted and dense nature of some real networks makes it impossible to count the isomorphism classes. Indeed in dense networks, the differences between the weights of the edges can be relevant to define the underlying structure. The peeling of the network is a possible solution to deal with such cases. The algorithm peels out a weighted network in many binary layers allowing the count of the isomorphism classes to measure the weighted triad census. As shown through the case study, it proves particularly helpful for small networks, even if it can be easily extended to medium and large ones. Indeed, the passing networks of the 32 teams of UEFA Champions League 2016-2017, are grouped meaningfully considering the different styles of play.

For what concerns further advances, an extensive simulation study should be conducted to properly compare conventional and weighted triad census, while probabilistic properties of weighted triads may be investigated under certain assumptions.

References

1. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. Proceedings of the national academy of sciences 101(11): 3747–3752 (2004)
2. Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. Stat. Modelling, 19(1), 28–54 (2019)
3. Fagiolo, G.: Clustering in complex directed networks. Phys. Rev. E, 76(2), 026107 (2007)
4. Faust, K.: A puzzle concerning triads in social networks: Graph constraints and the triad census. Soc. Networks, 32(3), 221–233 (2010)
5. Grund, T. U.: Network structure and team performance: The case of English Premier League soccer teams. Soc. Networks, 34(4), 682–690 (2012)
6. Holland, P. W., Leinhardt, S.: An omnibus test for social structure using triads. Sociol. Methods Res., 7(2), 227–256 (1978)
7. Ievoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. Knowl. Based Syst., 222, 106997 (2021)
8. Ievoli, R., Gardini, A., Palazzo, L.: The role of passing network indicators in modeling football outcomes: an application using Bayesian hierarchical models. AStA Adv. Stat. Anal., 107(1-2), 53–171 (2023)
9. Onnela, J. P., Saramaki, J., Kertesz, J., Kaski, K. Intensity and coherence of motifs in weighted complex networks. Phys. Rev. E, 71(6), 065103. (2005)
10. Palazzo, L., Ievoli, R., Ragozini, G.: Testing styles of play using triad census distribution: an application to men's football. *Working Paper*
11. Wasserman, S., Faust, K. Social network analysis: Methods and applications. Cambridge University Press (1994)
12. Watts, D. J., Strogatz, S. H.: Collective dynamics of 'small-world'networks. Nature, 393(6684), 440–442. (1998)

Solicited Session SS5 - *Advances in statistical learning from high-dimensional data*

Organizer and Chair: Fabrizio Maturo

1. *PCA approaches for vector functional time series* (Aguilera A.M., Alonso F.J. and Acal C.)
2. *Conformal Prediction for Functional Kriging Models* (Diana A., Romano E. and Adzic J.)
3. *Measuring Public-Private Connectedness in Financial Markets* (Sánchez García J. and Cruz Rambaud S.)
4. *An original approach to anomalies in intertemporal choices through Functional Data Analysis: Theory and application for the study of Hikikomori syndrome* (Martino R., Ventre V., Cruz Rambaud S. and Maturo F.)

PCA approaches for vector functional time series

Approcci PCA per serie storiche funzionali vettoriali

Ana M. Aguilera, Francisco J. Alonso and Christian Acal

Abstract Functional time series are characterized because the observations are continuous functions defined in some closed interval. These stochastic processes are very common in different areas of knowledge such as medicine or economy, among others. The current work is motivated by modelling and predicting the bivariate curves associated with the resistive switching processes generated inside the non-volatile memories operation. To solve this problem, a new approach for the study of vector functional time series is introduced within the context of functional data analysis. In particular, we propose to turn the functional autoregressive model into a vector autoregressive model by replacing the functional variables by their principal component representation.

Abstract *Le serie temporali funzionali sono caratterizzate perché le osservazioni sono funzioni continue definite in un intervallo chiuso. Questi processi stocastici sono molto comuni in diverse aree della conoscenza come la medicina o l'economia, tra le altre. Il lavoro attuale è motivato dalla modellazione e dalla previsione delle curve bivariate associate ai processi di commutazione resistiva generati all'interno del funzionamento delle memorie non volatili. Per risolvere questo problema, viene introdotto un nuovo approccio per lo studio delle serie temporali funzionali vettoriali nell'ambito dell'analisi dei dati funzionali. In particolare, proponiamo di trasformare il modello autoregressivo funzionale in un modello autoregressivo vettoriale sostituendo le variabili funzionali con la loro rappresentazione della componente principale.*

Key words: functional time series, autoregressive models, principal components

---

Ana M. Aguilera
Dept. of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: aaguiler@ugr.es

Francisco J. Alonso
Dept. of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: falonso@ugr.es

Christian Acal
Dept. of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: chracal@ugr.es

1 Introduction

Time series are composed by ordered observations of a magnitude sequentially obtained over the time. The inherent stochastic nature establishes a clear sequential dependence among observations that must be modelled. Besides, the magnitude to be predicted is random so that the uncertainty is controlled by different probabilistic models. A particular case of time series is the so called functional time series (FTS) in which the observations are functions instead of scalar vectors that evolve over some continuous argument. FTS can be also generated from a continuous-time stochastic process divided in segments of the same length, thus getting a sequence of functional data. However, the classical methodology based on autoregressive models proposed by Box-Jenkins (see, e.g. [1, 2]) is not adequate to model FTS because the observations are not vectors anymore. Then, some approach from Functional Data Analysis (FDA) must be considered to guarantee a suitable modelling.

FDA comprehends a wide variety of statistical tools to model and predict functional data (normally curves). General methodologies and aspects related to inference are discussed in [3, 4, 5]. Besides, and taking into account the high dimensionality in these studies (great number of correlated variables over small samples), Functional Principal Component Analysis (FPCA) plays a fundamental role in the FDA framework. In last two decades, FPCA has been subject of intensive research both for the univariate and multivariate case and has been applied to different areas from a theoretical and practical viewpoint (see, e.g. [6, 7, 8] and the references therein).

Depending on the final purpose, there are multiple works in the literature that address FTS through different functional methodologies. For instance, [9] developed a new clustering technique based on functional panel data modelling to capture homogeneity among FTS. On the other hand, [10] introduced a new functional approach to predict the unobserved part of those curves which have been partially observed using all the past trajectories. Likewise, different weighted regression models are considered to forecast functional time series in [11]. Finally, an exhaustive revision of functional autoregressive models can be checked in [12].

This paper is motivated by analyzing the bivariate curves generated by the reset and set processes related to non-volatile memories. In broad terms, these curves are generated consecutively cycle-to-cycle inside the process of formation (set) and rupture (reset) of a filament conductor (process that causes the resistive switching operation of these memories). Despite the stochastic nature, the usual statistical analysis carried out on this sector makes use of scalar time series tools on the voltages where the filament has been formed or destroyed.

In order to avoid inefficient and inadequate results, we consider a functional autoregressive model for vector functional time series. This model is subsequently turned into a vector autoregressive model by replacing the functional variables by their principal component representation. Thus, the resultant multivariate autoregressive model can be analysed by means of classical techniques on the vector of the most explicative principal component scores.

2 Analysis of vector functional time series

A functional time series is a collection of functions denoted as $\{X_i(t) : i \in \mathbb{N}; t \in T = [a,b]\}$ where i is the index of the series and t is in some continuous domain T. This definition can be generalised for a vector of more than one functional variable so that $\mathbf{X}_i(t) = (X_{i1}(t), X_{i2}(t), \dots, X_{iH}(t))^T$. Let us suppose that these trajectories are generated from a H-dimensional stochastic process with second order and continuous in quadratic mean, whose sample curves belong to the Hilbert space $L^2[T]$ of squared integrable functions.

The analysis focused on these data tries to study the sequential dependence among the trajectories, to model the evolution over time and to make predictions with enough precision. In this context, the functional autoregressive model (FAR) under Hilbert space is likely the most popular technique for this end. Next, the univariate functional autoregressive model of order p (FAR(p)) is adapted for the multivariate case as follows (MFAR(p)):

$$\begin{aligned}\mathbf{X}_{ij}(t) - \mu_j(t) &= \sum_{k=1}^{p} \rho_{jk}(\mathbf{X}_{i-k,j}(t) - \mu_j(t)) + \varepsilon_i(t) \\ &= \sum_{k=1}^{p} \int_T \phi_{jk}(t-s)[\mathbf{X}_{i-k,j}(s) - \mu_j(s)]ds + \varepsilon_i(t),\end{aligned} \tag{1}$$

where $\mu(t)$ is the mean function vector of $\mathbf{X}_i(t)$, $\phi_{jk}(t) \in L^2(T)$ is the kernel function of the Hilbert-Schmidt operator ρ_{jk}, $\mathbf{X}_{i-k}$ represents the kth lag of $\mathbf{X}_i$ and $\varepsilon_i(t)$ denotes white noise with zero mean and finite second moments.

The above model assume stationarity which means the dynamics of each series is stable over time. A revision of models under nonstationarity hypothesis can be seen in [12].

2.1 Multivariate Functional Principal Component Analysis

Multivariate Functional Principal Component Analysis (MFPCA) is the natural extension of the univariate dimension reduction technique called FPCA when there are more than one functional variable in the analysis. It is well known that FPCA aims to explain the main modes of variation of a functional dataset in terms of a reduced numbers of uncorrelated variables denominated principal components.

Let $\mathbf{X}_1(t), \dots, \mathbf{X}_n(t)$ be a multivariate sample of trajectories where $\mathbf{X}_i(t) = (X_{i1}(t), ..., X_{iH}(t))^T$ contains the information for the ith subject in each functional variable. Besides, let us assume that this sample of curves belong to a stochastic process with sample mean vector functions $\mu = (\mu_1(t), ..., \mu_H(t))^T$ and sample matrix covariance function $\mathbf{C}$ such that $\mathbf{C}(t,s) = (C_{h,h'}(t,s))$, $t,s \in T = [a,b]$

and $h, h' = 1, ..., H$. $C_{h,h'}$ will be covariance function if $h = h'$ and cross-covariance function otherwise.

The jth principal component is computed by means of the following expression

$$\xi_{ij} = \int_T (\mathbf{X}_i(t) - \mu(t))^T \mathbf{f}_j(t)dt = \sum_{h=1}^{H} \int_T (X_{ih}(t) - \mu_h(t)) f_{jh}(t)dt, \qquad (2)$$

where $\mathbf{f}_j(t) = (f_{j1}(t), ..., f_{jH}(t))^T$ are the eigenfunction vectors obtained as the solutions to the eigenequation system $\mathscr{C}\mathbf{f}_j = \lambda_j \mathbf{f}_j$. Here $\mathscr{C}$ is the covariance operator and $\{\lambda_j\}_{j\geq 1}$ is the decreasing sequence of non-null eigenvalues such that $\lambda_j = \mathrm{VAR}[\xi_j]$.

Finally, the sample curves can be approximated by truncating the Karhunen-Loève expansion (an orthogonal decomposition admitted by the stochastic process that generates the curves) in terms of the first q principal components as follows:

$$\mathbf{X}_i^q(t) = \mu(t) + \sum_{j=1}^{q} \xi_{ij} \mathbf{f}_j(t).$$

Note that in practice q must be chosen so that the explained cumulative variability is around 95% of the total variability in order to guarantee a good reconstruction of the sample curves.

3 A new autoregressive model for functional time series

Reducing the infinite dimension of the vector functional time series by using MF-PCA and replacing the functional variables by the representation in principal components (2), the functional autoregressive model expressed in (1) is equivalent to the classical autoregressive model for the vector with the most explicative principal component scores. Then, the new autoregressive model of order p based on principal components is

$$\boldsymbol{\xi}_i - \boldsymbol{\mu}_{\boldsymbol{\xi}} = \sum_{k=1}^{p} \Omega_k \boldsymbol{\xi}_{i-k} + \varepsilon_i^*,$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iq})^T$, ε_i^* is white noise and Ω_k is the matrix of coefficients (weights) of order $q \times q$.

Then, the initial problem would be reduced to estimate this vector autoregressive model for the most explicative principal components.

Acknowledgements This Research has been supported by project FQM-307 of the Government of Andalusia (Spain) and project A-FQM-66-UGR20 of the FEDER operative program of Government of Andalusia (Spain). Additionally, the authors acknowledge financial support by the IMAG María de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033 and the grant PID2021-128077NB-I00 financed by MCIN/AEI/10.13039/501100011033/FEDER.

References

1. Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1976)
2. Pankratz, A.: Forecasting with Univariate Box-Jenkins Models. Wiley (1983)
3. Ramsay, J.O., Silverman, B.W.: Functional data analysis (Second Edition). Springer-Verlag (2005)
4. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and practice. Springer-Verlag (2006)
5. Horvath, L., Kokoszka, P.: Inference for functional data with applications. Springer-Verlag (2012)
6. Jacques, J., Preda. C.: Model-based clustering for multivariate functional data. Comput. Stat. Data. Anal. 71, 92–106 (2014)
7. Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., Martin, P.: Clustering multivariate functional data in group-specific functional subspaces. Comput. Stat. 35, 1101–1131 (2020)
8. Acal, C., Aguilera, A.M., Sarra, A., Evangelista, A., Di-Battista, T., Palermi, S.: Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. Stoch. Env. Res. Risk A. 36, 1083-1101 (2022)
9. Tang, C., Shang, H.L., Yang, Y.: Clustering and forecasting multiple functional time series. Ann. Appl. Stat. 16(4), 2523-2553 (2022)
10. Jiao, S., Aue, A., Ombao, H.: Functional Time Series Prediction Under Partial Observation of the Future Curve. J. Am. Stat. Assoc. 118, 315-326 (2023)
11. Hyndman, R.B., Shang, H.L.: Forecasting functional time series. J. Korean Stat. Soc. 38, 199-211 (2009)
12. Chen, Y., Koch, T., Lim, K.G., Xu, X.: A review study of functional aut view study of functional autoregressive models with e models with application to energy forecasting. Wires Comput. Stat. 13(3), e1525 (2021)

Conformal Prediction for Functional Kriging Models

Conformal Prediction per modelli di Kriging funzionale

Andrea Diana, Elvira Romano and Jovanka Adzic

Abstract In this work we introduce a conformal prediction method for functional kriging. Conformal Prediction (CP) is a framework in machine learning and statistical inference that provides a principled way to quantify uncertainty and make predictions without relying on specific distributional assumptions. The approach we introduce provides additional information about the uncertainty associated with the kriging predictions by constructing prediction regions with a specified error rate. We apply CP for functional kriging to the analysis of spatially located network information. This allows us to obtain more reliable estimates and better assess the accuracy of predictions based on functional data.
Abstract *In questo lavoro proponiamo un metodo di Conformal Prediction per il kriging funzionale. La Conformal Prediction (CP) offre un approccio sistematico per quantificare l'incertezza ed effettuare previsioni senza fare affidamento su specifiche ipotesi distributive. L'approccio che introduciamo fornisce informazioni aggiuntive sull'incertezza associata alle previsioni di kriging mediante la costruzione di regioni di previsione con un tasso di errore specificato. In questo modo, possiamo ottenere stime più affidabile valutare meglio la precisione delle previsioni basate sui dati funzionali.*

Key words: Conformal prediction, functional data, kriging , spatial dependence.

Andrea Diana
University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: andrea.diana@unicampania.it

Elvira Romano
University of Campania Luigi Vanvitelli, Caserta, Italy, e-mail: elvira.romano@unicampania.it

Jovanka Adzic
TIM, Torino, Piemonte, Italy, e-mail: jovanka.adzic@telecomitalia.it.

1 Introduction

Functional data with spatial dependence refers to the situation where functional observations are collected at different spatial locations and exhibit correlation or dependence structure in addition to their inherent functional nature [2], [8], [9]. This type of data arises in various fields, such as environmental monitoring, geostatistics, remote sensing, and spatial epidemiology. In the framework of functional prediction, Functional Kriging is a method that combines ideas from geostatistics and functional data analysis to predict functional values at unobserved locations. The estimation is typically performed using a weighted average of the observed functional values, where the weights are determined based on the covariance structure. The weights are computed in a way that gives more importance to the observed values that are spatially closer and have higher functional correlation. The simplest form of kriging is known as ordinary kriging, which enables the prediction of a curve at an unmonitored location assuming a constant mean value. This method has been extensively studied and applied in various research works such as [2], [6], and [11]. In some cases, the mean function may vary with geographical coordinates, introducing spatial non-stationarity. Researchers such as [1], [10], and [12] have investigated the situation where the mean function depends on longitude and latitude, extending ordinary kriging to accommodate spatially varying means. Furthermore, [7] have explored more complex forms of non-stationarity in functional data analysis. They considered situations where the mean function may depend on exogenous variables, whether scalar or functional in nature. To address these scenarios, they developed a method called kriging with external drift, also known as regression kriging, specifically tailored for functional data analysis.

Despite significant advances in spatial prediction methods for functional data, the challenge of quantifying the uncertainty associated with predicted curves remains an ongoing issue. Unlike in traditional kriging, where the kriging variance provides a measure of uncertainty, there is currently no direct functional equivalent of the kriging variance. The absence of a distribution function specifically designed for functional data poses a challenge when it comes to estimating the uncertainty in functional predictions. Consequently, researchers have turned to resampling methods to tackle this issue and compute confidence bands for functional predictions. Resampling methods, such as bootstrap or permutation techniques, are employed to generate multiple resamples from the observed functional data. These resamples are used to create a distribution of possible functional predictions, from which confidence bands can be derived [5]. By resampling the functional data, the inherent uncertainty in the data and the prediction process can be captured, providing a means to quantify the uncertainty associated with the predicted curves.

While resampling methods provide a useful tool for quantifying uncertainty in functional predictions, ongoing research aims to develop alternative approaches that directly address the issue of uncertainty in the functional framework. These efforts involve exploring novel statistical techniques like Conformal Prediction (CP) methods we propose, which would ultimately improve the characterization of uncertainty in functional predictions.

A first proposal of Conformal Prediction for spatio-functional regression model can be found in [3]. The proposed approach is an extension of a CP method for functional regression model [4].

The main aim of this work is to propose a Conformal Prediction method for kriging models on functional data. The method does not rely on any specific distributional assumptions and can be used for a variety of kriging models. In addition it provides a valuable approach for uncertainty quantification in the context of georeferenced functional data. The method works by incorporating the principles of CP to estimate the uncertainty associated with the predictions. It uses a training dataset consisting of observed spatial data to build the kriging model. The kriging model then provides estimates of the spatial process and its uncertainty and is used to compute prediction intervals for new, unobserved locations. These prediction intervals are constructed in such a way that they guarantee a pre-defined level of confidence or validity. This means that the true value is expected to fall within the prediction interval with a certain probability. In the following, we will provide a brief introduction to the Universal Functional Kriging (UFK) model, which represents a more general approach to functional kriging. We then present the Conformal Prediction method that can be applied to enhance the kriging model.

2 Universal kriging for functional data

Geostatistical functional data $(X_{s_1}(t),\dots,X_{s_i}(t),\dots,X_{s_n}(t))$ are random functions $X_s(t)$ located in n points $(s_1,\dots,s_i,\dots,s_n)$ in $D \subseteq R^d$. Each function is defined on $T=[a,b]\subseteq R$ and is assumed to belong to a Hilbert space with the inner product $\langle X_{s_i},X_{s_j}\rangle = \int_T X_{s_i}(t)X_{s_j}(t)dt$ [13]. For a fixed site s_i, it is assumed that the observed functions can be expressed according to the model: $X_{s_i}(t)=\mu_{s_i}(t)+\varepsilon_{s_i}(t), \quad i=1,\dots,n$ where $\mu_{s_i}(t)$ describes the non-constant spatial mean variation and $\varepsilon_{s_i}(t)$ is supposed to be a zero-mean, second-order stationary and isotropic random field, i.e.:

- $\mathbb{E}[X_{s_i}(t)]=\mu_{s_i}(t), \ \forall i=1,\dots,n;$
- $\mathbb{E}[\varepsilon_{s_i}(t)]=0,, \ \forall i=1,\dots,n;$
- $Cov(\varepsilon_{s_i}(t),\varepsilon_{s_j}(t))=\mathbb{E}[\langle\varepsilon_{s_i}(t),\varepsilon_{s_{j(t)}}\rangle]=C(h), \ \forall i,j=1,\dots,n, \ h=||s_i-s_j||.$

Given n observations $\{X_{s_1}(t),\dots,X_{s_i}(t),\dots,X_{s_n}(t)\}$ the formulation of the Universal Kriging predictor of the variable X_{s_0} located in $s_0\in D$ is the best linear unbiased predictor (BLUP), according to [10]:

$$X^*_{s_0}=\sum_{i=1}^{n}\lambda_i^* X_{s_i} \tag{1}$$

whose weights $\lambda_1^*,\dots,\lambda_n^*$ minimize the global variance of the prediction error under the unbiasedness constraint:

$$(\lambda_1^*, \dots, \lambda_n^*) = argmin_{\substack{\lambda_1,\dots,\lambda_n \in \mathbb{R}: \\ X_{s_0}^\lambda = \sum_{i=1}^n \lambda_i X_{s_i}}} Var(X_{s_0}^\lambda - X_{s_0}) \tag{2}$$

such that $\mathbb{E}[X_{s_0}^\lambda] = m_{s_0}$.

3 Conformal Prediction for Functional Kriging

To evaluate the uncertainty of a predicted curve, $X_{s_0}^*(t)$, from a new site s_0 without making any assumptions about the distribution of the data, the Conformal Prediction (CP) method is proposed. Given nominal miscoverage level $\alpha \in (0, 1)$, we can define a prediction band $C \subset \mathscr{L}_2(T) \times D$ based on the observed data $(X_{s_1}(t), \dots, X_{s_i}(t), \dots, X_{s_n}(t))$ such that the probability of the true curve $X_{s_0}(t)$ falling within the band is at least $1 - \alpha$, as expressed by equation 3.

$$\mathbb{P}(X_{s_0}(t) \in C(s_0)) \geq 1 - \alpha, \tag{3}$$

Here, the probability is taken over s_0 independent and identically distributed (i.i.d.) draws by $X_{s_0}(t), X_{s_1}(t), \dots, X_{s_n}(t) \sim P$, and $C(s_0) = \{X(t) \in \mathscr{L}_2(T) : X_{s_0}(t) \in C\}$ represents the set of curves in $\mathscr{L}_2(T)$ where $X_{s_0}(t)$ is contained, for a given point $s_0 \in D$.

The concept of conformal prediction involves considering all possible curves for the test object and assessing their conformance to the set of training examples. By comparing the test curve, $X_{s_0}(t)$, to the data set, the likelihood that the conformal curve $C_{conf}(X_{s_0}(t))$ represents the class of $X_{s_0}(t)$ is obtained. In other words, it measures the extent to which $X_{s_0}(t)$ conforms to the training data set.

Following the approach described in [4], we define a prediction band as follows:

$$C_{conf}(s_0) = \left\{ X(t) \in \mathscr{L}_2(T) : X(t) \in \left[X_{s_0}^*(t) - k^S S(t),\, X_{s_0}^*(t) + k^S S(t) \right] \right\}, \tag{4}$$

where $X_{s_0}^*(t)$ represents the mean of the estimated kriging model and serves as the center of the prediction band. The value k^S determines the width of the prediction band, and $S(t)$ is a modulation function. Specifically, k^S is computed as the $(1 - \alpha)$ quantile of the values $\{R_{X_{s_0}^*(t), s_i} :\ i = 1, \dots, n\}$. Intuitively, a prediction band should adjust its width across T based on the local variability of the data, which is accomplished through the modulation function $S(t)$. The computation of k^S depends on $S(t)$, as a general $R_{X_{s_0}^*(t), s_i}$ is obtained using a nonconformity measure as follows:

$$R_{X_{s_0}^*(t), s_i} = \mathscr{D}_{sup}\left(X_{s_i}(t) - X_{s_0}^*(t) \right) = \sup_{t \in T} \left| \frac{X_{s_i}(t) - X_{s_0}^*(t)}{S(t)} \right|. \tag{5}$$

4 Real data analysis

To demonstrate the practical application of the discussed method in this work, this section is dedicated to a case study involving data provided from the Italian Telecomunication company (TIM). The number of mobile users connected to the 4G network was collected by TIM in an area surrounding the province of Milan. The data collection is spanned from March to August 2020, with measurements recorded at 15-minute intervals. As a result, a time series dataset is obtained for each specific location in Milan. We at first, perform a preprocessing step where we trasform spatially located time series in georeferenced functional data. Then with the aim to predict hypothetical number of connected users in an unobserved site, we perform a Conformal kriging prediction. The use of Conformal Prediction allows for the calculation of dependable prediction intervals and measures of uncertainty based on the available data. This approach improves decision-making and analysis, especially when exploring the potential benefits of integrating additional sensors into the 4G mobile network. In Figure 1, the conformal interval, with $\alpha = 0.05$ for a new spatial location is shown. The figure displays the conformity band for a location far from

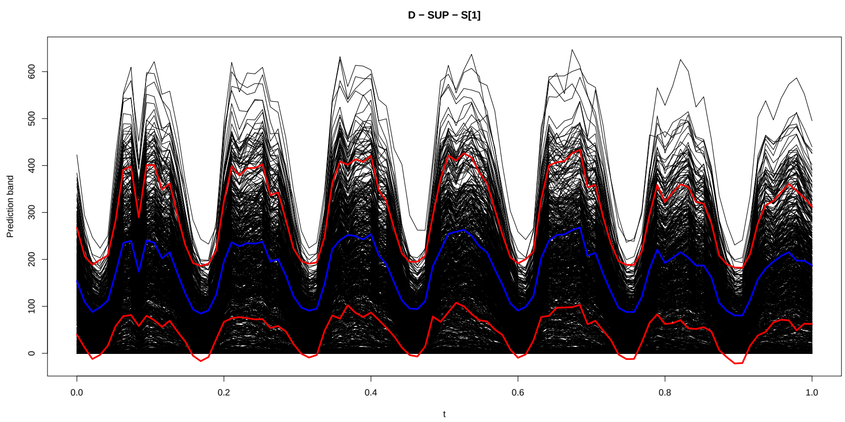


Fig. 1 Prediction Band on $s_0 = 2$, for one day of observation.

the others where data were collected. Consequently, it demonstrates how the prediction aligns with the estimation on the observed data. For the sake of brevity, we report only this result of the analysis. The extended version of the work contains a detailed description of the further results.

Acknowledgements The research leading to these results has received funding from the project aDversarial scenArios geneRation With dIgital twiNs In induSTry (DARWINIST), funded by the program of fundamental and applied research projects dedicated to young researchers, Università della Campania "Luigi Vanvitelli", D.R. 834 del 30/09/2022.

References

1. Caballero, W., Giraldo, R., Mateu, J.: A universal kriging approach for spatial functional data. Stoch. Environ. Res. Risk Assess. 27 (7), 1553-1563 (2013)
2. Delicado, P., Giraldo, R., Comas, C. and Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric 21, 224-239 (2010)
3. Diana, A., Romano, E., Irpino: Conformal prediction for spatio-functional regression models. In: Book of short papers SIS 2022. PEARSON, ISBN: 978-88-9192-736-1 (2022)
4. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal Prediction Bands for Multivariate Functional Data. Journal of Multivariate Data Analysis (2022)
5. Franco-Villoria, M., Ignaccolo, R.: Bootstrap based uncertainty bands for prediction in functional kriging, Spatial Statistics, 21A, 130–148 (2017). https://doi.org/10.1016/j.spasta.2017.06.005.
6. Giraldo, R., Delicado, P., Mateu, J.: Ordinary kriging for function-valued spatial data. Environ Ecol Stat 18, 411–426 (2011). https://doi.org/10.1007/s10651-010-0143-y
7. Ignaccolo, R., Mateu, J., Giraldo, R.: Kriging with external drift for functional data for air quality monitoring. Stoch. Environ. Res. Risk Assess. 28, 1171-1186 (2014)
8. Mateu, J., Romano, E.: Advances in spatial functional statistics. Stochastic Environmental Research and Risk Assessment, 31, 1-6 (2017)
9. Mateu, J., Giraldo, R. (Eds.). Geostatistical functional data analysis. John Wiley & Sons (2021)
10. Menafoglio, A., Secchi, P., Dalla Rosa, M.: A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electron. J. Statist. 7, 2209-2240 (2013). DOI: 10.1214/13-EJS84
11. Nerini, D., Monestiez, P., Manté, C.: Cokriging for spatial functional data. J. Multivariate Anal. 101, 409-418 (2010)
12. Reyes, A., Giraldo, R., Mateu, J.: Residual Kriging for functional spatial prediction of salinity curves. Commun. Stat. - Theory Methods 44 (4), 798-809 (2015)
13. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)

Measuring Public-Private Connectedness in Financial Markets

La Misurazione della Connessione Pubblico-Privato nei Mercati Finanziari

Javier Sánchez García and Salvador Cruz Rambaud

Abstract Knowledge about the interconnectedness of private and public financial markets is crucial in order to understand when collapses and bankruptcies in the financial sector can propagate and become sovereign debt crises. Similarly, the reverse is essential in order to prevent when uncertainty about the government debt interest rate path of a country due to factors such as idiosyncratic shocks, incertitude in the macroeconomic fundamentals, or political concerns can translate into volatility which, in turn, can propagate to its financial institutions leading to a systemic crisis. This paper analyzes the connectedness of public-private financial markets of several European countries by forming a dynamic network and analyzing its degree distribution, clustering and other relevant network statistics by using the TERGM.
Abstract *La conoscenza dell'interconnessione tra i mercati finanziari privati e pubblici è fondamentale per capire quando crolli e fallimenti, nel settore finanziario, possano propagarsi e trasformarsi in crisi del debito sovrano. Allo stesso modo, è essenziale evitare che l'incertezza sul tasso di interesse del debito pubblico di un paese, a causa di fattori quali shock idiosincratici, incertezza nei fondamentali macroeconomici o preoccupazioni politiche, possa tradursi in volatilità che può propagarsi alle istituzioni finanziarie portando a una crisi sistemica. Questo studio si focalizza sulla connessione dei mercati finanziari pubblico-privati di diversi paesi europei che formano una rete dinamica e analizza la sua distribuzione dei gradi, il raggruppamento e altre statistiche di rete rilevanti utilizzando il TERGM.*

Key words: Public-private contagion, financial connectedness, financial stress, financial shock

Javier Sánchez García
Department of Economics and Business, and Mediterranean Research Center for Economics and Sustainable Development (CIMEDES), University of Almería, La Cañada de San Urbano, s/n, 04120, Almería, Spain, e-mail: jsg608@ual.es

Salvador Cruz Rambaud
Department of Economics and Business, and Mediterranean Research Center for Economics and Sustainable Development (CIMEDES), University of Almería, La Cañada de San Urbano, s/n, 04120, Almería, Spain, e-mail: scruz@ual.es

1 Introduction

Due to globalization and technological development, the world is becoming increasingly interconnected each day, and so are financial markets. The internationalization of financial markets and financial globalization has facilitated simultaneous transactions between economic agents in many parts of the world. This has translated into an enormous flow of funds from countries with surpluses to countries with deficits [4, 5, 11, 15].

However, at the same time, this interconnectedness means that shocks to financial markets in one side of the world can also propagate to any other part in the world with ease. Since positions are taken at a multinational level, the financial system now has the form of a complex multilayered system [6, 10, 17]. This means that panics, liquidity shortages, overindebtness schemes or even defaults can propagate through the world without adequate macroprudential policies.

It is already known in the literature that many financial crises have been preceded by movements in asset prices [1, 2]. Detailed examples can be found in [18]. In fact, the 2007-2009 financial crisis turned into a public debt crisis in 2012 when Greece announced its inability to repay debt. This was the origin of what is now known as the European debt crisis [3, 13, 16].

However, little is known about how much public and private markets co-move, during which kind of shocks, and which factors do influence their interconnectedness. This is of upmost importance to researchers and practitioners, since knowing when public-private financial markets are highly interconnected can help understand if and when a shock to the financial sector can propagate to public debt markets, or on the other hand, when a shock to government debt can propagate to stock markets forcing companies to conduct asset fire sales or suffer from liquidity shortages.

This paper builds a network model to analyze the financial connectedness of public-private markets in Europe. From there, it employs key network statistics such as the degree distribution and clustering coefficient, and links the results to important economic and financial periods with a narrative approach. Finally, a dynamic network econometric Temporal Exponential Random Graph Model is proposed to analyze the effect of several macroeconomic and financial variables on that connectedness.

2 Data

The dataset employed in this paper are the daily closing log-returns of the stock and bond market indices of several European countries. The stationarity of all the series is tested by employing KPSS and ADF-GLS tests. The results of other relevant summary statistics are additionally reported. Nevertheless, the structural analysis tools for building the network which are succinctly presented in Section 3 are known to be robust to I(1) processes [12, 14].

For the inferential part, a set of relevant macroeconomic and financial variables of the Euro area have been chosen. These data are available to the public at the ECB statistics warehouse. Key summary statistics have been also reported.

3 Methodology

The methodological approach of this paper consists of three parts. Firstly, a directed network is constructed by employing the FEVD of a VAR model as in [8, 9, 7]. Secondly, relevant network statistics have been analyzed in order to study the topology of the network. Thirdly, an empirical model based on a Temporal Exponential Random Graph Model (TERGM) has been built and applied to the connectedness of the network considered as the dependent variable in order to analyze which macroeconomic and financial factors drive it.

4 Conclusion

Learning about the connectedness of public and private financial markets is of particular interest during periods of financial stress, as it could lead to implement macroprudential policies to avoid the propagation of a private sector shock to government debt and vice versa. In this paper, a network econometric analysis including several countries of Europe has been conducted in order to understand to which extend and under which conditions private and public financial markets are interconnected.

The results are of particular interest to researchers and practitioners in the fields of financial stability.

Acknowledgements Mediterranean Research Center for Economics and Sustainable Development (CIMEDES), University of Almería.

References

1. Amihud, Y., Mendelson, H., Pedersen, L.H.: Market Liquidity: Asset Pricing, Risk, and Crises. Cambridge University Press, New York (2012)
2. Baily, M.N., Litan, R.E., Johnson, M.S.: The origins of the financial crisis. YPFS Documents, 5825 (2009)
3. Beirne, J., Fratzscher, M.: The pricing of sovereign risk and contagion during the European sovereign debt crisis. Journal of International Money and Finance 34, 60—82 (2013)
4. Bernanke, B.S., Bertaut, C.C., Demarco, L., Kamin, S.B.: International capital flows and the return to safe assets in the United States, 2003-2007. FRB International Finance Discussion Paper, 1014 (2011)
5. Byrne, J.P., Fiess, N.: International capital flows to emerging markets: National and global determinants. Journal of International Money and Finance, 61, 82—100 (2016)
6. Correa, R., Goldberg, L.S.: Bank complexity, governance, and risk. Journal of Banking & Finance, 134, 106013 (2022) doi: 10.1016/j.jbankfin.2020.106013
7. Demirer, M., Diebold, F.X., Liu, L., Yilmaz, K.: Estimating global bank network connectedness. Journal of Applied Econometrics 33(1), 1-15 (2018)
8. Diebold, F.X., Yilmaz, K.: Better to give than to receive: Predictive directional measurement of volatility spillovers. International Journal of Forecasting 28(1), 57–66 (2012)
9. Diebold, F.X., Yılmaz, K.: On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics 182(1), 119–134 (2014)
10. Glasserman, P., Young, H.P.: Contagion in financial networks. Journal of Eco- nomic Literature 54(3), 779-831 (2016)
11. Iamsiraroj, S.: The foreign direct investment–economic growth nexus. International Review of Economics & Finance 42, 116–133 (2016)
12. Kilian, L., Lütkepohl, H.: Structural Vector Autoregressive Analysis. Cambridge University Press, New York (2017)
13. Lane, P.R.: The European sovereign debt crisis. Journal of Economic Perspectives 26(3), 49–68 (2012)
14. Lütkepohl, H.: Vector Autoregressive Models. In: Hashimzade, N., Thornton, M.A. (eds.), Handbook of Research Methods and Applications in Empirical Macroeconomics, pp. 139–164. Edward Elgar Publishing, Massachusetts (2013)
15. Matsumoto, H.: Foreign reserve accumulation, foreign direct investment, and economic growth. Review of Economic Dynamics, 43, 241–262 (2022)
16. Ongena, S., Popov, A., Van Horen, N.: The invisible hand of the government: Moral suasion during the European sovereign debt crisis. American Economic Journal: Macroeconomics, 11(4), 346-379 (2019)
17. Poledna, S., Molina-Borboa, J.L., Martınez-Jaramillo, S., Van Der Leij, M., Thurner, S.: The multi-layer network nature of systemic risk and its implications for the costs of financial crises. Journal of Financial Stability, 20, 70-81 (2015)
18. Reinhart, C.M., Rogoff, K.S.: This Time is Different: Eight Centuries of Financial Folly. Princeton University Press, New Jersey (2009)

An original approach to anomalies in intertemporal choices through Functional Data Analysis: Theory and application for the study of Hikikomori syndrome

Un approccio originale per le anomalie nelle scelte intertemporali attraverso l'Analisi Funzionale dei Dati: Teoria e applicazione per lo studio della sindrome di Hikikomori

Roberta Martino, Viviana Ventre, Salvador Cruz Rambaud and Fabrizio Maturo

Abstract The pattern of intertemporal preferences is related to critical behavioural aspects involving the individual emotional and cognitive spheres. The increased precision in the structure of the discount function provided by recent studies increases the complexity of analysing data. In this regard, Functional Data Analysis (FDA) offers great support by improving data interpretation for two reasons: it allows the investigation of mathematical properties of the discount function; and it enables the identification of interrelationships among factors characterising the decision maker. Therefore, the present study shows how the integration of FDA increases the descriptive power of the empirical data in qualitative and quantitative terms. The case study is the condition of Hikikomori, a new social phenomenon.

Abstract *L'andamento delle preferenze intertemporali è legato ad aspetti comportamentali critici che coinvolgono la sfera emotiva e cognitiva individuale. La maggiore precisione nella struttura della funzione di sconto fornita da studi recenti aumenta la complessità dell'analisi dei dati. A questo proposito, l'analisi funzionale dei dati (FDA) offre un grande supporto poichè migliora l'interpretazione dei dati per due motivi: permette di indagare le proprietà matematiche della funzione di*

Roberta Martino
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Viale A. Lincoln, 5, 81100, Italy, e-mail: roberta.martino@unicampania.it.

Viviana Ventre
Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Viale A. Lincoln, 5, 81100, Italy e-mail: viviana.ventre@unicampania.it.

Salvador Cruz Rambaud
Department of Economics and Business, and Mediterranean Research Center for Economics and Sustainable Development (CIMEDES), University of Almería, La Cañada de San Urbano, s/n, 04120, Almería, Spain, e-mail: scruz@ual.es.

Fabrizio Maturo
Universitas Mercatorum, Rome, Italy, e-mail: fabrizio.maturo@unimercatorum.it.

sconto; consente di identificare le interrelazioni tra i fattori che caratterizzano il decisore. Pertanto, il presente studio mostra come l'integrazione della FDA aumenti il potere descrittivo dei dati empirici in termini qualitativi e quantitativi. Il caso studio è la condizione di Hikikomori, un nuovo fenomeno sociale.

Key words: Hikikomori, Discount Function, Intertemporal Choice, Inconsistency, Functional Data Analysis, Hyperbolic Factor

1 Introduction

Intertemporal choices are the kind of decisions whose effects are only evident over time. The research aims to understand how individuals interact with alternatives spread over several periods. The primary reference model for studying intertemporal preferences is the Discounted Utility model [1], in which the utility of a future outcome is obtained by reducing the present utility by a factor expressed by the discount function. The discount function quantifies the individual's perception of the indeterminacy of the future and is described by two elements: the discount factor [2] and impatience [3].

In line with the principles of economic rationality, the trend of the discount function was initially assumed to be exponential, characterised by a constant discount factor and a constant degree of impatience, necessary to preserve the consistency of preferences over time [4]. However, empirical evidence proved that the exponential model is far from the behaviour of individuals in contrast with hyperbolic models, closer to empirical preferences and characterised by a degree of impatience and a discount factor which are not constant over time [5]. Due to the critical link between behaviour and intertemporal preferences, research has turned to intertemporal choice to interpret and quantify the mechanisms of different disorders. In particular, the delayed discounting (DD) task is the most widely used measure of individual behaviour and neurobiology indicators [6].

Recently, an even more precise method for empirically defining the intertemporal discount function of the DD task has been proposed by [7, 8, 9]. As the descriptive capacity of the discount function increases, it becomes more difficult to detect subtle individual differences. Thus, the fundamental problem of interpreting intertemporal preferences for behaviour analysis is understanding how to distinguish those discount functions which express more significant degrees of non-rationality and different behavioural properties.

The starting idea of this research is that Functional Data Analysis (FDA) could support the study of discount function behaviour because the cognitive mechanisms underlying decision-making can be expressed and represented as curves over time. The application of FDA to intertemporal decision-making can provide an in-depth description of the diversity of behavioural profiles. It could have a decisive impact on two different aspects: firstly, FDA allows for the investigation of the essential properties of discount functions, such as first and second derivatives, curvature, the radius of curvature and arc length [10]; secondly, FDA could identify relationships between factors which *a priori* are unknown. This aspect is fundamental, especially in studying new multifactorial social phenomena. In addition, FDA allows the study of data even when they have high dimensionality [11, 12, 13, 14].

This research concentrates on the phenomenon known as the Hikikomori syndrome, a condition of isolation which has become especially worrying in recent decades. The conceptualisation of Hikikomori is proposed through a biopsychosocial model [15], which aims to include personality dysfunction, social disturbance, psychosis, bullying, anxieties, paranoia, and obsessions. The expansion of this phenomenon suggests more needs to know about its causes and characteristics to develop therapeutic and rehabilitative strategies. The aim of this study is to understand how Hikikomori subjects behave with respect to the discount function.

The data needed for the analysis were collected by means of a questionnaire divided into two tasks: the first one collects both the values of the discount function at different times and the individual degrees of impatience decrease; the second one contains the results related to the 25-item Hikikomori questionnaire [16].

2 Methods

2.1 Non-rationality in intertemporal choice

The DD task consists in giving two different options: one with an early (SS) and one with a delayed (LL) payoff. The normative model does not define the best alternative, since it depends only on individuals' impatience and their perception of future uncertainty. The normative assumption is only that the preference remains constant over time. Temporal inconsistency indicates that preferences are not constant over time creating a gap between the empirical evidence and the normative model. Two different measures of inconsistency are used in the present paper. The first measure is the degree of decrease in impatience through the hyperbolic factor [17]. This measure quantifies the direct impact of the emotional factor on intertemporal decisions [8], related to specific anomalies in the discounted utility model: delay effect [18], size effect [19], interval effect, sign effect [20, 21]. The second measure of inconsistency directly relates to individual temporal distortions, as proven and defined by [8]. The proposed approach constructs a unique discount function for each individual by interpolating values over the time intervals [8, 9]. Unlike the DD task, where the alternatives are established *a priori* and the discount function is constructed using approximation techniques, this approach is more authentic in terms of behaviour expression and retains much more information, allowing for a more in-depth study.

The questionnaire to obtain the hyperbolic function for each individual is based on discovering the delay effect $U_D(x(0))$ fixed $t_0 = 0$, using the following question: "you have to receive $U\left(x\left(t_i\right)\right)$ euros today, how much do you want to receive in t_{i+1} days to consider the offer equivalent?". The function is given by:

$$f(t) = \begin{cases} f(0) = 1 \\ f(t_{i+1}) = f(t_i)\frac{U(x(t_i))}{U(x(t_{i+1}))} \end{cases}$$

2.2 FDA for non-rationality in intertemporal choice

Functional Data Analysis (FDA) [11] is proposed to fit the available data to a wide range of discount functions which reflect different degrees of inconsistency. Each vector of indifference pairs, according to individuals' responses, can be seen as a function in the time domain rather than a finite-dimensional vector. Indeed, it is possible to represent the sequences of individual discrete observed data as functions, analyzing them as single entities. Assuming that sample paths belong to a finite-dimension space spanned by a basis the function $f(t)$ could be represented by a basis expansion $f(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t),\ i = 1,\ldots,n$ where: $f_i(t)$ is the reconstructed function for the i-th unit (time points could be different for each unit); $\phi_k(t)$ are linearly independent and known functions, called basis functions; a_{ik} are coefficients that link each basis function together in the representation of $f_i(t)$ To be considered a discount function, a b-spline approximation must also satisfy the conditions: $f(0) = 1$; $f(t) > 0$; and $f(t)$ is strictly decreasing. Therefore, the monotone smoothing approach proposed by Ramsay and Silverman [11] is considered. Moreover, when functions must satisfy one or more constraints, linear combinations of basis functions are difficult to constrain for this purpose. Ramsay and Silverman [11] suggested to transform the problem into one where the curve being estimated is unconstrained.

3 Results

The first results are already evident using the boxplots, shown in Figure 1 and Figure 2. In general, the boxplot provides a summary view of the main characteristics of the distribution of a dataset, such as its dispersion, symmetry or skewness, and the presence of outliers. In Figure 1, the median, represented by a horizontal line in each rectangle, is decreasing over time and its decreasing trend is hyperbolic. This confirms the efficiency of the data collected and emphasises that, as the outliers increase with time, the values that do not follow the hyperbolic trend deviate significantly from the main distribution of the data. Figure 2 uses boxplots to compare the distributions between the category of Hikikomori and non-Hikikomori. In particular, it is possible to observe how the median of the Hikikomori deviates from the time interval $t = 7$ days, assuming generally larger values. This result, with respect to the decision profile, is in line with a different perception of time between Hikikomori and non-Hikikomori subjects. The different box lengths also indicate that in general Hikikomori subjects apply a gentler discount than Hikikomori subjects, an attitude also confirmed by the lower presence of outliers. Figure 3 shows the discount functions represented via the monotone smoothing approach.

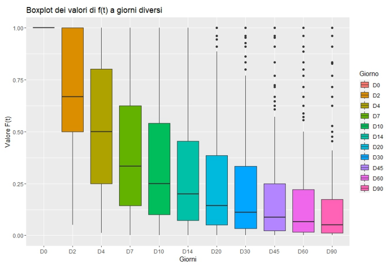


Fig. 1 Boxplot of the distribution of $f(t)$ values at different days.

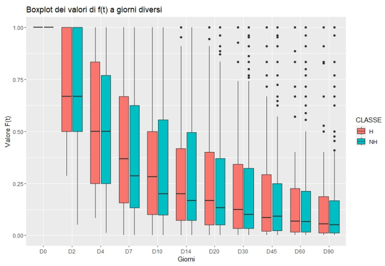


Fig. 2 Boxplot of the distribution of $f(t)$ values at different days for Hikikomori (H) and non-Hikikomori (NH) condition.

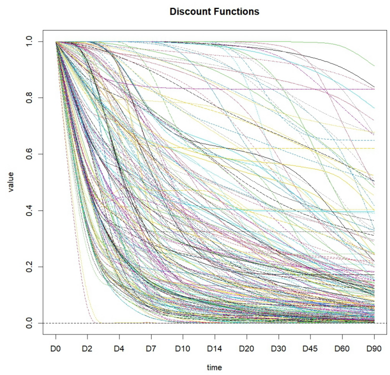


Fig. 3 Smoothed monotone discount functions.

References

1. Samuelson, P. A.: Probability, utility, and the independence axiom. Econom.: Journal of the Econom. Soc., 670-678 (1952)
2. Read, D.: Blackwell Handb. of Judgm. and Decis. Mak.; John Wiley & Sons: Hoboken, NJ, USA, 424–443 (2008)
3. Cruz Rambaud, S., & Muñoz Torrecillas, M. J.: Measuring impatience in intertemporal choice. PLoS One, 11(2), e0149256 (2016)
4. Green, L., Myerson, J., & McFadden, E.: Rate of temporal discounting decreases with amount of reward. Mem. & cogn., 25(5), 715-723 (1997)
5. Prelec, D.: Decreasing impatience: a criterion for Non-stationary time preference and "hyperbolic" discounting. The Scand. J. of Econ., 106(3), 511-532 (2004)
6. Kable, J. W.: Valuation, intertemporal choice, and self-control. Neuroeconomics, 173-192 (2014)
7. Ventre, V., Cruz Rambaud, S., Martino, R., & Maturo, F.: An analysis of intertemporal inconsistency through the hyperbolic factor. Qual. & Quant., 1-28 (2022)
8. Ventre, V., & Martino, R..: Quantification of Aversion to Uncertainty in Intertemporal Choice through Subjective Perception of Time. Math., 10(22), 4315 (2022)
9. Ventre, V., Martino, R., & Maturo, F.: Subjective perception of time and decision inconsistency in interval effect. Qual. & Quant., 1-26 (2022)
10. Di Battista, T., Fortuna, F., and Maturo, F.: BioFTF: An R package for biodiversity assessment with the functional data analysis approach. Ecol. Indic., 73, 726-732 (2017)
11. Ramsay, J, Silverman, B.: Functional Data Analysis. 2nd ed. New York, NY: Springer (2005)
12. Maturo, F., & Verde, R. Pooling random forest and functional data analysis for biomedical signals supervised classification: Theory & application to electrocardiogram data. Statistics in Medicine, 41(12), 2247-2275 (2022)
13. Maturo, F., & Verde, R. Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. Computational Statistics (2022)
14. Maturo, F., & Verde, R. Supervised classification of curves via a combined use of functional data analysis and tree-based methods. Computational Statistics, 38, 419-459 (2023)
15. Kato, T. A., Shinfuku, N., Sartorius, N., & Kanba, S.: Are Japan's Hikikomori and depression in young people spreading abroad?. The Lancet, 378 (2011)
16. Amendola, S., Presaghi, F., Teo, A. R., & Cerutti, R.: Psychometric properties of the Italian version of the 25-item Hikikomori Questionnaire. Int. j. of envir. Res. and public health, (2022)
17. Rohde, K. I.: The hyperbolic factor: A measure of time inconsistency. J. of Risk and Uncertain., 41, 125-140 (2010)
18. Thaler, R.: Some empirical evidence on dynamic inconsistency. Econ. Lett., 8(3), 201-207 (1981)
19. Noor, J.: Intertemporal choice and the magnitude effect. Games and Econ. Behav., 72(1), 255-270 (2011)
20. Loewenstein, G., & Thaler, R. H.: Anomalies: intertemporal choice. J. of Econ. Perspect., 3(4), 181-193 (1989)
21. Cruz Rambaud, S., & Sánchez Pérez, A. M.: The magnitude and "peanuts" effects: Searching implications. Front. in Appl. Math. and Stat., 4, 36 (2018)

Solicited Session SS6 - *Labour market: trends, perspectives and new challenges*

Organizer: Matilde Bini
Chair: Andrea Cutillo

1. *Enriching Job Vacancy Official Information with Online Job Advertisements: Chances and Limits* (Lucarelli A. and Righi A.)
2. *Innovation in Management: towards the Open Manager* (Bruttini P., Gallo M., Mariani P. and Menini T.)

Enriching Job Vacancy Official Information with Online Job Advertisements: Chances and Limits

Arricchire le informazioni ufficiali sui posti di lavoro vacanti con annunci di lavoro dal web: possibilità e limiti

Annalisa Lucarelli and Alessandra Righi

Abstract Thanks to the increasing use of portals, online job advertisements (OJAs) are a powerful source of information on job requirements. After explaining the contents of the current official job vacancy survey and the new OJA information system collecting data from the web, we show some examples of how these data can enrich the official job vacancy indicators. Assessing the limitations of the web information explains why the OJA cannot replace other types of labour market information but can provide comprehensive, detailed and timely information on labour market trends in terms of occupations, skills, and types of education required.
Abstract *Grazie al crescente utilizzo dei portali di annunci di lavoro online, questi sono una rilevante fonte di informazione sulle richieste di lavoro. Dopo aver illustrato i contenuti della fonte ufficiale sui posti di lavoro vacanti e del nuovo sistema informativo degli annunci di lavoro sul web, si mostrano esempi di come questi dati possano arricchire gli attuali indicatori ufficiali. La valutazione dei limiti delle informazioni derivate dal web spiega perché gli annunci, pur non potendo sostituire altre fonti sul mercato del lavoro, possono fornire informazioni dettagliate e tempestive sui cambiamenti in termini di professioni, competenze e tipo di formazione richieste.*

Keywords: job search, labour demand, web-data

Annalisa Lucarelli
Istat; e-mail: anlucare@istat.it

Alessandra Righi
Istat; e-mail: righi@istat.it

1 Introduction

Online job advertisements (OJAs) create opportunities to enrich and supplement job vacancy official indicators – as required by the EU Regulations – thanks to the development of online job boards and portals, which may support and deepen labour market statistics.

In the context of the official statistics covering the demand side of the labour market, there are already job vacancy surveys that supply quarterly information on the unmet labour demand. A vacancy is defined as an unoccupied workplace that a company wants to have filled. Company representatives must take action to hire a worker. This makes job vacancy statistics (JVS) highly susceptible to economic expectations. It also connects vacancies to short-run economics and makes job vacancy statistics leading indicators of the economic cycle. As job vacancies measure employment intentions that have materialized in candidates' searches, they can give "early warnings" on the dynamics of jobs in the near future. Data on job vacancies are used by the European and the European Central Bank to monitor short-term developments in the business cycle and the labour market. The vacancy indicator currently used at the European level – the job vacancy rate – is one of the Principal European Economic Indicators PEEIs on the labour market. A new integrated EU Regulation on labour market statistics on the business side, is in the process of being approved, allowing the use of innovative sources (web scraping) to obtain detailed information on labour shortages by occupation and region. The collection of this type of information through surveys would entail additional costs for NSIs and an increased statistical burden for enterprises.

OJAs provide not only highly detailed information regarding the job positions advertised but also at high frequency (on a daily basis). Before complementing official statistics, and/or developing estimator(s) for the number of job vacancies from OJAs, issues on coverage and representativeness of online job offers need to be addressed, accounting for the differences in the statistical unit and coverage [1].

After illustrating the characteristics of the vacancy sources, the official JV survey, on the one hand, and the OJA Cedefop data system, on the other, we consider some limits of the OJA information and explain why OJAs cannot replace JV official statistics but can provide comprehensive, detailed, and timely insights into labour market trends. Then, we present two possible uses of OJAs that, despite their coverage and representative limits, can complement job vacancy official indicators. Some final considerations conclude the work.

2 Data sources

European Regulation No 453/2008 defines a job vacancy as: "a paid post that is newly created, unoccupied, or about to become vacant: a) for which the employer is taking active steps and is prepared to take further steps to find a suitable candidate from outside the enterprise concerned; and (b) which the employer intends to fill either immediately or within a specific period of time." The active search for a suitable candidate must have already started but not yet ended at the reference time at which job vacancies are measured and it is not sufficient for the enterprise to express the intention to undertake these actions in the future. The active steps to find a suitable candidate include: (i) notifying the job vacancy to public employment services; (ii) contacting a private employment agency/head hunters; (iii) advertising the vacancy in the media (for example the Internet, newspapers, magazines); (iv) advertising the vacancy on a public notice board; (v) approaching, interviewing or selecting possible candidates/potential recruits directly; (vi) approaching employees and/or personal contacts; (vii) using internships.

The job vacancy rate, the main indicator disseminated at the European level, measures for how many jobs these search/selection activities are in progress out of every 100 positions, which are either already occupied or for which a recruitment process is taking place. The Istat job vacancy rate measures vacancies that were open on the last day of the quarter and covers enterprises in Industry and Services with at least one employee. It is available only at the national level and at the level of the Nace Rev. 2 activity sections.

OJAs refer to advertisements published on several online sources as job portals, company sites, social networks, employment websites, employment agencies, job search engines, online newspapers, public employment services, and employers organizations. OJA data offer detailed information regarding the characteristics of the job (e.g. occupation, location, type of contract, working time, and salary); the characteristics of the employer (e.g. economic activity sector); the job requirements (e.g. education, skill, and experience); and also information on the advertisement (e.g., job portal and publishing and the expiring date of the ad). There is no exact correspondence between an advertisement and a job because each advertisement can relate to different job positions or cannot report the number of positions requested.

The European Center for the Development of Vocational Training (Cedefop) collects OJA data at the European level by 2018. Data are released quarterly although they are available on a daily basis. All the data collected are classified by well-known international standard classifications.

The Cedefop team has developed a well-structured system, not only for collecting but also for processing data, which consists of several steps ranging from data cleaning, standardisation, classification, and the application of validation and plausibility rules [2]. These data, although still experimental, are undergoing a gradual process of improvement in order to achieve higher quality standards and to meet the requirements of Eurostat Big Data Task Force and the European Statistical Systems Network (ESSnet). The comparison of the two sources shows that online job advertisements form a specific fraction of the job vacancy market. The concept of OJAs does not correspond to that of JVS, due to different factors. In particular, online job vacancies should cover all those job vacancies, as defined by the EU Regulation, for which the active steps carried out by the employers to find a suitable candidate include also advertising on the Internet job portals. Whilst there is a general growing trend in the number of job vacancies being offered online, many vacancies continue to be filled through traditional channels, such as newspapers, employment agencies, noticeboards, or personal contacts. Furthermore, job portals used by the employers may not be totally covered by data ingestion activities of Cedefop or other organisations and, even after collecting all the advertisements published on all existing online portals it is still unclear if they refer to total job vacancies that actually exist at a specific reference date. A job advertisement is only a proxy measure for the existence of a job vacancy within a company. For example, delays in the communication between enterprises and online portals could cause the presence of expired job vacancies on portals. In addition, OJAs may not represent a job vacancy in the scope of the official survey as a company might potentially only try to investigate the market. Analysing the relation between OJAs and JVS published by Eurostat, Beręsewicz, and Pater [3] found that the OJA dataset identified 40% - 91% job offers more than JVS did and that the OJAs show a lot more variation than JVS. Besides differences in the distribution shapes, the relations between OJAs and JVS differ significantly among countries, industries, and occupations.

3 How OJAs can be used to complement JVS

In this section, we present descriptive analyses, which exploit the detailed and high-frequency information included in the OJAs, to suggest different ways in which the OJAs can enrich the current official JVS information. The first analysis refers to the use of monthly OJAs data to compute a monthly basis JV official rate from the quarterly one; the second one refers to the use of the OJAs by occupation to study changes in the types of occupation over time.

The job vacancy rate is the primary source of information used to analyse and monitor short-term cyclical economic developments within the EU/euro area and at the national level. Even if the current EU Regulation requires the job vacancy rate on a quarterly basis, a monthly job vacancy rate, indirectly derived from the OJAs, can be used to better focus the dynamics of the cycle within the quarter, besides other monthly macroeconomic indicators.

We used the monthly average number of OJAs to estimate the vacancy rate of the survey on a monthly basis for Italy, as follows:

$$Monthly\ JV\ rate = quarterly\ JV\ rate * \frac{monthly\ mean\ OJAs}{quarterly\ mean\ OJAs}$$

Our monthly estimates are based on the monthly OJAs data from July 2018 to June 2022 and on the current not seasonally adjusted job vacancy rate quarterly estimates, derived from the Italian official survey (Figure 1). Since there is no evidence of a clear daily trend of OJAs within each month, we preferred to use the monthly and quarterly average of OJAs instead of the number of OJAs on a specific day (e.g. the last day of the month or quarter). The graphic analysis shows for the two series quite similar results, apart from some divergences in the second and third quarters of 2019 and in the first quarter of 2022.

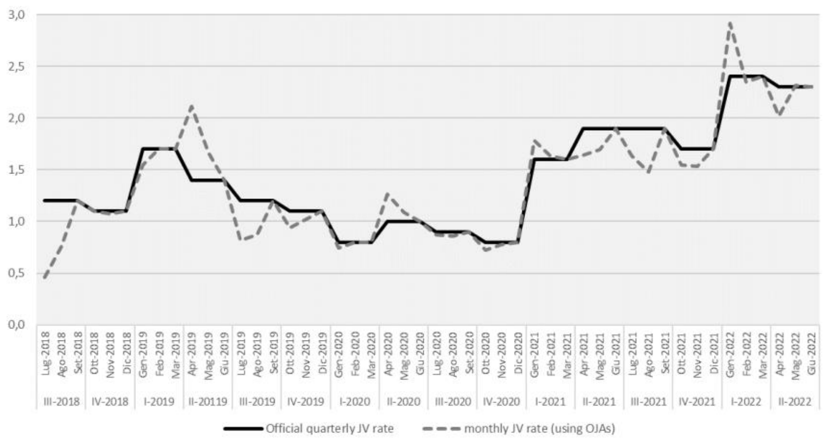


Fig. 1 Quarterly official job vacancy rate (Istat, Vela survey) *vs* a monthly basis rate (Cedefop OJA data), III 2018 - II 2022, Italy

Figure 2 shows the quarterly and monthly job vacancy rate year-on-year changes (percentage points differences) calculated by comparing the vacancy rates of each quarter and the month to the same quarter and month of the previous year, respectively. With respect to the quarterly year-on-year change, which is an average of the monthly year-on-year changes, the monthly change gives additional information on the dynamics inside the quarter. We can note that the drop in the job vacancy rate in the first quarter of 2020 (-0.9) depends not only on a fall in the month of March – when the Italian lock-down to address the COVID-19 health emergency started –, but on a fall beginning also earlier in the month of January. While, the decrease in the second quarter of 2020 is mainly due to that observed in April, coherently with the effect and duration of the emergency government interventions.

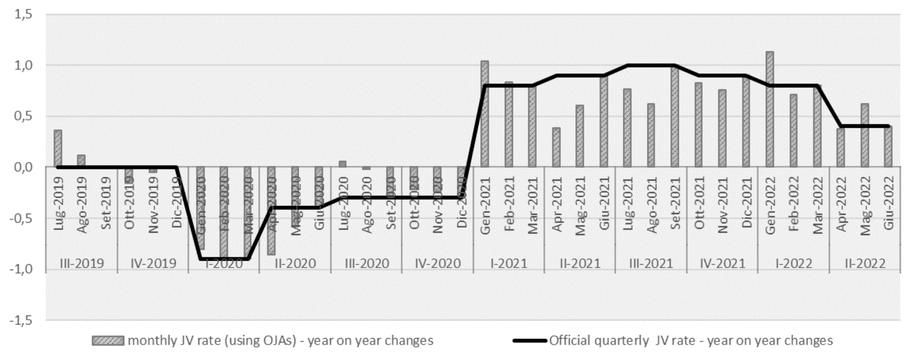


Fig. 2 Quarterly official job vacancy rate (Istat, Vela survey) *vs* monthly basis rate (Cedefop OJA data), year on year changes, III 2019-II 2022, Italy

OJAs enrich the labour market information also providing the dynamics of occupations required by employers through online ads. Even if the production of job vacancy official indicator broken down by this characteristic is still not required as obligatory by the current EU Regulations, Eurostat already invites Member States to disseminate and transmit these data, which for some Member States are already available in the Eurostat online data warehouse. In the new EU Regulation under approval, the use of innovative sources (web scraping) to obtain information on labour shortages by occupation and region will be allowed.

We also analysed the occupations included in OJAs by ISCO-08 major occupational group in the period from the third quarter of 2019 to the second quarter of 2022 for Italy. The evidence emerging from the analysis gives information consistent with the period under analysis characterised by the pandemic crisis. Elementary occupations and professionals in the second quarter of 2020 show an acceleration in line with the introduction of the policy measure of the "Construction bonus" since June 2020. We observe a general contraction of all the occupations in 2021, due to the economic consequences of the strengthening of (second) lockdown measures, and in the second quarter of 2022. The analysis of this information could be carried out, of course, also on a monthly basis to refine the understanding of current trends.

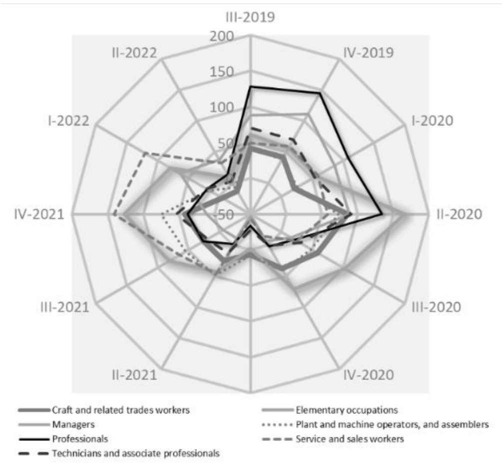


Fig. 3 Cedefop OJA data by ISCO-08 major occupational group, year on year changes, III 2019 - II 2022, Italy

4 Conclusions

We highlighted some possible uses of OJA data to support the current official production of job vacancy statistics, even if some quality aspects of OJAs need still to be improved [1]. Different activities aimed at improving the data quality are being developed within the Cedefop Data Lab project and the ongoing ESSnet Web Intelligence Network Project. Istat is also thinking about inserting a question in the current official JV survey questionnaire on the share of vacancies that pass through the online recruitment channel, to better understand the coverage of the OJAs and explain, at least in part, the differences in the estimates that are recorded at the sectoral and employment level between the two sources.

Regarding to the possible uses of the OJA data in official statistics in the short to medium term, we proposed experimental estimates of the monthly distribution of the job vacancy rate and the exploitation of qualitative data on the type of occupation. This latest information has a great level of detail and could be very useful to calculate skills rankings or to analyse the classifications of skills across occupations and the skill changes over time.

OJA data could also give the opportunity to focus on the dynamics of the unmet labor demand, providing this information by the main characteristics of both the job position (e.g., occupation and type of contract) and job requirements (e.g., education and skills), strategic information that does not currently exist in official statistics.

References

1. Alexandru C., Aprile D., Chianella D., Columbano A., Dumesnil de Maricourt C., Elezović S., Grahonja C., de Lazzer J., Lucarelli A., Maślankowski J., Necula M., Rengers M., Saucy F., Schmassmann S., Sorrentino M., Špeh T., Stateva G., Wu D.: Methodological framework for processing online job adverts data for Official Statistics V.2, In ESSnet Project on Big Data II - WPB Online job vacancies Deliverable B3 (2020)
2. Cedefop: Online job vacancies and skills analysis: a Cedefop pan-European approach (2019)
3. Beręsewicz M., Pater R.: Inferring job vacancies from online job advertisements. Statistical Working Papers, Eurostat, Luxembourg (2021)
4. Amarone M., Aprile D., Chianella D., Lucarelli A., Rocci F., Sorrentino M.: Supplementary indicators for official statistics from OJAs: the Italian case. ESSnet on Big Data II Project - WPB Online job vacancies national report (2020)
5. Branka J., Kvetan V., Napierala J.: From the online job advertisements to official statistics – the aspects of quality assurance. Q2022, Vilnius (2022)

Innovation in management: towards the open manager

Innovazione nel management: verso l'open manager

Paolo Bruttini, Michele Gallo, Paolo Mariani and Tullio Menini

Abstract This study focuses on the manager's professional work. In particular, the main focus is to detect the possible new approach in the managerial behavior able to define this professional figure and a first idea of 'open manager'. Kindness, empathy and sharing of objectives are characteristics that could revolutionize the figure of the leader. A transformation that moves away from the old models in favour of a horizontal and participatory organization of power. For this reason, the successful leader is able to interact the human dimension of employees and guide them towards a shared goal.

Abstract *Questo studio si concentra sul lavoro professionale del manager. In particolare, l'obiettivo principale è quello di individuare il possibile nuovo approccio nel comportamento manageriale in grado di definire questa figura professionale e una prima immagine di "manager aperto". Gentilezza, empatia e condivisione degli obiettivi sono caratteristiche che potrebbero rivoluzionare la figura del leader. Una trasformazione che allontana i vecchi modelli a favore di un'organizzazione orizzontale e partecipativa del potere. Per questo motivo, il leader di successo è in grado di intercettare la dimensione umana dei collaboratori e di guidarli verso un obiettivo condiviso.*

Key words: Labour Market, Open Manager, Rasch Analysis

Paolo Bruttini
Università degli Studi di Udine; e-mail: bruttini.paolo@spes.uniud.it

Michele Gallo
Università degli Studi di Napoli L'Orientale; e-mail: mgallo@unior.it

Paolo Mariani
Università degli Studi di Milano-Bicocca; e-mail: paolo.mariani@unimib.it

Tullio Menini
Università degli Studi di Napoli L'Orientale; e-mail: menini@unior.it

1 Scenario

The great wave of globalization has produced a profound impact on organizations from many points of view. A change as essential as ever to enable those who lead a business today to cope with the difficulties, rethinking management in entirely new ways and logics. Moreover, with the new century, one of the most interesting strategic perspectives in strategic and industrial development research has been developing, namely the phenomenon called Open Innovation [3]. Indeed, it has been realized that there can be an open way of developing innovation, through connections and collaborations with research centers, professionals and companies outside the organization, with a view to creating a mutually beneficial alliance.

In recent years, it is also applied to the enterprise in its complexity, to relationships with employees and to the way management interprets its role [2]. Thus, Open Organization is understood as a complex of practices that can be traced back to organizational models, systems of work team functioning and managerial behaviors that seek to provide a concrete response to the need for companies to adapt quickly and evolve, based on market needs (customers and tensions with competitors). It is a fragmented, multifaceted, largely unstructured movement [4] that identifies itself in various "buzz words" such as heterarchy, agility, teal and openness precisely.

2 Data description

Data was collected by Fondirigenti and Confindustria in 2020 through a structured questionnaire distributed to a set of Italian companies and filled in by a managerial internal figure. The total number that responded was equal to 383 managers (non-probabilistic sample). The questionnaire was made up of two sections:

- in the first section there were questions concerning the context in which the firms operate, such as economic sector, dimension, geographical area, as well as the main social demographic characteristics of managers, such as gender, age, education level, respectively.
- in the second section there were thirty items describing the managers' business behaviors and attitudes, useful for defining the concept of 'openness' characterizing the figure of the open manager. So, in the second section of the questionnaire, there are 30 different statements. Items were formulated as a 4point Likert scale, with responses ranging from 1 to 4 where 1 stands for "totally disagree" and 4 stands for "totally agree".

3 Methodology

The Rasch dichotomous model can be applied wherever discrete data is obtained with the intention of measuring an attribute or qualitative trait. The Rasch dichotomous model specifies the probability, P, that person *n* of ability B_n succeeds on item *i* of difficulty D_i. "Success" means "exhibits more of our intended latent variable. "Failure" means "exhibiting less of our intended variable". So, we must score the observations in accordance with this intention, no matter what values are assigned to the observation during data collection. P is the probability of success, and 1- P is the probability of failure. Success or failure must always happen, when we add their probabilities, they must sum to 1. In other words, success is a score of "1", and failure is a score of "0" on an item. Then the Rasch dichotomous model specifies the probability P_{ni1}, of the person *n* of ability B_n scores 1 on item *i* of difficulty D_i while with P_{ni0} the probability of scoring 0. For "qualitatively ordered data" "Success" means "more of what we are looking for" "Failure" means "less of what we are looking for". The difference between "Success" and "Failure" is qualitative. The ordering of these different qualities is indicated by scoring them "1" and "0". "1" "indicates more of the latent variable". "0" "indicates less of the latent variable". In the Rasch model, the probability of a correct answer is modeled as a logistic function of the difference between the person and item parameter. We have already talked about dichotomous data, such as "Right or Wrong". "Dichotomous" means "two cuts" in Greek. In performance assessment and attitude surveys, we encounter rating scales, such as the first "none, some, plenty, and all" and the second "strongly disagree, disagree, agree, and strongly agree" [1,5].

4 Results

After processing the fit statistics report how well the data corresponded to the measure estimates. The results of the analysis are summarized and reported in the table 1. Reliability statistics report the reproducibility of the measures.
The concept of reliability is defined by the ratio we now express as:

Reliability = True Variance / Observed Variance.

Kuder-Richardson KR-20, Cronbach Alpha, etc. are all estimates of this ratio.
They are estimates because we can't know the "true" variance, we must infer it in some way. In Rasch situations, we also have an item reliability.
This reports how reproducible the item difficulty order is for the set of items for this sample of people.
OUTFIT means "Outlier-sensitive fit statistic". OUTFIT is a conventional Pearson chi-square fit statistic divided by its degrees of freedom. This is more sensitive to unexpected remarks by people on items that are relatively very easy or very difficult for them.
The INFIT mean-square is the information-weighted average of the squared residuals.

INFIT means “Inlier-pattern-sensitive fit statistic”, or more technically, “Information-weighted fit statistic”. This is more sensitive to unexpected patterns of people's observations of items that are roughly targeted at them (and vice versa).

Table 1 Summary statistics

SUMMARY OF 383 MEASURED PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	99.4	30	1.46	0.3	1.03	0.1	0.99	-0.1
S.D.	7.2	0	0.62	0.04	0.37	1.3	0.35	1.2
MAX.	117	30	3.88	0.6	2.68	4.4	2.93	4.6
MIN.	73	30	-0.3	0.23	0.32	-3.6	0.31	-3.6

CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = 0.76

SUMMARY OF 30 MEASURED ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	1269	383	0	0.08	1	0.1	0.99	0
S.D.	119.3	0	0.59	0.01	0.12	1.6	0.15	1.9
MAX.	1457	383	1.42	0.12	1.33	4.6	1.33	4.6
MIN.	958	383	-0.99	0.06	0.88	-1.6	0.81	-2.2

ITEM RELIABILITY = 0.98

The parameters of the model characterize the competence of the interviewees and the difficulty of the items as collocations on a continuous latent variable. The proposed representation of the results allows us to have at the same time the measure of the behaviors considered prevalent in the definition of OM, and of the adherence of the managers interviewed to these behaviors. In the variable map (Fig.1) it is possible to notice three groups of items: the uppermost ones, circled in red, which are distant from the others represent the statements less condivisible by the OM (the most difficult).

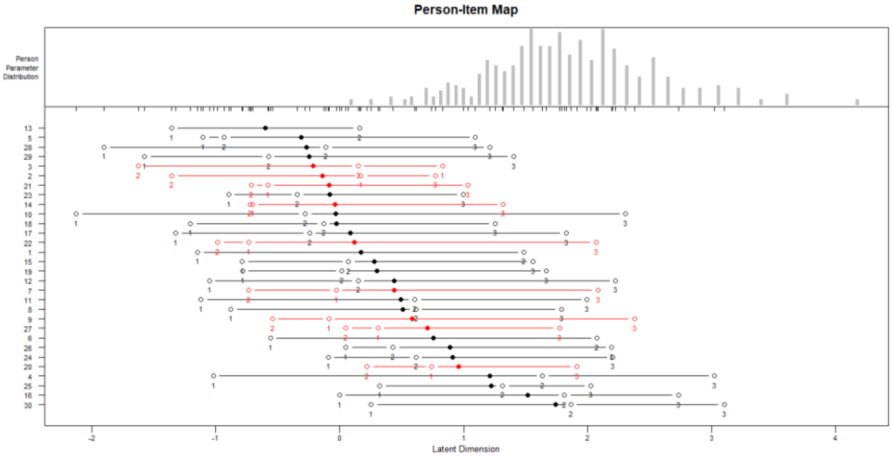


Fig. 1 Person and item parameters (items 1-30)

There are several diagnostics for the analysis of the individual items. By way of example, the measures of the thresholds between categories and curves are calculated. The baseline value of 1.0 for INFIT and OUTFIT is very close to the value in the table (Fig.2).

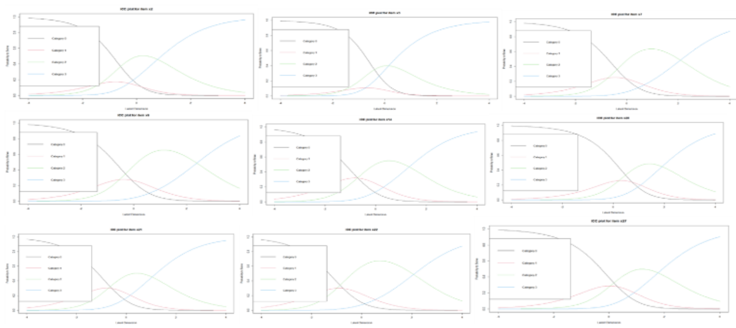


Fig. 2 Category probabilities - item x2, x3, x7, x9, x14, x20, x21, x22 e x27. Black (0): "Strongly disagree"; red (1): "Partially disagree"; green (2): "Partially agree"; blue (3): "Strongly agree"

5 Conclusions and Further extension

The results of the analysis can be used to predict how likely the manager will be to be open in his activities, based on the pattern of responses to the questionnaire. The guidelines for the managerial staff in the selection phase are useful for a fruitful collaboration or for calibrating training interventions on certain aspects of possible improvement of the company management. Using the data collected in the first part of the questionnaire, it is possible to compare groups of subjects with different personal characteristics. The study lends itself to further analysis and application in management and training.

References

1. Bond, T., Yan, Z., and Heene, M.: Applying the Rasch model: Fundamental measurement in the human sciences", Routledge (2020)
2. Bruttini, P.: Città dei capi. Milano. IPSOA (2014)
3. Chesbrough, H.: The logic of open innovation: managing intellectual property. California management review, 45(3), 33-58 (2003)
4. Laloux, F., Poireaux, G. N., & Blanchard, P.: Reinventing Organizations-Vers des communautés de travail inspirées. Diateino (2015)
5. Rasch, G.: Probabilistic models for some intelligence and attainmenttests, MESA Press (1993)

Solicited Session SS7 - *Data analysis for web sources*

Organizer and Chair: Andrea Marletta

1. *Enhancing SMEs default prediction with web-scraped data* (Crosato L., Domenech J. and Liberati C.)
2. *Web data as enabler for informed decisions in Labour Market* (Maggioni G.)
3. *The metaverse & luxury fashion brands: strategic communication exercise* (Forciniti A. and Zavarrone E.)
4. *Increasing the Geographical Granularity of Economic Indicators with Google Trends* (Domenech J. and Marletta A.)

Enhancing SMEs default prediction with web-scraped data.

Migliorare la previsione del default per le PMI con dati web-scraped

Lisa Crosato, Josep Domenech and Caterina Liberati

Abstract The objective of a Credit Risk model is to develop an accurate rule that can distinguish between good and bad instances. In this work instances are firms, specifically we focus on Small and Medium Enterprises (SMEs), as most of the papers in the financial literature, thanks to their contribution to the European Union economy for both value added and the creation of jobs. The aim of the paper is twofold: to explore the usage of the web-scraped indicators to improve firms' default predictions and to compare the models' performances fed with the web-based indicators with the ones built with the standard financial information. The results, obtained on a sample of Spanish companies, will also include a focus about the ability of the web features in recognizing uncommon defaulted firms.
Abstract *L'obiettivo di un modello di rischio di credito è quello di sviluppare una regola accurata in grado di distinguere tra osservazioni buone e cattive. In questo lavoro le aziende sono le nostre osservazioni, in particolare ci concentriamo sulle Piccole e Medie Imprese (PMI), come la maggior parte dei lavori nella letteratura finanziaria, dato il loro contributo all'economia dell'Unione Europea sia per il valore aggiunto sia per la creazione di posti di lavoro. Lo scopo dell'articolo è duplice: esplorare l'uso degli indicatori ottenuti dal web per migliorare le previsioni di insolvenza delle imprese e confrontare le prestazioni dei modelli alimentati con tali indicatori rispetto a quelli costruiti con le informazioni finanziarie standard. I risultati, ottenuti su un campione di aziende spagnole, includeranno anche un focus sulla capacità degli indicatori web di riconoscere le aziende insolventi non comuni.*

Key words: Credit Risk Modeling, Default Prediction, SMEs, Web-scraped Data, Balance sheets

---

Lisa Crosato
Department of Economics, Ca' Foscari University of Venice
e-mail: lisa.crosato@unive.it

Joseph Domenech
Department of Economics and Social Sciences, Universitat Politecnica de Valencia
e-mail: jdomenech@upvnet.upv.es

Caterina Liberati
Department of Economics, Management and Statistics, University of Milano-Bicocca
e-mail: caterina.liberati@unimib.it

1 Introduction

The objective of a Credit Risk model is to develop an accurate rule that can distinguish between good and bad firms [3] when they apply to get financial support from the banks. The task is particularly challenging when Small and Medium Enterprises (SMEs) are involved. In the last twenty years, SMEs is the firms segment who has received the greatest attention in financial prediction studies, because it represents 99% of all EU private businesses, and it generates more than half of the EU jobs [8]. The main issue of bankruptcy prediction on SMEs is the information employed to evaluate companies. Most models are generally run on indicators derived from firms' balance sheets, because they detail all the firms' economic and financial structure [1, 9, 6, 2]. Unfortunately, it is also known that SMEs balance sheets may show opacities that complicate the assessment of creditworthiness. Moreover, the access to these data is not free, requiring a fee subscription by data-providers as Bureau van Dijk, and are available for statistical purposes with a 18-24 month delay with respect to the reference period. Accordingly, studies in this field are shifting the interest on new sources of information to complement the standard indicators.

In our work we propose the usage of a new set of indicators based on the information extracted from firms websites. This approach, proposed by Crosato et.al. [7] allows to collect firms updated data that are freely available to all the users, although the preparation process to retrieve, storage and clean them requires specific skills in data quality assessment. We rely our study on a sample of about 900 Spanish companies extracted from the SABI -Sistema de Análisis de Balances Ibéricos (Bureau van Dijk)- and carry out the analysis using both standard accounting indicators and new web-scraped indicators. Performance results of the estimated models will be illustrated. We will also highlight a characterization of the two groups of firms in terms of standard/web information.

2 Research setting and data modeling

As illustrated in Section 1, the data of our study collects two sets of indicators: the ones derived from the accounting balance sheets (offline data) and the ones computed from the firms' website (online data). The offline data are easily obtained from the SABI platform, while the online data are built from the scratch with a three step procedure, described as follows. First, we downloaded the firms' homepages via the Wayback Machine[1] and we extracted HTML code, then we built a first set of indicators just registering the presence/absence of specific tags that generally describe the layout, the technology or the complexity of the website. Finally, we obtained a second set of online indicators by employing Natural Language Processing (NLP)

[1] The Wayback Machine is a free digital library of Internet sites. It allows to see how websites looked in the past even if they are defunct web pages not available anymore.

techniques on sites' contents, to identify the most frequent tokens or stems and their presence/absence in the texts of each firm website.

The offline data collects the variables *Number of Employees, Debt amount, Economic Profit* and *Productivity* referring to 2013. The online data consist of 50 dummies[2] referring to 2014. The reference time in the offline/online predictors is different due to the availability of the information that delays the access on the balance sheets. The target variable of our study indicates the status of the firm: defaulted ($y = 1$) or alive ($y = 0$) in 2015.

Credit risk literature is full of successful examples of Machine Learning methods applied to predict firms bankruptcy. Baesens et.al. [3] provided a comprehensive study about performances and reliability of several algorithms comparing standard statistical models as Logistic Regression and Discriminant Analysis with Support Vector Machines (SVM) and k-Nearest Neighbors. In our study we applied Least-Square Support Vector Machines (LS-SVM), introduced by [11], because it simplifies the solution of the convex problem in the SVM formulation without losing the classification capability. We estimated several solutions based on different kernel maps (Tab. 1) using a grid search to set the width value, c, whereas the degree of the Polynomial kernel is fixed to $d = 2,3$.

Table 1 Kernel functions

Kernel Mapping	k(x,z)
Cauchy (CAU)	$\frac{1}{1+\frac{\|\mathbf{x}-\mathbf{z}\|^2}{c}}$
Laplace (LAP)	$\exp(-\sqrt{\frac{\|\mathbf{x}-\mathbf{z}\|^2}{c^2}})$
Multi-Quadric (MULTIQ)	$\sqrt{\|\mathbf{x}-\mathbf{z}\|^2+c^2}$
Polynomial (POLY)	$(\mathbf{x}\cdot\mathbf{z})^d$
Gaussian (RBF)	$\exp(\frac{-\|\mathbf{x}-\mathbf{z}\|^2}{2c^2})$

The performances of the LS-SVM classifiers will be compared to Logistic Regression, that is considered the benchmark of the industry applications and to the Random Forests [5] that are indicated as the reference model for supervised classifications in credit risk [10].

3 Upcoming Results

Preliminary analyses of the data highlight interesting results on web-scraped variables in predicting SMEs default: whatever model we apply, the Area Under the ROC Curve is higher than 0.5, with a maximum value of 0.71. The models trained

[2] We reduced the 50 dummies to quantitative factors by means of a Multiple Correspondence Analysis [4].

on offline indicators show better classification performances: in the best case the AUC reaches a value equal to 0.847. Combining the online and offline sets of indicators, we obtain even better outputs, increasing the level of AUC up to 0.855. Detailed illustration of the results will be provided in the oral presentation.

Acknowledgements This work was partially supported by the Ca' Foscari University of Venice, Italy and by grant PID2019-107765RB-I00 (funded by MCIN/AEI/10.13039/501100011033).

References

1. Altman, E.I., Sabato, G.: Modelling credit risk for SMEs: Evidence from the US market. Abacus 43, 332–357 (2007)
2. Andreeva, G., Calabrese, R., Osmetti, S.A.: A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. European Journal of Operational Research 249, 506–516 (2016)
3. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54, 627–635 (2003)
4. Benzécri, J.P.: Sur l'analyse des tableaux binaires associés à une correspondance multiple. Les Cahiers de l'Analyse des Données 2, 55–71 (1977)
5. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
6. Ciampi, F.: Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of Italian firms. Journal of Business Research 68, 1012–1025 (2015)
7. Crosato, L., Domenech, J., Liberati, C.: Predicting SME's default: Are their websites informative? Economics Letters , 109,888 (2021)
8. European Commission: Annual Report on European SMEs 2018/2019. Tech. rep. (2019)
9. Fantazzini, D., Figini, S.: Default forecasting for Small-Medium Enterprises: Does heterogeneity matter? International Journal of Risk Assessment and Management 11, 138–163 (2009)
10. Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247, 124–136 (2015)
11. Suykens, J., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9, 293–300 (1999)

Web data as enabler for informed decisions in Labour Market

Il dato da Web come abilitatore per decisioni informate nel mercato del lavoro

Gabriele Maggioni

Abstract The contribution shows how within the context of Labour Market, particularly in The Adecco Group, date retrieved from the Web have a highly significant impact. In this category of information reside data coming from Google Trends, Job Crawling tools, public information related to companies registered data and indications shared by institutional entities. These data help market interpretation and are a crucial support to take data driven informed decisions.
Abstract *Il contributo presenta come all'interno del contesto del mercato del lavoro, e in particolare in The Adecco Group, i dati recuperati dal web abbiano un impatto estremamente rilevante. In questa categoria di informazioni rientrano i dati provenienti da Google Trends, strumenti di Job Crawling, informazioni pubbliche relative a dati camerali delle aziende e indicazioni condivise da enti istituzionali. Queste informazioni aiutano la lettura del mercato e sono un supporto cruciale per la presa di decisioni informate e basate sui dati*

Key words: Labour Market, Web data, Google Trends, Job Crawling, data driven decisions

1 Web Data for Labour Market

Labour Market is continuously evolving, especially in these years, where regulatory and exogenous elements have strongly affected it. For this reason and to have the most

---

Gabriele Maggioni
Head of Data Science & Research, The Adecco Group; e-mail: gabriele.maggioni@adeccogroup.com

accurate and detailed visibility of actual and future scenarios of this complex context, data is crucial.

There is a heterogenous and wide set of data that can be retrieved on the web and can help from this point of view.

1.1 Google Trends

Google has become in the latest 15 years the first access point of the Web for the large majority of Internet users; also, a big subset of these users is using the Web to purchase or to search-to-purchase, becoming consumers, therefore directly or indirectly the reason for an item/service to be requested. Adecco, as a Job Agency, helps companies dealing (for example) with peak requests or with talent scarcity, being therefore affected at some point by consumers interest trends.
Thanks to a detailed monitoring of user interest (through selected keywords grouped by company sectors), Adecco is able to have almost real-time full visibility of emerging trend, also with a forecast in the short term.

1.2 Job Crawling

One of the most recurring questions in Labour Market is "How is the demand evolving over time?". This point includes a large set of sub-topics regarding for example locations, type of jobs, type of profiles. A potential way to indirectly track these KPIs is to monitor Job Posts: knowing what has been publicly exposed as a demand is a very valuable proxy of the real evolution of the market.
Job Crawling tools are able to automatically inspect web pages where job ads are posted and summarize them, while also reconducting them to fixed taxonomies, for example Job Categories and Skills (this is where the real value of these tools is evident).

1.3 Public and Institutional Information

The third section of interesting information includes those falling in the following categories:

- Institutional data (i.e. from ISTAT)
  These include data on population and its evolution over time according to several parameters, on country official KPIs, like GDP (Gross Domestic Product), Unemployment rate, on financial information, like revenues by sector;

- Company data (i.e. from “Camera di Commercio”)
  Revenues, number of employees, number of temporary contracts, industrial sectors are few of the available information;

- Staffing Market data (i.e. from Ebitemp or Formatemp)
  Thanks to the data that are mandatory to be sent to these institutions and subsequently broadcasted in aggregated form by these institutions themselves, figures on the market are available for analyses and to enable strategies.

2 Use Cases

The wide amount and variety of available data enable for The Adecco Group a relevant set of use cases, that leverage the depth and complexity of the information to build complex solutions for the company.

Below you can see three of these tools as an example.

2.1 Adecco Trend Catcher

A few years ago, Adecco had the idea to leverage Google Trends data to predict user interests; to reach this goal, a set of keywords has been declined by industry, defining baskets of research strings conceptually representing topics/businesses. Then these by-industry baskets have been correlated with Adecco business data and shifted, to obtain correlations in the past enabling predictions of future impacts on the business.

2.2 Job Demand Prediction

Leveraging Job Crawling data, enriched with public data on Employment and population, it is possible to have a reasonably accurate prediction of profiles searched by companies; this can be obtained mainly projecting past job ads information (from job crawling), properly weighting them by province and considering unemployment and population trends. These forecasted insights can help Adecco operations and branches to proactively define the best delivery and candidate sourcing strategies, declined by single area and type of profile.

2.3 Targeted Business Development

Staffing and company data we mentioned previously, if properly managed and aggregated, can help having a precise enough picture of the market. Staffing data lead to a detailed (even if delayed) outline of how well Adecco is positioned in the market, by area and by industrial sector. Leveraging then updated information of companies (including the data of which ones are growing the most) Adecco sales people are supported by receiving concrete indications on which client/prospect to address with priority. This data-driven business development approach has a really high effectiveness, as it allows to visit companies exactly when they might need to work with a job agency.

The metaverse & luxury fashion brands: strategic communication exercise

Il metaverso & i brand di lusso della moda: esercizio di comunicazione strategica

Alessia Forciniti and Emma Zavarrone

Abstract The paper analyzes the luxury fashion brands' strategic communication on Twitter for the metaverse fashion weeks (MVFWs) of 2022 and 2023. A three-step methodological approach has been developed: 1) textual analysis; 2) bipartite graph clustering to detect patterns of communication among brands; 3) word vectors representations for investigating linguistic regularities. The results showed that MVFW 2022 was based on a shared narrative, probably because of its innovative nature, while the event 2023 has brought out a heterogeneous and diversified communication where each brand is a cluster. Also, the linguistic regularities allowed us to observe different semantic sub-structures over time, and a mature introjection of the phenomenon in 2023.

Abstract *Il paper analizza la comunicazione strategica su Twitter dei brand della moda di lusso per le metaverse fashion weeks (MVFWs) del 2022 e 2023. Un approccio metodologico a tre fasi è stato sviluppato: 1) analisi testuale; 2) clustering di grafo bipartito per rilevare modelli di comunicazione tra i brand; 3) rappresentazioni di vettori di parole per indagare regolarità linguistiche. I risultati hanno mostrato che la MVFW 2022 era basata su una narrativa condivisa, probabilmente per la sua natura innovativa, mentre l'evento 2023 ha messo in luce una comunicazione eterogenea e diversificata dove ogni brand è un cluster. Anche le regolarità linguistiche hanno consentito di osservare diverse sub-strutture semantiche nel tempo, ed una maturata introiezione del fenomeno nel 2023.*

Key words: metaverse fashion week, luxury brands, network analysis, word vectors representations, word embeddings

Alessia Forciniti
Iulm University, Via Carlo Bo, e-mail: alessia.forciniti@iulm.it

Emma Zavarrone
Iulm University, Via Carlo Bo, e-mail: emma.zavarrone@iulm.it

1 Introduction

The technology revolution 4.0 deeply influenced business models by developing new forms of marketing, buying-selling, and decentralized systems, where ownership is shared amongst users; there is free access; the monetization is without infrastructures (banks and third parties); the users have a digital identity. In the last few years, the worldwide attention was captured by an 'integrated immersive ecosystem' called *metaverse* which operates through systems of Artificial Intelligence (AI), Virtual Reality (VR), Augmented Reality (AR), and virtual economies based on blockchain and cryptocurrency (NFTs), also in sectors of the creative industry, which is making a natural transition toward it. In particular, after the pandemic of Covid - 19, the fashion industry is undergoing an expedited digital transition [6] by introducing 3D fashion, virtual models, online fashion weeks [8], and by approaching the metaverse both for buying-selling and marketing, and for events such as the *metaverse fashion week* (MVFW). Fashion Week is the most expected two-year fashion industry event, where designers and brands display their latest collections for the Spring/Summer or Fall/Winter seasons. Since the first fashion week of New York in 1943, the event has always been held in the world's fashion capitals, New York, London, Milan, Paris, but in Spring 2022, a collection dedicated to the metaverse was launched. The inaugural event, held from March 24th to 27th, 2022, on Decentraland - a digitalized microcosm - recorded 18,000 users and several luxury brands. In 2023, the MVFW returned from March 28th to 31st, with 165,000 wearables.

The goal of our contribute is to study the strategic communication of luxury fashion brands to advertise their participation to MVFW. We explored two research questions (RQs).
RQ_1: Can we detect patterns of strategic communication among fashion brands during the MVFWs 2022 and 2023?
RQ_2: Are there linguistic regularities in communicating the MVFWs?

2 Textual data collection and Methodology

2.1 Textual data collection

Social media communication has been analyzed as the best perspective for exploring topics and events of the virtual community [9], and more properly Twitter because of its popularity and data extraction facilitated by Twitter *API (application programming interface)*.
Data collection consists of the population of tweets published on the official accounts by the main companies anticipated by Decentraland for the MVFW 2023. We selected only luxury brands' tweets that repeated their participation in MVFW 2022 and that 2023: Clarks Shoes, Diesel, Donna Karen New York (DKNY), Dolce

& Gabbana, Monnier Freres, Tommy Hilfiger, Vogue Singapore magazine.
The download was based on two slots: 1) from March 1st to 31st, 2022 to investigate the MVFW 2022 that took place from March 24th to 27th, 2022; 2) from March 1st to 31st, 2023 to study the last event proposed from March 28th to 31st, 2023. Data extraction returned 1,057 tweets, which were filtered by keywords *"metaverse fashion week", "MVFW","MFW"*. The filtered corpus is 144 tweets, of which 47 in 2022 and 97 in 2023. The lexical diversity calculated with one of the most used measures, the *type-token ratio (TTR)*, shows smaller lexical richness in the event 2023 (TTR = 22.63%) than opening event (TTR = 39.21%), with a difference of +16.58% in 2022. For details concerning the corpus of each brand refers to Table 1.

Table 1 Lexical statistics on brands' corpora 2022 and 2023

	2022			2023		
Brand	Types	Tokens	TTR	Types	Tokens	TTR
ClarksShoes	22	35	62.86%	43	51	84.31%
Diesel	49	54	90.74%	73	174	41.95%
DKNY	22	46	47.83%	15	43	34.88%
Dolce & Gabbana	80	99	80.80%	186	140	49.46%
Tommy Hilfiger	28	60	46.67%	232	1092	21.24%
Vogue Singapore	166	642	25.86%	232	1482	15.58%

2.2 Methods

Our methodological flow is composed of three steps which were replicated separately on both corpora 2022 and 2023:

1. *Textual analysis*
2. *Social Network Analysis & Graph clustering*
3. *Word vectors representation for linguistic regularities*

To answer RQ_1, we combined textual analysis to network analysis and graph clustering.

1. *Textual analysis*: The pre-treatment was performed by normalization, lemmatization, removal of emojis, URLs, hashtags, usernames, special characters of Web and stop words for the English language based on Smart system [4]. We used the coding of *bag-of-words* for feature extraction and built the lexical table $\mathbf{T}_{ixj}$ which records the frequency, n_i, of the term i^{th}, where $i = 1,..., p,$ in the brand's corpus j^{th}, where $j = 1,..., q$, where each frequency represents the weight of relationship between brand and word. The lexical tables of two slots are indicated such as **A** and **B**.

2. *Network Analysis & Graph clustering*: The matrices **A** and **B** may be represented through the social network properties [11]. Specifically, by a weighted bipartite graph G_B (V_1, V_2, E), where V_1 and V_2 are the nodes corresponding to two classes of entities, V_1 for the class of words and V_2 for the class of brands, and E is the frequency-based linkage which unites V_1 to V_2 [1]. To look for communication patterns, an agglomerative hierarchical clustering approach was used, which takes shape through groups of nodes called communities [5]. We adopted a well-known clustering algorithm considered the latest, fastest, and most computationally efficient among others for detection, the Leiden algorithm [10]. It efficiently divides the entire network into smaller clusters of nodes, performs well for networks from small to large, and may be effective for bipartite graphs [3]. It is an improvement of the Louvain algorithm by means of three steps: 1) the modularity optimization *(Q)* [5] in an empirical range [0.3;0.7] ; 2) the refinement of partitions; and 3) the community aggregation [11]. Its effectiveness is demonstrated for real-world networks and for textual data (e.g., [2]).

3. *Word vectors representation for linguistic regularities* was adopted to answer RQ_2. We used a word embedding approach through an unsupervised learning algorithm called *GloVe* [7] which is able to capture semantic and syntax regularities using vector arithmetic with a performance of 75%, by combining the global matrix factorization and local context window methods. Its training is performed on elements other than zero in a *co-occurrence 'word x word' matrix* which tabulates how frequently words co-occur with one another in a corpus, rather than on the entire sparse matrix or on individual context windows. GloVe is a log-bilinear model with a weighted least-squares objective whose dot product equals the logarithm of the words' co-occurrence probability. These ratios of co-occurrence probabilities have the potential for encoding forms of meaning such as linear sub-structures of the word vector space. The substructures are captured by the similarity metric of cosine generated for closest neighbor assessments result in a single scalar that measures how closely two words are connected. The model is outlined as the vector difference between two word vectors (e.g., Paris - France + Germany = Berlin) in order to capture as much as possible the meaning specified by the juxtaposition of two words.

3 Results and discussion

The textual analysis returned two lexical tables: for 2022, $\mathbf{A}_{126x6}$ and for 2023, $\mathbf{B}_{260x6}$ denoting more terms for 2023 in the investigation of the same six brands.

The Fig. 1 answers RQ_1 and shows the bipartite graph clusterings 2022 and 2023.

The partitioning of MVFW 2022 detected four clusters with Q equals 0.57. We can see Clark Shoes as an isolated node, and a cluster that groups Diesel, DKNY,

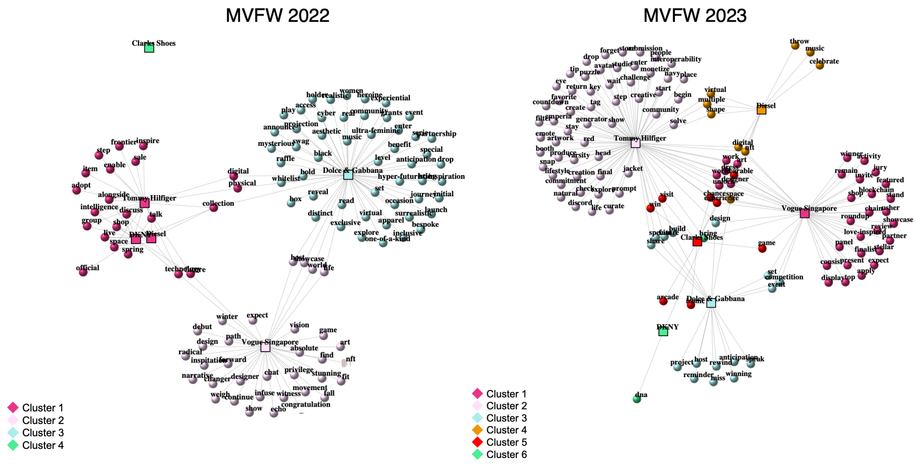


Fig. 1 Graph clustering: communities MVFWs 2022 and 2023 with most frequent 100 terms

and Tommy Hilfiger that share futuristic and technological aspects of the dynamics of the metaverse, the combination of physical and virtual. Dolce & Gabbana is the most numerous cluster and its communication inaugurates in the hyper-futuristic cyber world a collection dedicated to femininity that, through aesthetics, highlights its heroism. Vogue Singapore talks of inspiring and radical narrative based on gaming, art, in which fashion fits the world of NFTs. The MVFW 2023 shows Q equal to 0.49 and six clusters. There are no isolated nodes and brands that share the same grouping. Each brand identifies a cluster. Vogue Singapore's communication is love-inspired "spectacularization" and monetization. Dolce & Gabbana shows little disclosure compared to the opening event, with themes related to futuristic design and competition. Clark Shoes identifies itself through design and experience, while, Diesel celebrates the virtual and wearable NFTs. Tommy Hilfiger - who is the largest cluster - communicates future, design, creativity, colors, emotionality, interoperability, virtual community, and avatars.The event 2022, showed similar features within some brands, while that 2023 detected communication strategies that do not go in the same direction. Probably, this is due to the innovative nature of the inaugural event, which led brands to follow the same topics than 2023, where the event has been consolidated.

To detect the linguistic regularities in MVFWs 2022 and 2023 (RQ2), we proposed the same word vectors representations for both corpora. The purpose is twofold: 1) to test the use of linguistic or semantic regularities to communicate this type of events, and 2) to determine changes in regularities from 2022 to 2023. From the vocabularies of corpora 2022 and 2023, we selected seven lemmas with standardized degree centrality greater or equal to 0.55 that most characterize the nature of the event, namely: 'design', 'metaverse', 'nft', 'digital', 'future', 'experience', and 'technology'. From results, we observed, as in the case of graph clustering, the presence of different sub-structures between 2022 and 2023 (Tab.2), and the consolidation and maturity of the metaverse phenomenon in the fashion industry in 2023. More properly, design and cryptocurrency (NFT) in the first event expressed the expectation, while in 2023 indicated the experience and the collections. In 2022, digital design was synonymous of technological innovation and raffle, acquiring in 2023 a mastery of virtual fashion and the event through wearable models, challenges, and competition between brands and designers. The future connected to technology in the week 2022 sees the virtual experience as main feature of the linguistic regularity of strategic communication. On the contrary, in the MVFW 2023 the communication connotes the phenomenon as a natural acquisition of transformative context.

Table 2 Word vectors representation for linguistic regularities of MVFW 2022 and 2023

	2022		2023	
Word vectors difference	Sub-structure	Probability	Sub-structure	Probability
Design - metaverse + nft	occasion/expectation	0.55/0.50	collection/experience	0.66/0.61
Design - metaverse + digital	raffle/technology	0.68/0.65	wearable/challenge	0.63/0.60
Future - experience + technology	journey/cyber	0.61/0.58	host/natural	0.60/0.53

References

1. Borgatti, S.P: Two-Mode Concepts in Social Network Analysis. In: Meyers R. (eds.) Encyclopedia of Complexity and Systems Science. Springer, New York, NY (2009)
2. Boy, J.D. *et al.*: textnets: A Python package for text analysis with networks 5 (2020)
3. Kelly, S.T.: leiden: R implementation of the Leiden algorithm. R package version 0.4.3 (2022) Available via DIALOG. https://github.com/TomKellyGenetics/leiden Cited April 2023
4. Lewis, D. D., *et al.*: Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research 5, 361–397 (2004)
5. Newman, M. E. & Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E. 69 (2004)
6. Noris, A., Nobile, T., Kalbaska, N., Cantoni, L.: Digital Fashion: A Systematic Literature Review: A Perspective on Marketing and Communication. Journal of Global Fashion Marketing, 69, 32-–46 (2021)
7. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for Word Representation (2014) Available via DIALOG. https://nlp.stanford.edu/pubs/glove.pdf Cited April 2023
8. Ramadan, Z. and Nsouli, N.Z.: Luxury fashion start-up brands' digital strategies with female Gen Y in the Middle East. Journal of Fashion Marketing and Management, 26, 247–265 (2022)
9. Ribeiro-Navarrete, S., Saura, J.R., Palacios-Marqués, D.: Towards a new era of mass data collection: assessing pandemic surveillance technologies to preserve user privacy. Technol. Forecast. Soc. Change, 16 (2021)
10. Traag, V.A., Waltman, L. & van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep, 9 (2019)
11. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)

Increasing the Geographical Granularity of Economic Indicators with Google Trends

Aumentare la Granularità Geografica degli Indicatori Economici con Google Trends

Josep Domenech and Andrea Marletta

Abstract This paper proposes a method to compute a regional Economic Sentiment Index (ESI) using Google Trends (GT) data. The ESI is a leading indicator of macroeconomic magnitudes, and GT offers a high-frequency and real-time measure of information demand. The proposed method consists of retrieving the search volumes for terms potentially correlated with the ESI, selecting those highly correlated, conducting a principal component analysis (PCA), and estimating a linear regression model. The method was applied to the ESI in Spain, and the results showed that it is possible to estimate the ESI at a regional level using the PCA factor loadings of the national data.

Abstract *Questa ricerca propone un metodo per calcolare un Indice del sentimento economico regionale (ESI) utilizzando i dati di Google Trends (GT). L'ESI è un indicatore anticipatore del prodotto interno lordo e i dati GT offrono una misura ad alta frequenza e in tempo reale della domanda di informazioni. Il metodo proposto consiste nel recuperare i volumi di ricerca per i termini potenzialmente correlati con l'ESI, selezionare quelli altamente correlati, condurre un'analisi delle componenti principali (PCA) e stimare un modello di regressione lineare. Il metodo è stato applicato all'ESI in Spagna e i risultati hanno mostrato che è possibile stimare l'ESI a livello regionale utilizzando i coefficienti PCA dei dati nazionali.*

Key words: Economic sentiment, Google Trends, Non-traditional data sources.

Josep Domenech
Department of Economics and Social Sciences, Universitat Politècnica de València, e-mail: jdomenech@upvnet.upv.es

Andrea Marletta
Department of Economics, Management and Statistics, University of Milano-Bicocca, e-mail: andrea.marletta@unimib.it

1 Introduction

The Economic Sentiment Index (ESI) is a composite indicator representing the confidence of the different economic agents on the evolution of the economy. It is computed monthly based on regular surveys conducted by the Directorate-General for Economic and Financial Affairs (DG ECFIN) of the European Commission. Since the expectations of economic agents play an important role in consumer and producer decisions, the ESI is a leading indicator of other macroeconomic magnitudes such as GDP growth or industrial production [5, 4].

Despite its potential as a leading indicator, economic sentiment indicators at a regional level are less popular due to the limited resources of statistical offices. Recent research has shown that estimating economic sentiment with Internet searches is not only feasible, but it offers a number of advantages, such as real-time and high-frequency availability [6]. Time series on the popularity of Internet searches provided by Google Trends (GT) offer an aggregate measure of the demand for information [1] that correlates with a variety of economic magnitudes such as unemployment, inflation, touristic travel, car sales, etc.

Nowcasting economic indicators with GT usually departs from the selection of some keywords or search categories whose popularity could be related to the evolution of the economic variable of interest. Since the evolutions of the popularity of searches for different keywords are usually correlated, a common solution is to use a principal component analysis (PCA) to reduce the dimensionality and avoid multicollinearity issues [3, 7]. This way, common trends from the selected keywords and search categories are extracted. After that, a regression model is estimated to quantify the relation between the number of searches and the evolution of the economic indicator.

Since GT provides data in almost real-time, it is possible to have a flash estimate of the economic indicator under study. Not only this but the high frequency of GT data also enables the creation of daily or weekly estimates of indicators, such as unemployment or GDP, which are only released with a monthly or quarterly frequency.

While previous research works focused on the possibilities of GT for creating a high-frequency indicator of other underlying magnitudes, this paper proposes to use Google Trends Series to increase the geographical granularity of economic indicators, namely, the ESI.

2 Methods

The method for computing a regional ESI is represented in Figure 1. It departs from the estimation of the ESI at the national level based on the volume of searches in Google (as provided by GT), similarly to the process described by [3, 7]:

1. The search volumes for some terms potentially correlated with the ESI are retrieved and seasonally adjusted.

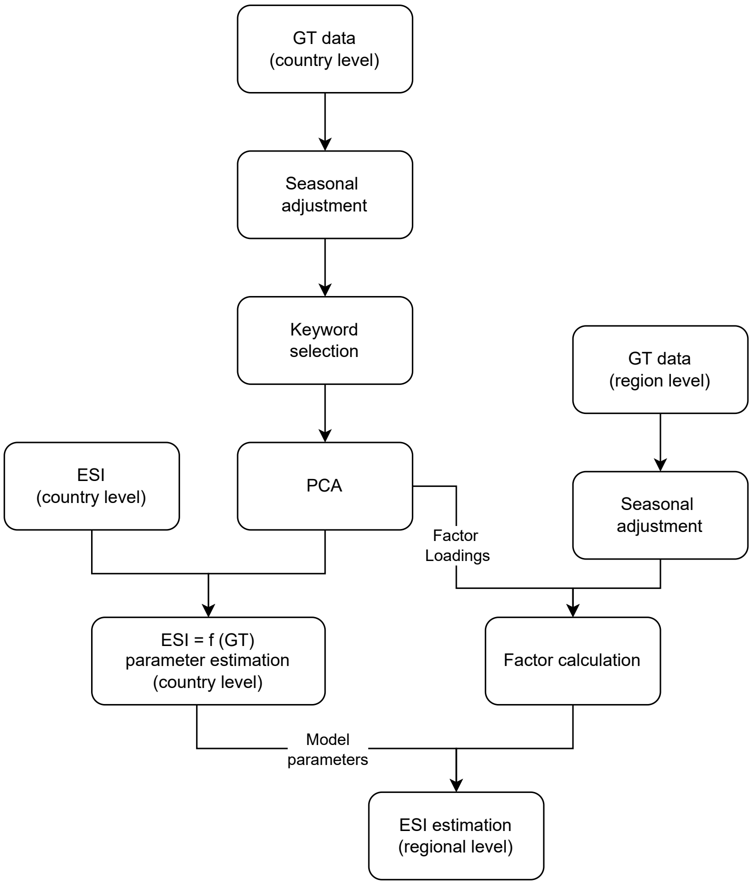


Fig. 1 Process for creating an Economic Sentiment Index at a regional level

2. Among all terms, only those highly correlated with the ESI are selected.
3. A principal component analysis (PCA) of the selected series is conducted to extract the common signal in them.
4. The first Principal Component is used to estimate the parameters of a linear regression model on the ESI (as provided by DG ECFIN).

After the model for the national ESI is estimated, the search volumes at the regional level for the same terms are retrieved from GT seasonally adjusted. Using the PCA factor loadings of the national data (step 3 above), factors at the regional level can be computed. Then, using the model parameters estimated in step 4 (above), it is possible to estimate the ESI at a regional level.

3 Results

The method described above has been applied to the ESI in Spain. A list of terms[1], comprising keywords and topics related to the economic sentiment, was requested to GT. Those with a Pearson's correlation coefficient higher than 0.7 were used as input of the PCA. The first principal component was used to estimate:

$$ESI_t = \beta_0 + \beta_1 GT_t + u_t.$$

Where ESI is the DG ECFIN's survey-based sentiment index and GT is the first principal component extracting the common signal of the selected search term volumes.

Figure 2 shows the cross-correlation between the first difference of the ESI and the first difference of its estimation based on search volume data. The highest cross-correlation value is in lag 1. This evidences that changes in the search behavior anticipate the changes in the sentiment index (as also reported by [3]).

The same set of search terms was requested to GT using a geographical filter to obtain the search volumes in each region of Spain. These series are combined using

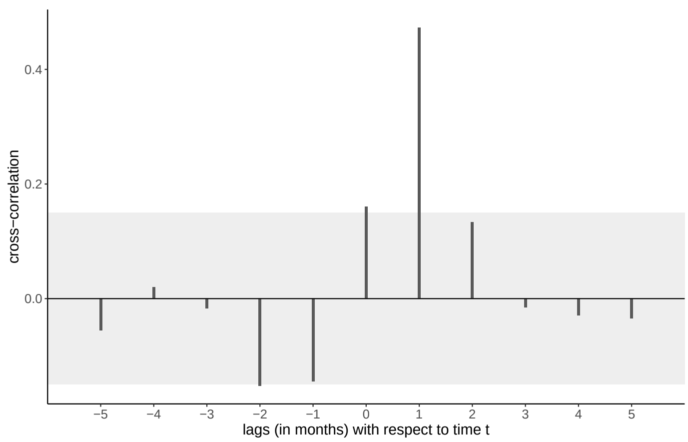


Fig. 2 Cross-correlation of the Economic Sentiment Index for Spain (DG ECFIN) and its estimation based on Google Trends series for t=Jan-2017 to Sep-2022. The grey area represents the 95% confidence interval.

[1] The complete list of terms includes the following keywords: crisis, quiebra (failure), infojobs, comprar (purchasing), desempleo (unemployment), LinkedIn, Idealista, construcción (building), emprendedor (entrepreneur), rebajas (sales), desempleo (unemployment), recesión (recession), paro (unemployment), bancarrota (bankrupcy), ahorrar (saving), trabajo a tiempo parcial (part-time job), hipoteca (mortage), cobrar paro (to receive the unemployment subsidy); and the following topics (topic ID): Gross Domestic Product (/m/039mk), saving (/m/0hbm6), business (/g/121jnq1m), economic crisis (/g/1211cg58), economy (/m/0gfps3), crisis (/m/02gyy_).

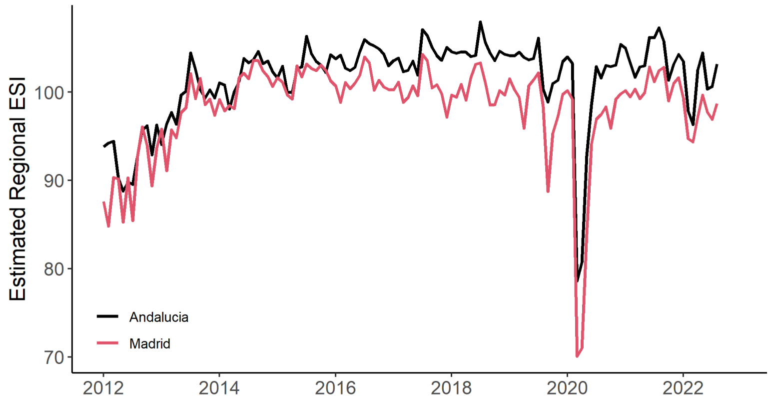


Fig. 3 Estimated Economic Sentiment Index for two regions in Spain.

the factor loadings resulting from the PCA at the national level. Finally, the ESI at the regional level is computed using the estimation of β_0 and β_1.

The created regional ESI series allow us to compare the evolution of the economic sentiment in different regions, as illustrated in Figure 3. As one can observe, the ESI in Andalucia and Madrid have diverged since 2016.

4 Conclusions

Google Trends is a rich source of information that offers high-resolution results both in terms of time-frequency and location. While the use of search volume data has long been used to increase the time granularity of economic indicators, this work has provided a procedure to extend also the geographic granularity.

It must be noted that the use of GT data is not exempt from limitations, such as the lack of accuracy [2], and time inconsistency [3]. Both are derived from the underlying sampling process and become more important in low-populated geographic areas.

Acknowledgements

This work was partially supported by grant PID2019-107765RB-I00, funded by MCIN/AEI/10.13039/501100011033.

References

1. Askitas, N., Zimmermann, K.F.: The internet as a data source for advancement in social sciences. International Journal of Manpower 36, 2 – 12 (2015). DOI 10.1108/IJM-02-2015-0029
2. Cebrián, E., Domenech, J.: Is Google Trends a quality data source? Applied Economics Letters , 1–5 (2022). DOI 10.1080/13504851.2021.2023088
3. Eichenauer, V.Z., Indergand, R., Martínez, I.Z., Sax, C.: Obtaining consistent time series from Google Trends. Economic Inquiry 60, 694–705 (2022). DOI https://doi.org/10.1111/ecin.13049
4. Ferreira, E., Martínez Serna, M.I., Navarro, E., Rubio, G.: Economic sentiment and yield spreads in Europe. European Financial Management 14, 206–221 (2008)
5. Gelper, S., Croux, C.: On the construction of the European economic sentiment indicator. Oxford Bulletin of Economics and Statistics 72, 47–62 (2010)
6. Van der Wielen, W., Barrios, S.: Economic sentiment during the COVID pandemic: Evidence from search behaviour in the EU. Journal of Economics and Business 115, 105,970 (2021)
7. Woo, J., Owen, A.L.: Forecasting private consumption with Google Trends data. Journal of Forecasting 38, 81–91 (2019). DOI https://doi.org/10.1002/for.2559

Solicited Session SS8 - *Methodological and applicative contributions for evaluating sustainable development*

Organizer and Chair: Ida Camminatiello
Discussant: Antonio Lucadamo

1. *Evaluating sustainable development in EU countries through synthetic indicators* (Alaimo L.S. and Cucci M.)
2. *Naples and tourism sustainability: A survey of citizens' perceptions* (Aria M., Pagliara F., D'Aniello L. and Della Corte V.)
3. *Modelling inequalities for sustainable development in Italy countries* (Musella M., Borrata G., Camminatiello I. and Lombardo R.)
4. *Food Security and Sustainability: A Science Mapping Analysis* (Piscitelli A.)

Evaluating sustainable development in EU countries through synthetic indicator

Valutare lo sviluppo sostenibile attraverso indici compositi: un'analisi nei paesi UE

Leonardo Salvatore Alaimo and Marianna Cucci

Abstract Sustainable development is a concept that has gained increasing popularity in recent years due to the increasing concern about the impacts of human activities on the environment. Parallel to the interest in this topic, the need of measuring it has also grown, providing appropriate frameworks and tools that can supply a simple and immediate reading of the phenomenon. The objective of this paper is to propose a synthetic index of sustainable development, starting from the framework proposed by Eurostat for measuring the sustainable development goals.
Abstract *Lo sviluppo sostenibile è un concetto che ha acquisito sempre più popolarità negli ultimi anni a causa della crescente preoccupazione per l'impatto delle attività umane sull'ambiente. Parallelamente all'interesse per questo tema, è cresciuta anche l'esigenza di misurarlo, fornendo quadri teorici e strumenti adeguati che possano dare una lettura semplice e immediata del fenomeno. L'obiettivo di questo lavoro è proporre un indice sintetico di sviluppo sostenibile, utlizzando il quadro di riferimento proposto da Eurostat per la misurazione degli obiettivi di sviluppo sostenibile.*

Key words: Sustainable development, EU countries, Synthetic indicators, Sustainable development index - SDI

Leonardo Salvatore Alaimo
Department of Social Sciences and Economics, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy, e-mail: leonardo.alaimo@uniroma1.it

Marianna Cucci
Department of Social Sciences and Economics, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy, e-mail: cucci.1813734@studenti.uniroma1.it

1 Introduction

Sustainable development is a concept that has gained increasing popularity in recent years mainly due to the increasing concern about human society, its future and the impacts of its activities on the environment [1]. Since its debut in the international debate in 1987 via the so-called Brundtland Commission and its report, *Our Common Future*, the term has not had a univocal definition in the literature, also because different disciplines have contributed to the sustainability debate. It can be defined as a *contested concept* [3]; it has an ambiguous and vague definition leading to different interpretations and meanings. The first and probably the best known and most accepted definition of sustainable development is that of *Our Common Future*: "sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs" [7, 41]. Sustainable development is not an aim in itself, but an instrument that must ensure the achievement of actual and future generations' needs. The only way to achieve sustainable development is to conceive it as a multidimensional concept consisting of three key pillars: economic, social, and environmental[1]. Nowadays, the definition focuses on this *holistic approach* [6]. Detecting trends in sustainable development is important to determine whether progress is being made towards achieving sustainable development goals and to identify areas that need improvement. It requires ongoing monitoring, evaluation, and improvement to ensure that we are on track to create a more sustainable future. The need to assess the sustainable development of societies has grown together with the importance of this issue in the international debate, in public opinion and among stakeholders. Several frameworks and indicators have been developed to assess its progress. The so-called *Agenda 2030*, adopted at the United Nations Sustainable Development Summit in September 2015, is one of the main framework. The Sustainable Development Goals form a part of the Agenda 2030: they are a framework of 17 goals and 169 targets selected to cover the three traditional dimensions of sustainable development. In this paper, we want to examine the EU countries situation with regard to the achievement of these objectives. Using the yearly time series of the indicators selected by Eurostat to measure the sustainable development in the EU countries in the period 2015-2021, we adopt a two steps procedure. Firstly, we construct a synthetic index for each dimension of sustainable development. Secondly, we define a synthetic index of sustainable development starting from the previous synthetic measures. The paper is organized as follows. Section 2 presents the indicators and the methods used. In Section 3 the application and the main findings are shown. Conclusions in Section 4 summarise the obtained results.

[1] Some authors have also criticised the three-way framework, estimating that the pillars that underpin sustainable development are more than three. For a review, please see: [1].

2 Data description and methods

All the indicators are selected from the Eurostat data-warehouse, in the Section on ”Sustainable Development Indicators”[2]. The choice of indicators has been based on two criteria: the availability of data for each EU country and in time series from 2015 to 2021. For this reason, it was not possible to include Goals 14 and 15 because none of the indicators in the Eurostat framework met the two criteria. Table 1 reports the indicators selected for each goal.

Table 1 Indicators of sustainable development for EU countries: goal, Eurostat code; indicator, polarity. Yearly time-series 2015-2021 (unless otherwise indicated).

Goal	Eurostat code	Indicator	Polarity
		Environmental dimension	
Goal 6	SDG_06_10*	Population having neither a bath, nor a shower, nor indoor flushing toilet	NEG
Goal 13	SDG_13_10*	Net greenhouse gas emissions	NEG
		Economic dimension	
Goal 7	SDG_07_40	Share of renewable energy in gross energy consumption	POS
Goal 8	SDG_08_30	Employment rate	POS
Goal 9	SDG_09_10	Gross domestic expenditure on R&D	POS
Goal 10	SDG_10_41	Income distribution	NEG
Goal 12	SDG_12_41	Circular material use rate	POS
		Social dimension	
Goal 1	SDG_01_10	People at risk of poverty or social exclusion	NEG
Goal 2	SDG_02_40	Area under farming	POS
Goal 3	SDG_03_11	Healthy life years at birth	POS
Goal 4	SDG_04_10	Early leavers from education and training	NEG
Goal 5	SDG_05_60	Positions held by women in senior management positions	POS
Goal 11	SDG_11_11	Severe house deprivation rate	NEG
Goal 16	SDG_16_20	Population reporting occurrence of crime, violence or vandalism	NEG
Goal 17	SDG_17_50	Share of environmental taxes in total tax revenues	NEG

*Data available up to 2010.

For the construction of the synthetic indices, we adopt the so-called aggregative-compensative approach, the dominant framework in the literature[3], which consists in the aggregation, by means of a mathematical function, of the elementary indicators. These methodologies are defined *composite* indicators [1]. The construction of a composite indicators is a step-by-step process. After the definition of the phenomenon and the selection of the basic indicators, we must normalise the basin indi-

[2] All data are freely accessible at https://ec.europa.eu/eurostat/data/database.

[3] Despite its success, it poses some conceptual and methodological questions. for a review, see for instance: [1, 2, 4].

cator and aggregate the indicators normalised. Normalization is required to make the indicators comparable, because they often present different measurement units and ranges. In the normalisation, it is necessary to define the *polarity* of the basic indicators, i.e. the sign of the relation between the indicator itself and the phenomenon. Therefore, the type of composite we want to construct defines polarity. In other words, some indicators may be positively related with the phenomenon to be measured (positive polarity), whereas others may be negatively related with it (negative polarity). After the normalisation, all the indicators must have positive polarity, i.e. an increase in the normalised indicators corresponds to an increase in the composite index [1]. In this paper, we use the re-scaling or Min-Max, which normalises indicators to be bounded in $[0,1]$ by subtracting the minimum value and dividing by the range of the indicator values. Let us suppose to have a bi-dimensional matrix, $\mathbf{X} \equiv \left\{x_{ij} : i = 1,\ldots,n; j = 1,\ldots,p\right\}$, where x_{ij} represents the determination of the j-th indicator in the i-th unit, the Min-Max normalisation is given by:

$$r_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\max_i(x_{ij}) - \min_i(x_{ij})} \tag{1}$$

where $\min_i(x_{ij})$ and $\max_i(x_{ij})$ are, respectively, a minimum and a maximum value (commonly the observed ones) that represent the possible range of the indicator j[4]. At this point, the normalised indicators are aggregated for each dimension to create the three dimensional indicators by using the Min-Mean function - MMF [5] given by

$$MMF_i = \mu_{r_i} - \alpha\left(\sqrt{\left(\mu_{r_i} - \min_j\{r_{ij}\}\right)^2 + \beta^2} - \beta\right) \qquad (0 \leq \alpha \leq 1; \beta \geq 0) \tag{2}$$

where μ_{r_i} is the mean of the normalized values (through any method of normalization) for unit i, and the parameters α and β are respectively related to the intensity of penalization of unbalance and degree of complementarity between indicators. Reasonable values for the parameters are $\alpha = \beta = 1$, which allow an intermediate easy case of adjustment with incomplete and progressive compensability. The last step is to aggregate the three dimensional indicators to construct the SDI, using the Min-Mean function with the same parameterization.

3 Application and results

Figure 1 shows the trends of the dimensional indices and the SDI in the 27 EU countries from 2015 to 2021. For all indices, we can observe trends trending upward,

[4] It applies if indicator has positive polarity, otherwise we compute the complement to respect to 1 to equation 1.

albeit slowly, in almost all EU countries. It appears clear that in all the countries, the highest values and trends are in the environmental dimensional index (EVI) and the lowest in the economic one (ECI). Looking at the EVI, we see that all countries show trends with values greater than 0.5 (the average value), except Romania (RO), Poland (PO) and Luxembourg (LU). To understand these values it is necessary to look at the elementary indicators used to construct the synthetic index: RO and PO show the worst trends in the SDG_06_10 indicator and LU in the SDG_13_10 indicator. With respect to the ECI, the situation appears split between countries with trends below the average value (for instance, Romania - RO, Greece - EL, Spain - ES), others with trends above the average (for instance, Sweden - SE, Austria - AT) and some with average trends (for instance, Czechia - CZ, Finland - FI). Similar considerations apply to the SCI, with RO and BG having the lowest trends and SE, CZ and AT the highest ones. Now let us focus on the Sustainable Development Index. Not surprisingly, RO has the lowest trend among European countries while SE has the highest. It is noticeable that the index values increased between 2015 and 2021 for all European countries. This is evident by looking at Figure 2, where the SDI values for the two years are compared. In particular, we can see that EL and BG go from values below 0.4 to values above; DE, BE, CZ, SI and SK passes from values below 0.6 to values above.

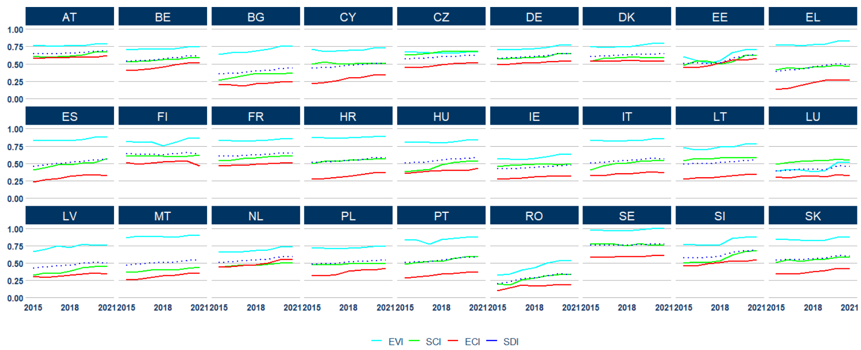


Fig. 1 Composites of environmental (EVI), social (SCI) and economic (ECI) dimensions and composite index of sustainable development (SDI): European countries; time series 2015–2021.

4 Conclusions

The concept of sustainable development and sustainability is gradually gaining importance in the international debate: political, economic and public. Meanwhile, the increasingly urgent search for solutions to respond to the emergence of the consequences of human activity on the environment and climate change calls the entire scientific community into the field in an attempt to define and measure sustainable

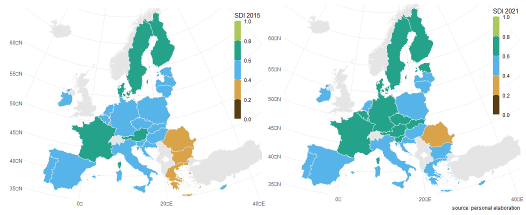


Fig. 2 Values of the sustainable development index (SDI) in the European countries: years 2015 and 2021.

development. In this paper, we use the definition and the framework proposed in the Agenda 2030 and the indicators selected by Eurostat to measure each goal for the European countries to construct synthetic measure of different dimensions and of the concept of sustainable development. Due to practical issues of data availability, not all the sustainable development goals (specifically 14 and 15) were taken into account. The results returned a clear and broad picture of the current European situation. Each country had an improvement in its sustainable development levels in 2021 compared to 2015.

References

1. Alaimo, L.S.: Complexity of Social Phenomena: Measurements, Analysis, Representations and Synthesis. Sapienza University Press, Roma (2022)
2. Alaimo, L.S.: Open issues in composite indicators construction. In: A. Balzanetta, M. Bini, C. Cavicchia, R. Verde (eds.), Book of Short Papers SIS 2022, 176–185. Pearson, Milano (2022)
3. Alaimo, L.S., Maggino, F.: Sustainable Development Goals Indicators at Territorial Level: Conceptual and Methodological Issues — The Italian Perspective. Soc Ind Res 147(2), 383–419 (2020)
4. Alaimo, L.S., Seri, E.: Measuring human development by means of composite indicators: open issues and new methodological tools. Qual Quant (2023). doi: 10.1007/s11135-022-01597-1
5. Casadio Tarabusi, E., Guarini, G.: An unbalance adjustment method for development indicators. Soc Ind Res 112, 19–45 (2013)
6. Sachs, J.D.: The Age of Sustainable Development. Columbia University Press, Columbia University (2015)
7. WCED - World Commission on Environment and Development: Our Common Future. The Brundtland Report. Oxford University Press, Oxford (1987)

Naples and tourism sustainability: A survey of citizens' perceptions

Napoli e la sostenibilità del turismo: Un'indagine sulla percezione dei cittadini

Massimo Aria, Francesca Pagliara, Luca D'Aniello and Valentina Della Corte

Abstract Tourism development can significantly affect local communities, including their economic, social, cultural, environmental, and political aspects. This study examines residents' perceptions of tourism development in Naples, Italy, across ten municipalities. Using a questionnaire, we investigate how residents perceive tourism's impact on economic, environmental, cultural, social, and political factors. Our findings offer insights for developers and policymakers to address any negative impacts of tourism development on local communities.

Abstract *Lo sviluppo turistico può avere un impatto significativo sulle comunità locali, inclusi i loro aspetti economici, sociali, culturali, ambientali e politici. Questo studio esamina le percezioni dei residenti sullo sviluppo turistico a Napoli, Italia, in dieci comuni. Utilizzando un questionario, abbiamo indagato su come i residenti percepiscano l'impatto del turismo su fattori economici, ambientali, culturali, sociali e politici. I nostri risultati offrono spunti per gli sviluppatori e i responsabili delle*

Massimo Aria

Department of Economics and Statistics, University of Naples Federico II, Italy; e-mail: massimo.aria@unina.it

Francesca Pagliara

Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II, Italy; e-mail: fpagliara@unina.it

Luca D'Aniello

Department of Social Sciences, University of Naples Federico II, Italy; e-mail: luca.daniello@unina.it

Valentina Della Corte

Department of Economics, Management and Institutions, University of Naples Federico II, Italy; e-mail: valentina.dellacorte@unina.it

politiche per affrontare eventuali impatti negativi dello sviluppo turistico sulle comunità locali.

Key words: Tourism development, Tourism impact, Citizens' perceptions

1 Introduction

Tourism is a rapidly growing and massive industry that contributes significantly to the global GDP [11, 4]. Previous studies have explored the attitudes of local residents toward tourism development, revealing their awareness of both the positive and negative impacts of tourism [16]. While tourism can bring about increased income, job opportunities, better infrastructure, and the promotion of local culture, it can also result in higher living costs, increased property prices, congestion, crime, and drug abuse [3, 9]. Residents' perceptions of tourism development are critical, as they influence their support or opposition to it and numerous factors influence them, including economic gain, economic involvement, community attachment, environmental attitudes, level of involvement in the planning process, and demographic factors, which have been extensively studied in the literature [1, 12].

Naples is a popular destination for Italian and foreign tourists, drawn to the city's unique blend of art and everyday life. Despite the pandemic's impact on the tourism industry, Naples was recently recognized by CNN as one of the 22 dream destinations for 2022, emphasizing the importance of understanding residents' perspectives on tourism development in the city. As tourism plays a significant role in the city's economy, examining the impacts of this industry on residents is critical for policymakers and developers.

This study aims to explore the perceptions of Naples residents on the impacts of tourism development on their community, with a focus on economic, social, cultural, environmental, social, and political aspects.

2 Tourism in Campania: Impact and Forecast

Tourism in Campania is expected to surpass pre-pandemic levels in 2023, with the most notable case being Pompeii, where the total number of visitors at the archaeological site has already exceeded 2019 levels in the first two months of the current year. The overall success of tourism is predicted to see a further increase in tourist flows in 2023, in line with the national trend. According to the Demoskopika Institute's "Tourism Forecast 2023" report, Italy is expected to reach a record-high of 442 million accommodations, with Campania estimated to receive 5.7 million visitors. As tourism remains the region's primary source of income, signs of recovery for the sector are recorded on the incoming side. Demoskopika estimates nearly 6 million arrivals for the current year, representing an increase of 13.1% compared to last year. Currently, Naples is the primary attractor of regional tourist flows. In December, for instance, 1.5 million tourists visited the city, occupying 80% of available hotels and *b&b* rooms in the first half of the month and 98% during the Christmas period [8].

The main objective of this study is to examine the impact of tourism in Naples, with a particular emphasis on the perceptions of the city's residents towards this significant influx of visitors and its effects on five critical factors: economic, environmental, cultural, social, and political. These factors were investigated through a survey by administrating a questionnaire to the citizens of Naples.

3 Questionnaire design

A structured questionnaire was utilized for conducting the survey. It was submitted to the Naples' citizens between December 2022 and January 2023 through mixed modes of administration (CAWI and CAPI).

The questionnaire was organized into two sections. In the first one, several questions were provided aiming at investigating the factors that influence the respondent's perception of tourist development in the city of Naples. This section focused on five important factors, including economic, social, environmental, political, and cultural ones. Participants were asked to rate their agreement with statements using a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The second section, on the other hand, pertained to the socio-demographic characteristics of the residents. In addition to the typical socio-demographic information, such as gender, age, and education level, respondents were also asked to indicate their municipality of residence in Naples. Although municipalities are not tourism-specific institutions, they are responsible for the planning and development of tourism in the Campania region [13]. This additional question aims to provide further insights into the respondents' perspectives and experiences, allowing for a more nuanced analysis of the data collected. Two questions were asked at the beginning of the questionnaire to filter respondents between those who reside in the city of Naples from those who do not. Out of 3017 distributed questionaries, 2563 were valid ones (85% of the response rate), by considering only residents, i.e., people who live, work, or study in the city.

4 Data description

Sample profile

The sample analyzed in this study included 2563 residents of Naples, with 47.6% women, 50.41% men, and 1.99% who preferred not to answer about their gender. Regarding age, most of the respondents fell between 18-44 years old (57.63%), while only 0.55% were over 85 years old. Education-wise, 25.32% of participants held a degree, while 22.32% completed middle school. In terms of occupational status, the largest percentage of respondents (33.63%) were employed, followed by students

(15.33%), and retirees (7.84%). Only 14.05% of the participants worked in the tourism industry, while the remaining 85.95% did not.

To assess the general perception of tourism in Naples, participants were asked whether they perceived it as a cost or a benefit. The findings revealed that a substantial majority (95.67%) of respondents viewed tourism in Naples as a benefit.

Distribution of responses among municipalities

Asking respondents which of the 10 municipalities they resided in was used to investigate potential differences in perceptions of tourism development across different areas of Naples. By examining the questionnaire response rate in each district, it was found that approximately 10% of residents in each area responded, indicating a balanced distribution of responses across the city.

The results of a questionnaire-based study investigating how residents of the ten municipalities in the city of Naples perceive tourism and its impact on five different factors are displayed in Figure 1.

Questions about the impact of tourism has on the economy include issues related to the economic benefits and costs of tourism, such as the impact on employment, income, and local businesses. The median score for this factor ranged from 3 to 4, indicating a moderate level of agreement among participants. The environmental factor covers topics related to the impact of tourism on the natural environment, including pollution, waste management, the deterioration of water quality in the port area, and the loss of green spaces in the city. The median score for this factor ranged from 2 to 4, with a higher level of variability in participants' responses. The cultural factor reflects the perceived effects of tourism on the preservation and promotion of local culture and heritage. Questions explore whether tourism has increased residents' cultural pride, helps maintain local culture and identity, positively impacts residents' lives, and contributes to the preservation of historical sites and artistic events in the city. The median score for this factor ranged from 4 to 4.7, indicating a relatively high level of agreement among participants.

The questions relating to the political aspect aim to measure the role of local authorities in regulating and managing tourism activities. It is asked whether the policy cares for the interests of the community, manages the relationship between citizens' needs and tourism development, and takes into account citizens' voices in tourism-related choices. The goal is to evaluate whether tourism policy is active and responsive to citizens' needs and whether it can balance tourism development with the community's well-being. The median score for this factor ranged from 4 to 4.5, indicating a relatively high level of agreement among participants. The social factor covers topics related to the impact of tourism on the local community, including social cohesion, cultural exchange, and quality of life. The questions aim to measure the impact of tourism on the social aspect, concerning vandalism, crime rates, public transportation, and quality of public services. The median score for this factor ranged from 2.86 to 4, indicating a moderate to a high level of agreement among participants.

Overall, the results suggest that residents of the ten municipalities in Naples have varying perceptions of the impact of tourism on different factors, with higher levels of agreement observed for the cultural and political factors.

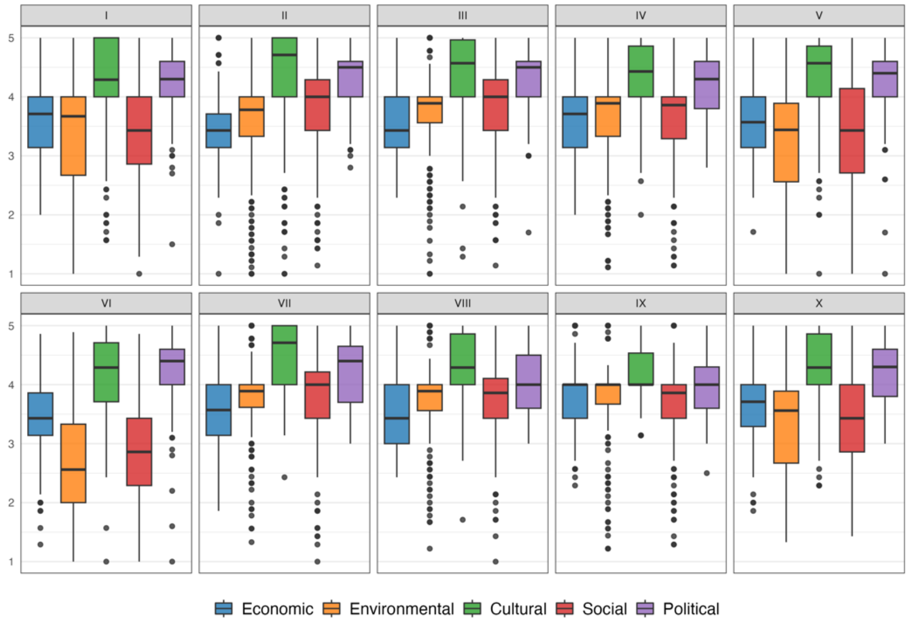


Fig. 1 Boxplot of responses about the five factors of tourism impact among ten municipalities.

5 Findings and future developments

The preliminary findings of this study suggest that residents perceive tourism as a positive force for the city's economic growth, cultural life, and identity. The presence of tourists from around the world is also seen as a positive experience, contributing to a sense of pride in their community. The findings of this study can provide valuable insights for policymakers and tourism stakeholders in developing strategies that can mitigate negative impacts and enhance positive ones, ultimately leading to more sustainable and responsible tourism practices.

To test the relationships among the theoretical models hypothesized in this study, future research will focus on testing these relations using the Structural Equation Modeling (SEM) statistical approach. This will allow for a more comprehensive understanding of the complex relationships between residents' perceptions of tourism development and its impacts on the economic, social, cultural, environmental, and political aspects of the community. This will provide a valuable contribution to the existing literature on the topic and inform policymakers and developers to promote sustainable tourism practices in a way that benefits both residents and tourists.

References

1. Andereck, K.L., Nyaupane, G.P.: Exploring the nature of tourism and quality of life perceptions among residents. Journal of Travel Research, 50(3), 248–260 (2011)
2. Broy, M.: Software engineering --- from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) Software Pioneers, pp. 10-13. Springer, Heidelberg (2002)
3. Deery, M., Jago, L., Fredline, L.: Rethinking social impacts of tourism research: A new research agenda. Tourism Management, 33(1), 64–73 (2012)
4. Dileep, M.R., Pagliara, F.: Transportation Systems for Tourism. Springer (2023)
5. Dod, J.: Effective substances. In: The Dictionary of Substances and Their Effects. Royal Society of Chemistry. Available via DIALOG (1999)
6. Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)
7. Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Annali di Matematica Pura ed Applicata, 169, 321-354 (1995)
8. Iuliano, V.: *Turismo, 2023 da record: In Campania previsto +12%*. https://www.ilmattino.it/napoli/cronaca/turismo_campania_napoli_boom_stranieri-7263534.html#:~:text=Il%202023%20%C3%A8%20destinato%20a,campane%20sono%205%2C7%20milioni. (2023)
9. Látková, P., Vogt, C.A.: Residents' attitudes toward existing and future tourism development in rural communities. Journal of Travel Research, 51(1), 50–67 (2012)
10. Leiper, N.: Tourism Management. Melbourne: RMIT Press (1995)
11. Morrison, A. M.: Marketing and managing tourism destinations, 2nd ed. London: Routledge (2019)
12. Nicholas, L.N., Thapa, B., Ko, Y.J.: Residents' perspectives of a world heritage site: The pitons management area. Annals of Tourism Research, 36(3), 390–412 (2009)
13. Pagliara, F., Aria, M., Russo, L., Della Corte, V., and Nunkoo, R.: Validating a theoretical model of citizens' trust in tourism development. Socio-Economic Planning Sciences 73, 100922. (2021)
14. Sharpley, R.: Host perceptions of tourism: A review of the research. Tourism Management, 42, 37–49 (2014)
15. Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. doi: 10.1007/s001090000086. (2000)
16. Vareiro, L.M.D.C., Remoaldo, P.C., Cadima Ribeiro, J.A.: Residents' perceptions of tourism impacts in guimarães (Portugal): A cluster analysis. Current Issues in Tourism, 16(6), 535–551 (2013)

Modelling inequalities for sustainable development in Italy countries

Modellare le disuguaglianze per lo sviluppo sostenibile nei paesi italiani

Mario Musella, Gianmarco Borrata, Ida Camminatiello and Rosaria Lombardo

Abstract In this paper we aim to investigate the socio-economic determinants that contribute to inequalities across the 20 regions of Italy through some suitable multiple regression models. Given the high number of socio-economic determinants, the problem of multicollinearity could affect the accuracy and reliability of our analysis. So we consider and compare three well-known regression models, i.e. the ridge regression, the least absolute shrinkage regression and the partial least squares regression, which aim to reduce the variance of the regression estimators while accounting for multicollinearity. We identify the most critical determinants of inequalities and compare the effectiveness of these regression models. The results of our analysis provide insights into the socio-economic determinants that significantly contribute to regional inequalities in Italy and how to best model them.

Abstract *In questo articolo ci proponiamo di investigare le determinanti socioeconomiche che contribuiscono alle disuguaglianze tra le 20 regioni d'Italia attraverso l'analisi di regressione. Tuttavia, siamo di fronte al problema della multicollinearità, cioè le variabili indipendenti sono altamente correlate e questo può influire sull'accuratezza e sulla affidabilità della nostra analisi. Per affrontare questo problema, valutiamo l'efficacia di tre modelli di regressione: ridge regression, least absolute shrinkage regression e partial least squares regression, che mirano a ridurre la varianza degli stimatori di regressione tenendo conto della multicollinearità. Identifichiamo le determinanti più critiche delle disuguaglianze e mettiamo in evidenza l'efficacia dei diversi modelli di regressione nell'identificare*

Mario Musella
University of Napoli, Federico II, Napoli, Italy, e-mail: mario.musella2@unina.it

Gianmarco Borrata
University of Napoli, Federico II, Napoli, Italy, e-mail: gianmarco.borrata@unina.it

Ida Camminatiello
University of Campania, Capua (CE), Italy, e-mail: camminat@unina.it

Rosaria Lombardo
University of Campania, Capua (CE), Italy, e-mail: rosaria.lombardo@unicampania.it

queste determinanti. I risultati della nostra analisi forniscono spunti sui fattori socioeconomici che contribuiscono significativamente alle disuguaglianze regionali in Italia e sul modo migliore per analizzarli.

Key words: Inequalities, multicollinearity, shrinkage methods, Sustainable Development Goals

1 Introduction

The "2030 Agenda for Sustainable Development" is a worldwide strategy that was approved by the United Nations in 2015. The plan focuses on 17 Sustainable Development Goals (SDGs) which tackle various global challenges and aim to promote sustainable and inclusive development.

Our research focuses on studying the dependence relationship between the indicators of Goal 10 which concern the *social inequality* and the indicators of Goals 3, 4, and 8 related to socio-economic determinants. This analysis is based on data collected from 20 regions in Italy.

To analyze how the different determinants (independent variables) impact on economic inequality across the different Italian regions, we consider the data produced by ISTAT regarding the 2030 Agenda. The independent variables result highly correlated, so we consider and compare the shrinkage estimators of three regression models, i.e. the Partial Least Squares (PLS) regression [12, 13, 14, 15], the Ridge Regression [6, 7] and the Least Absolute Shrinkage and Selection Operator (LASSO) regression [10].

2 Methodology

The Ordinary Least Squares (OLS) is the most simple method for studying the dependence relationship between one or more quantitative response variables and a set of predictors. While widely used in many fields, the model is susceptible to the multicollinearity problem, which arises when two or more predictors are highly correlated. Several shrinkage methods have been proposed over the years to address this problem, which attempt to shrink the coefficients to reduce their variances, while adding some bias.

In presence of multicollinearity, the stepwise selection of the predictors can be performed. For example, the Ridge regression [6, 7] and its modifications, the continuum regression [9], the least absolute shrinkage regression [10], the partial least squares regression [12, 13, 14, 15], the principal components regression and the latent root regression [11] have been proposed as alternative tools for dealing with multicollinearity. Also, Frank et al. [2] compare variable selection when performing Principal Components Regression (PCR), Ridge Regression (RR), and Partial Least

Squares Regression (PLSR). Although the results are conditional on the simulation design used in the study, it results that RR, PCR, and PLSR have similar proprieties, perform similarly, and are highly preferable for the variable selection. However, many criticisms have been made to PCR because the principal components are calculated without considering the response variable. Indeed, it can happen that the three methods carry out different results depending on data features.

In this paper, we choose to compare the main findings of RR, LASSO, and PLSR models when studying social inequalities. The PLSR model can be traced back to Wold [13] in its origins. During the initial stage of partial least squares regression, uncorrelated latent variables are generated, which are linear combinations of the original predictors.

The number of latent variables to be included in the model can be determined by applying several tools, among which cross validation [15]. The main objective of this procedure is to ensure that the weights used for determining the linear combinations of the original predictors are proportional to the covariance between the independent and dependent variables, as suggested by [5]. Subsequently, a least squares regression is conducted on the subset of extracted latent variables. This approach results in a biased but low-variance estimation of the regression coefficients compared to the OLS regression.

The RR model considers a known and small amount of bias into the regression equation estimation, with the aim of substantially reducing the inflated variances caused by multicollinearity. This reduction in variance helps to stabilize the estimates of the regression coefficients [6, 7].

The RR coefficients minimize a penalized residual sum of squares given by

$$\beta_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

where $\lambda \geq 0$ is a penalty parameter that controls the amount of shrinkage, the larger the value of λ, the greater the amount of shrinkage [4]. There are various procedures for choosing the ridge parameter. The aim of such procedures is of course finding the best λ values, but we cannot be sure to be successful in finding them [1]. The RR techniques have been shown to have a beneficial impact on point estimation theory. By selecting an appropriate value of λ, these techniques can produce estimators with smaller mean square errors compared to those of the least squares estimator [3].

The LASSO is a shrinkage method similar to the RR, but with subtle and important differences. The LASSO estimator is defined by

$$\beta_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The main difference is that the L2 ridge penalty $\sum_j^p \beta_j^2$ is replaced by the L1 LASSO penalty $\sum_j^p |\beta_j|$ [4].

The L1 constraint imposes a restriction on the absolute sum of the regression coefficients, rather than on their sum of squares as in the case of the L2 penalty considered in the RR. Because of the absolute value, the objective function of the LASSO regression is not differentiable at zero and becomes discontinuous. This makes the objective function of the LASSO regression non-convex and non-linear with respect to the regression coefficients. It means that, unlike the RR, the LASSO regression problem must be solved using numerical optimization methods as there is no a closed-form solution. A higher value of lambda results in a stronger penalty term, which allows the model to select a smaller subset of variables which are most relevant to the outcome variable. On the other hand, a lower value of lambda results in a weaker penalty term, which permits more variables to be included in the model.

3 Reducing inequalities

In recent years, sustainable development goals (SDGs) have become increasingly important. We aim to evaluate the most important determinants of inequalities across the Italian regions, looking at the ISTAT indicators of the SDGs (https://www.istat.it /it/ benessere-e-sostenibilit%C3%A0/obiettivi-di-sviluppo-sostenibile/gli-indicatori-istat). We consider three multivariate regression models, where the responses are two simple indicators of Goal 10, i.e. the *Income* and the *Poverty*. The correlation between them is negative and very high, equal to -0.89.

To evaluate the reliability of the three considered models, Table 1 reports the values of MSE (Mean Squared Error) for the complete, training and test datasets (MSETotal, MSETR and MSETE, respectively), and the R^2 (Coefficient of Determination) values for the complete, training and test datasets (R^2Total, R^2TR and R^2TE, respectively) in relation to the *Income* response, while Table 2 concerns the *Poverty* response. The results indicate that the PLSR model has the best performance among the three models considered, with lower MSE and higher R^2 compared to the other two models. However, it should be noted that the RR model has better performance than the LASSO model for some performance measures (e.g. R^2 of the training dataset). Overall, it results that the PLSR model is the best among the three models considered for predicting the *Income* and the *Poverty*, but it can be important to also consider other factors (e.g., model complexity, ease of interpretation, number of observations, etc.) before choosing the best model for other applications.

To provide a decision-making and executive tool for policymakers, we have analyzed the determinants which influence the socioeconomic inequalities. For the sake of brevity, we look at the first four most important regression coefficients of the three models for the complete dataset. Table 3 shows the best determinants of the *Income*:

- *QualityEdu1* (inadequate literacy, numerical and English skills);
- *QualityEdu6* (basic digital skills);

Table 1 Performance measures for the *Income* response variable

	PLS	RIDGE	LASSO
MSETotal	0.06	0.55	0.49
MSETR	0.04	0.28	0.73
MSETE	0.19	0.24	0.19
R^2Total	0.93	0.44	0.51
R^2TR	0.94	0.71	0.27
R^2TE	0.83	0.76	0.81

Table 2 Performance measures for the *Poverty* response variable

	PLS	RIDGE	LASSO
MSETotal	0.02	0.51	0.52
MSETR	0.02	0.14	0.54
MSETE	0.03	0.17	0.36
R^2Total	0.97	0.49	0.48
R^2TR	0.96	0.86	0.27
R^2TE	0.97	0.83	0.64

- *WorkGrowth2* ((unemployment/employment rate).
- *WorkGrowth3* (young people who do not work and not study- NEET).

From a logical point of view, given the high negative correlation between the two dependent variables, the determinants of one response should have an inverse impact on the other response. The results of the analysis confirm this, but with some small differences. Indeed when analyzing the *Poverty*, the second most important predictor is *QualityEdu3* (places authorized in socio-educational services for 100 children aged 0-2), negatively affecting the *Poverty* while *WorkGrowth2* is not included among these first four predictors; see (Table 4).
Furthermore, note that, for the *Income*, the LASSO regression selects only three predictors, e.g. *QualityEdu1, QualityEdu6*, and *WorkGrowth3*, differently from the RR and PLSR models.
In conclusion, this study shows that for reducing inequalities across the Italian regions, the social area of main interest for policymakers should be the Education quality.

Table 3 The first four most important coefficients regression for the *Income* variable

	QualityEdu1	QualityEdu6	WorkGrowth2	WorkGrowth3
PLS	-0.18	0.23	-0.15	-0.16
RIDGE	-0.17	0.21	-0.14	-0.15
LASSO	-0.18	0.60	- - - -	-0.05

Table 4 The first four most important coefficients regression for the *Poverty* variable

	QualityEdu1	QualityEdu3	QualityEdu6	WorkGrowth3
PLS	0.17	-0.20	-0.18	0.18
RIDGE	0.15	-0.18	-0.16	0.17
LASSO	0.10	-0.17	-0.07	0.56

References

1. Camminatiello, I.: Robust methods for partial least squares regression: Methodological contributions and applications in environmental field Doctoral thesis, pp. 11-12 (2006)
2. Frank, I.E., Friedman, J.H., Wold, S., Hastie, T. & Mallows, C.: A statistical view of some chemometrics regression tools. Technometrics 35(2), 109-148 (1993)
3. Halawa, A.M. & El Bassiouni, M.Y.: Tests of regression coefficients under ridge regression models. Journal of Statistical Computation and Simulation 65, 341-356 (2000)
4. Hastie, T., Tibshirani, R. & Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer (2009)
5. Helland, I.S.: On the structure of partial least squares regression. Communications in Statistics - Simulation and Computation 17, 581-607 (1988)
6. Hoerl, A.E. & Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55-67 (1970a)
7. Hoerl, A.E. & Kennard, R.W.: Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12, 69-82 (1970b)
8. ISTAT. SDGS. Report: Statistical Information for the 2030 Agenda in Italy (2021)
9. Stone, M. & Brooks, R.: Continuum Regression: Cross Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. Journal of the Royal Statistical Society Series B 52(2), 237-269 (1990)
10. Tibshirani, R.: Regression shrinkage and selection via Lasso. Journal of Royal Statistical Society Series B 58, 267-288 (1996)
11. Webster, J.T. Gunst, R.F. & Mason, R.L.: Latent Root Regression Analysis. Technometrics 16(4), 513-522 (1974)
12. Wold, H.: Estimation of principal components and related models by iterative least squares. Multivariate Analysis. (Eds.) P.R. Krishnaiah, 391-420. New York: Academic Press (1966)
13. Wold, H.: Soft modelling by latent variables: Non linear Iterative Partial Least Squares approach. Perspectives in Probability and Statistics: Papers in honour of Bartelett. (Eds.) J. Gani, pp. 117–142. London: Academic Press (1975)
14. Wold, H.: Partial Least Squares. Encyclopedia of Statistical Sciences (vol. 6). (Eds.) S. Kotz, N. L. Johnson, New York: Wiley, 581-591 (1985)
15. Wold, S.: Cross-validation estimation of the number of components in factor and principal components analysis. Technometrics 24, 397-405 (1978)

Food Security and Sustainability: A Science Mapping Analysis

Sicurezza alimentare e sostenibilità: un'analisi della mappatura scientifica

Alfonso Piscitelli

Abstract: The concepts of food security and food (system) sustainability are deeply interrelated, and their understanding enables interesting approaches for different fields of scientific research. To better understand both the relationship between people's access to sufficient, safe, and nutritious food and the sustainability of food production, as well as its characterization, a bibliometric study of international papers on this subject has been developed. A total of 5,468 documents in the last twenty-five years have been selected and analysed to discover the research topics in this field and the main associations related to the terms "food security" and "sustainability".

Abstract *I concetti di sicurezza alimentare e sostenibilità (del sistema) alimentare sono profondamente correlati e la loro comprensione rappresenta interessanti modelli di studio per diversi campi della ricerca scientifica. Per comprendere meglio sia la relazione tra l'accesso delle persone a cibo sufficiente, sicuro e nutriente e la sostenibilità della produzione alimentare, sia la sua caratterizzazione, è stato sviluppato uno studio bibliometrico di articoli internazionali su questo argomento. Un totale di 5.468 documenti negli ultimi venticinque anni è stato selezionato e analizzato per scoprire i temi di ricerca in questo campo e le principali associazioni legate ai termini "sicurezza alimentare" e "sostenibilità".*

Key words: Food security, Sustainability, Science mapping, Thematic evolution, Bibliometric studies.

Alfonso Piscitelli
Department of Agricultural Sciences, Federico II University of Naples; e-mail: alfonso.piscitelli@unina.it

1 Introduction

Food security is realized at the individual, household, national, regional and global levels when all people, at all times, have access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life [1].

Nowadays the need for achieving security in a sustainable framework is imperative and, being the two goals of security and sustainability sometimes contrasting, asks for finding a suitable balance in the specific context [2, 3].

This work aims to give an analysis of the thematic evolution of the food security and sustainability domain by science mapping approach. The thematic evolution of the field is analysed using Bibliometrix R-tool [4] on scientific articles, proceedings and book chapters indexed on Scopus bibliographic repository from 1997 to 2022. The bibliometric methods for science mapping allow us to quantify and visualize the thematic evolution of the studied field, detecting and visualizing conceptual subdomains (particular themes or general thematic areas). Bibliometric analysis, performed by Bibliometrix R-tool, has allowed us: (i) to identify the knowledge base of food consumption and well-being literature and its intellectual structure; (ii) to examine the research front (or conceptual structure) of the field; and (iii) to visualise the thematic evolution of the main topics. Finally, the interpretation based on the results obtained from the analysis is discussed.

2 Study Methodology

In recent years, increasing attention has been paid to the systematic study of scientific literature. Today it is possible to do in-depth research on a specific search domain thanks to the availability of online databases and the development of effective tools capable of performing automatic analyses. This study was developed from a bibliometric research perspective, aimed at discovering the main research dimensions and keywords employed in scientific publications linked with the theme of people's access to sufficient, safe, and nutritious food in a sustainable way.

Bibliometric analysis is a quantitative approach for the analysis of academic literature using statistical tools to provide the description, evaluation and monitoring of the published research [5]. The methodological aim is to analyse publications, citations and sources of information for science mapping, and it can be successfully applied to any scientific domain [6]. Practically, bibliographic data are processed through a well-defined workflow encompassing these steps: study design, data collection, data analysis, data visualization and interpretation. For this study, the Bibliometrix package [4] in the R programming language (https://www.r-project.org/) was used for the analysis and visualization of the bibliographic data from the Scopus database.

3 Data Collection

With the aim of understanding how the research on food security and sustainability has evolved, the data were retrieved from Scopus databases. We extracted documents published between 1997 and 2022 (incl.) which contained the topic “Food security” with the keyword “Sustainability” in the title or abstract (“Food security” AND “Sustainability”). The data were downloaded in January 2023 and we collected 6,057 documents, relating to 2,001 sources and 18,344 authors. Of our collection of documents, 3,670 are articles, while other types of documents are book, book chapter, conference paper, conference review, letters, and short survey. Generally, documents are scrutinized and included in the corpus only if they meet certain criteria related to the quality, reliability and validity of the document; hence all the conference reviews, letters, and short surveys are often excluded from the study, as documents in these typologies did not undergo a peer-review process [7]. In this analysis instead we have decided to include these document types. About authors, there are 17,524 authors of multi-authored documents, underlining the interdependence of the collaboration between authors, even from different countries and/or research fields.

4 Analysis and Discussion

In the first years of the analysis (1997- 2007) the number of publications is very low (276), underlying that the phenomenon was novel and as such probably not very well know and addressed by researchers. Only in the last decade, there has been a significant increase in the number of publications per year. The year with the highest number of publications is 2022 (1007 documents published). Figure 1 emphasizes the dense network of worldwide collaboration. The authors who have distinguished themselves in terms of the number of publications related to this topic come mainly from the USA, China and United Kingdom, but we can note that there is a strong collaboration between USA, Canada, Australia, and almost all European countries (the strength of the collaboration is indicated by the thickness of the links). This underlines how this issue is of a global nature and does not concern only some nations.

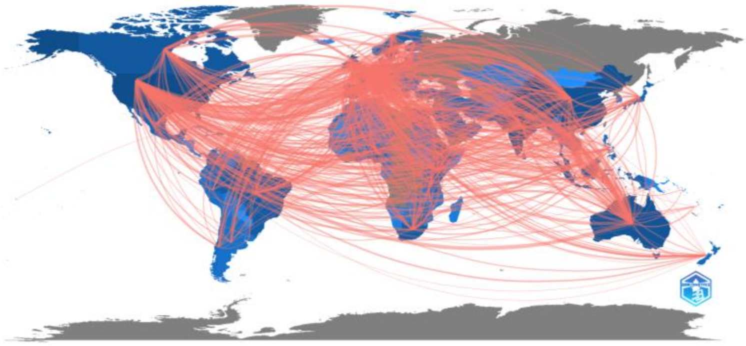


Fig. 1 Country collaboration map

The visualization in Figure 2 shows the network of KeyWords Plus (KWP) co-occurrence. The thickness of the connecting line between any two keywords represents the strength of their co-occurrence. The size of the KWP represents the index of their centrality. The nodes “Food security” and “Food supply” are both linked to “Sustainability”, and “Sustainable development”, but also “Agriculture”, and “Climate change”, which are the KWP most frequently adopted, along with “Human”, "Environmental sustainability" and “Diet”. We performed a hierarchical clustering using co-occurrence network to identify clusters of documents that express common concepts. Results show how food security is not only a food production problem but also has a significant link to social life of people and environmental ecosystem.

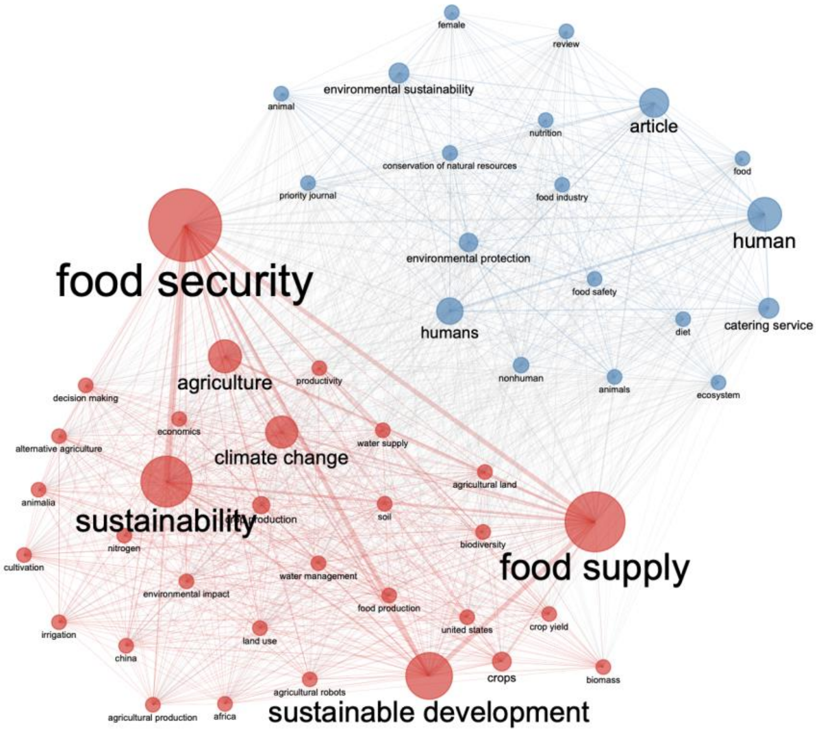


Fig. 2 Co-occurrences network of Keywords plus

Finally, Figure 3 shows the KWP considered as themes, classified by different levels of density and centrality in the network of scientific KWP. In the strategic diagram presented in Figure 3, the vertical axis measures the density, namely the strength of the internal links within a cluster represented by a theme, and the horizontal axis the centrality, namely the strength of the links between the theme and other themes in the map [7].

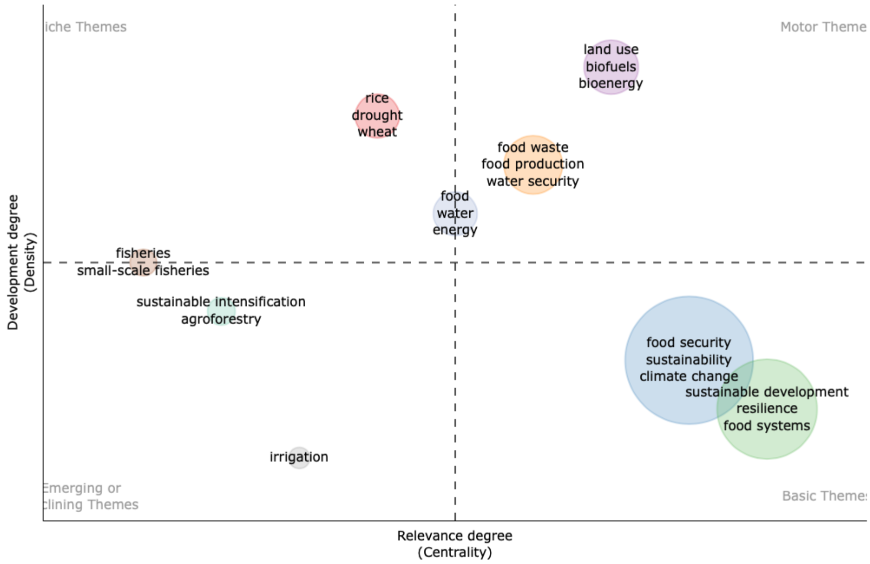


Fig. 3 Thematic map of Keywords plus

As can be seen, land use, food waste, and water security appear as motor themes, while irrigation or fisheries policy are emerging themes or niche themes. This result underlines how researchers, in addition to studying food security and its basic dimensions such as the availability, accessibility and utilization of food, also try to study aspects related to the policies of the sustainable food system.

5 Conclusions

This study analyses the knowledge structure of publications on food security and sustainability over the past 25 years with a quantitative approach. Our results demonstrate that research on food security and sustainability is mainly driven by two different clusters that showed a moderate degree of interconnection, requiring a higher level of synergy to promote more translational research in this field. Finally, the mapping of thematic areas helped to identify main research interest and provided some clues into future research direction.

References

1. FAO: *Declaration on World Food Security and World Food Summit Plan of Action*. World Food Summit 13-17 November 1996. Rome (1996)
2. Berry, E. M., Dernini, S., Burlingame, B., Meybeck, A., and Conforti, P.: Food security and sustainability: can one exist without the other? Public health nutrition 18(13), 2293-2302 (2015)
3. Ferranti, P., Berry, E., and Jock, A. (2018). *Encyclopedia of food security and sustainability*. Elsevier.

4. Aria, M., and Cuccurullo, C.: Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975 (2017)
5. Garfield, E., Sher, I.H., Torpie, R.J.: *The use of citation data in writing the history of Science*. Institute for Scientific Information Inc Philadelphia PA. (1964)
6. Rodriguez-Soler, R., Uribe-Toril, J., Valenciano, J. D. P.: Worldwide trends in the scientific production on rural depopulation, a bibliometric analysis using bibliometrix R-tool. Land Use Policy, 97, 1047-87 (2020)
7. Marsh, H.W., Jayasinghe, U.W. and Bond, N.W.: Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. American psychologist, 63(3) (2008)
8. Pourkhani, A., Abdipour, K., Baher, B., Moslehpour, M.: The impact of social media in business growth and performance: a scientometrics analysis. International Journal of Data and Network Science, 3(3), 223-244 (2019)

Solicited Session SS9 - *Inequalities in the labour market*

Organizer and Chair: Francesca Adele Giambona

1. *Skill similarities across Italian regions: an analysis based on the online job advertisements* (Kahlawi A., Buzzigoli L., Grassini L., Martelli C. and Giambona F.)
2. *Italian Labour Market reform and gender inequalities* (Marini C. and Nicolardi V.)
3. *Intergenerational transmission of disadvantages in the Italian labour market: evidence from AD-SILC data* (Busetta A., Fabrizi E., Ragozini G. and Sulis I.)

Skill similarities across Italian regions: an analysis based on the online job advertisements

Similarità di competenze nelle regioni italiane: un'analisi degli annunci di lavoro online

Adham Kahlawi, Lucia Buzzigoli, Laura Grassini, Cristina Martelli and Francesca Giambona

Abstract The online job advertisements (OJAs) data allow us to improve the knowledge of the labour market with timely data about the demand of businesses and the skills required for each job position. In this contribution, we try to explore the variations in skills over time and the similarities in the profiles of skills at the regional level for Italy. In light of this, by using the Lightcast data for 2019 and 2020, we use two measures to represent the regional skill changes and regional skill similarities. Finally, by considering the profiles of regional skills jointly with some features of the local economy, an attempt is made to classify Italian regions.

Abstract *Gli annunci di lavoro online (OJA) offrono la possibilità di migliorare la conoscenza sul mercato del lavoro con la disponibilità di dati tempestivi sulla domanda proveniente dalle imprese, anche in termini di competenze richieste per la posizione lavorativa. In questo contributo, cerchiamo di esplorare le variazioni delle*

Adham Kahlawi
University of Florence; e-mail: adham.kahlawi@unifi.it

Lucia Buzzigoli
University of Florence; e-mail: lucia.buzzigoli@unifi.it

Laura Grassini
University of Florence; e-mail: laura.grassini@unifi.it

Cristina Martelli
University of Florence; e-mail: cristina.martelli@unifi.it

Francesca Adele Giambona
University of Florence; e-mail: francesca.giambona@unifi.it

competenze nel tempo e i profili regionali delle competenze. A tale scopo i dati Lightcast, per gli anni 2019 e 2020, sono stato usati per sintetizzare le similarità e il cambiamento delle competenze richieste a livello regionale. Infine, considerando sia i profili delle competenze regionali che alcune caratteristiche dell'economia locale viene proposta una classificazione delle regioni italiane.

Key words: online job ads, Italian regional inequalities, occupations, skills.

1 Introduction and theoretical framework

The use of web data to address socio-economic issues and integrate existing sources of information is on the rise. Online platforms and websites produce a wealth of data that can provide valuable and multidimensional information for various applications. In particular, with the growth of the internet and knowledge, the availability of online tools for job searching, candidate searching, and job matching – the so-called Online Job Ads (or Job Posting) data (OJA data or OJAs) – provides a richness of job-related information. Indeed, OJAs cannot replace traditional official labour market information. Still, they can provide additional comprehensive, detailed (also at a geographical level), and timely insights into labour market trends by allowing the analysis of skills requested in occupations by job seekers.

Our research uses the dataset produced by Lightcast for Italy, which collects millions of online job postings daily from various online resources such as dedicated portals and company websites. The dataset creation process is complex and involves collecting data from different sources using multiple methods, eliminating noise, outliers, and duplicate entries through pre-processing, and coding the content of the ads using categories based on reference taxonomies using text classification algorithms. The dataset contains dozens of variables, such as opening and closure date of publication, identification and description of occupation and related skills, geographic job location, economic sector of the company, and educational level required. When possible, the classification of data items refers to official standards, like the European NUTS (up to the LAU1 level) for the territorial units, the NACE 2008 (up to the second level) for the economic sector, the European multilingual standard classification language (ESCO) v.1.1.1 for occupations and skills. In particular, for occupations, the maximum detail is ESCO 4 digit, which corresponds to the fourth level of ISCO 1.08.

This detailed information can be particularly relevant to analyse the differences in skills demand in Italian regions. For example, the 2021 Unioncamere Report has highlighted the increasingly wide misalignment, or skill mismatch, between job supply and demand, with considerable territorial differences (Unioncamere, 2021). To analyse Italian OJAs data from a regional perspective, we use: i) a measure of regional skill similarity to assess whether skills required for a specific occupation are similar between Italian regions; ii) the index proposed by Deming and Noray (2020) to evaluate the regional skill changes between 2019 and 2020; iii) information about some features of the regional economic context to better interpret the results of the analysis i) and ii).

1.1 Lightcast data

The data used in the analysis refers to the OJVs posted on 239 online job portals in Italy from January 2019 to December 2020. 421 out of the 511 ISCO occupations occur in the data. The number of skills is 1,200, of which 917 have been traced back to ESCO classification (the total number of skills in ESCO is 13,485).

The number of job ads slightly differs between 2019 and 2020, with 1,311,833 and 1,448,808 job ads, respectively. Over 150 thousand job ads each year do not include skills, while approximately 1,150 job ads are not classified under any job title (i.e. without ISCO occupation id). Note that the single job ad may demand more than one job position, even in different places (different provinces, regions). The information on the region is missing in 15.9% of total job ads in 2019 and 7.2% in 2020, denoting a slight improvement in the textual data processing and/or quality of the raw data. The higher percentages of job ads occur in the regions with the largest populations (such as Lombardia or Veneto). The same happens for the average number of job ads per 100 inhabitants: Lombardia outstands from all with 29.1 and 27.7. Emilia Romagna and Veneto follow at a considerable distance, while the other regions are below 10. The average number of job ads per 100 inhabitants has grown in all regions. An interesting feature is the variety of skills and occupations within regions. These numbers may express the complexity of the region's labour market, but the territory's size still influences them: large numbers comparable with those by Lombardia occurs for Veneto, Emilia Romagna, Lazio and Piemonte. The correlation between the two sets of data (number of occupations and number of requested skills) across regions is very high, settling at around 0.94 for the two years.

2 Methods and empirical findings

We analyse regional skills using a multistep methodological approach.
In the first step, we summarise regional skill similarities by specifying a skill similarities index (SSI); in the second step, we consider the regional skill change by using the skill change index (SCI) proposed by Deming and Noray (2020) and, finally, we propose a classification of Italian regions based on cluster analysis (CA). In the following, we present a brief description of the methodology jointly with the main empirical findings.

Skill Similarities Index (SSI)
The empirical analysis using *SSI* aims to understand if, in Italy, the labour market is affected by regional skills similarities. We construct a matrix where each column corresponds to a region, and each row represents the combination of occupation and

skills mentioned in job advertisements. Then, we employ the skill importance indicator (*SI*) in the specific occupation and region in formula (1) to compute the values of each cell in the matrix:

$$SI_{R,O,S} = \frac{\#JobAds_{R,O,S}}{\#JobAds_{R,O}} \tag{1}$$

where $\#JobAds_{R,O,S}$ denote the count of job advertisements demanding skill *S* for occupation *O* (ESCO-4 detail) in region *R*, while $\#JobAds_{R,O}$ represents the total number of job ads demanding *O* in region *R*. $SI_{R,O,S}$ represents the proportion of job ads in region *R* and for occupation *O*, that necessitate skill *S*.

The indicator of skill importance implements a comparison between regions with the mediation of the occupation *O*: in fact, the skill importance is conditioned to a specific occupation.

In assessing the similarity between regions, we combine traditional statistical methods with data mining and machine learning techniques, which can solve the computational complexity of large and sparse datasets (Koren et al. 2009). Firstly, we factorise the matrix, and secondly, we calculate a similarity measure between regions. Operationally, firstly we create a vector representing each region's profile using the Collaborative filtering algorithm (Bhumichitr et al., 2017; Jiang et al., 2019; Paleti et al., 2021), and then apply the cosine formula (2) to measure the (cosine) similarity between each couple of factorised regional profiles *A* and *B*

$$sim(A,B) = \frac{A'B}{\|A\| \cdot \|B\|} \tag{2}$$

As a result, we obtain 19 similarity values for each region. Subsequently, we compute the arithmetic mean and standard deviation of each region's similarity values. Considering the average similarity, the values are very low, showing a low similarity between regions, especially with respect to the Southern regions, for example, Sardegna, Umbria and Calabria for which the regional mean value is about 0.07. Higher regional mean similarity values are present for some Center regions and (especially) for Northern regions as, for example, for Piemonte (0.22), Emilia-Romagna (0.24), Veneto (0.27) and Lombardia (0.29); in these regions, there is also the highest variability.

Skill Change Index (SCI)

The *skill change index (SCI),* obtained for each region, allow us to assess if skill changes occurred between 2019 and 2020. To measure the shift in skill demand from 2019 to 2020, we use the index developed by Deming and Noray (2020), defined as:

$$SCI_R = \sum_{s=1}^{S} \left| \left(\frac{\#JobAds_{Rs}}{\#JobAds_R}\right)_{2020} - \left(\frac{\#JobAds_{Rs}}{\#JobAds_R}\right)_{2019} \right| \tag{3}$$

where *R* stands for each Italian region: $\#JobAd_{Rs}$ is the number of job ads in region *r* requiring skill *s*, and $\#JobAd_{sR}$ is the number of job ads in region *R.* Note that in this analysis there is no mediation of the occupation as we compare data of the same region over time. Results reveal higher *SCI* values for Molise, Calabria and Lazio, and lower values for Friuli Venezia Giulia, Marche and Emilia Romagna.

Cluster Analysis (CA)
Finally, we try to group the Italian regions to detect if there is territorial clustering in terms of skill similarity, skill change and local economic context. In particular, we assess if the well-known Italian north-central-south tri-partition is confirmed through a cluster analysis based on the *k*-means algorithm (Everitt *et al.*, 2011).
For clustering regions, we choose the six variables listed below:

- the regional context is described by three variables that could describe the dynamism of the local business context: 1) the rate of youth entrepreneurship, ***YE*** (total owners of sole proprietors with less than thirty years of age as a percentage of the total owners of sole proprietorships); 2) the net rate of registration in the business register, ***RR*** (percentage of companies registered minus companies that have ceased to exist, out of the total number of companies registered in the previous year); 3) the degree of use of internet in businesses, ***INT*** (percentage of employees of businesses with more than ten employees that use computers connected to the internet). Data have been selected by Istat for 2020 (Istat, 2020).
- the regional skill similarity, that is: 4) the skill similarities index (***SSI***) mean values and 5) the standard deviation (***SSIsd***)
- 6) the skill change index (***SCI***).

The main results of the cluster analysis are in Table 1. Cluster 1 (**C1**), includes eight regions: Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Sardegna; cluster 2 (**C2**) groups six regions: Piemonte, Valle d'Aosta, Trentino Alto Adige, Lombardia, Liguria, Lazio; finally, cluster 3 (**C3**) includes six regions: Molise, Campania, Puglia, Basilicata, Calabria, Sicilia.

Table 1 CA, *main results*

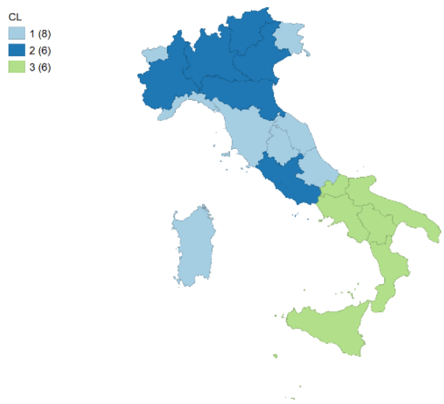
Method: KMeans

Distance function: Euclidean

Cluster centers

Variable	C1	C2	C3
YE	4.856	5.331	6.054
INT	48.796	55.258	40.224
RR	-0.441	-0.381	0.660
SSI	0.096	0.238	0.094
SSIsd	0.046	0.230	0.062
SCI	1.880	2.062	2.489

Fig. 1 *Cluster map*



Overall, from Table 1, the cluster center values increase as the cluster grows, except for the degree of internet use in business (INT), which assumes the highest value for C2 for which we observe greater SSI and SSIsd. C1, which groups mainly Center regions, is characterised by lower center values except for INT. In the South, C3, we observe the highest center values for all indicators except for INT and for variables related to regional skill similarities (SSI and SSIsd). Cluster 2 is very cohesive internally in terms of skill similarity. In fact, the average value of the paired similarities between regions is 0.502 while the average similarity is about 0.09 within C1 and C3.

3 Concluding remarks

The OJAs data give us a chance to improve information about the labour market with the availability of timely data about the demand of businesses and the skills required for each occupation: skills perspective helps to provide a more focused, comprehensive and detailed understanding of the labour market trends, helping to recover a vision more adherent to the modern complexity of work and going beyond the traditional hierarchical perspective of occupation and economic sectors.

This is relevant for regions due to local implications on the labour market as regional socio-economic context and business environment clearly affect skills demand.

In light of this, by considering the skill similarity index and the skill change index jointly with some regional variables related to the regional context, we have attempted to classify Italian regions into groups. Surely other variables could be chosen, or other classification methods could be used, but this first proposal confirms the territorial tri-partition, North-Center-South, with some exceptions.

References

1. Azar, J., I., Marinescu, E., Steinbaum, M., Taska, B.: Concentration in US Labor Markets: Evidence from Online Vacancy Data. Labor Economics. 66 (2000)
2. Bhumichitr, K., Channarukul, S., Saejiem, N., Jiamthapthaksin, R., Nongpong, K.: Recommender Systems for university elective course recommendation. 14th International Joint Conference on Computer Science and Software Engineering. (2017) doi: 10.1109/JCSSE.2017.8025933.
3. Deming, D.J., Noray, K.: Earnings Dynamics, Changing Job Skills, and STEM Careers. Quarterly Journal of Economics. Forthcoming.
4. Everitt. B. S., Landau, S., Morven, L., Stahl, L.: Cluster Analysis. Wiley, New York (2011)
5. Istat.: Indicatori per le politiche territoriali. Roma (2020)
6. Jiang, L., Cheng, Y., Yang L., Li, L., Yan, H., Wang, X.: A trust-based collaborative filtering algorithm for E-commerce recommendation system. Journal of Ambient Intelligence and Humanized Computing. 10, 8, 3023–3034 (2019)
7. Koren, Y., Bell, R., Volinsky, C.: Matrix Factorization Techniques for Recommender Systems. Computer (Long Beach Calif). 42, 30–37 (2009)
8. Paleti, L., Radha Krishna, P., Murthy, J. V. R.: Approaching the cold-start problem using community detection based alternating least square factorisation in recommendation systems, Evolutionary Intelligence. 14, 2, 835–849 (2021)
9. Unioncamere.: Le Competenze digitali. Analisi della domanda di competenze digitali nelle imprese. Roma (2021)

Italian Labour Market reforms and gender inequalities

Riforme del Mercato del Lavoro italiano e disuguaglianza di genere

Caterina Marini and Vittorio Nicolardi

Abstract The reforms of the Italian job market occurred over a period of 18 years since 1997. The effects of reforms on the occupational levels are complicated to be depicted because of the overlapping with those of the unpredictable episodes of the global crises. In this paper, we analyse the economic statistical effects on the labour forces of the governmental interventions. Over a period of 21 years, starting from 2004, we present a partial analysis based on the I.Stat data warehouse that provides a first and aggregated information on the labour forces. In this first attempt of analysis, we found out that the Italian entrepreneurial system reacted positively to all reforms in a gender dimension. Therefore, somehow reforms preserved the female employment for some aspects, and restarted the male employment for some others.

Abstract *Le riforme del mercato del lavoro italiano sono prodotte in un periodo di 18 anni dal 1997. Gli effetti delle riforme sull'occupazione sono di complessa descrizione a causa della sovrapposizione con quelli delle imprevedibili crisi mondiali. In questo lavoro analizziamo gli effetti economici sulle forze di lavoro degli interventi dello Stato italiano. A partire dal 2004, per 21 anni, presentiamo una prima analisi basata sul data warehouse I.Stat che offre una prima e aggregata informazione sulle forze di lavoro. In questo primo tentativo di analisi, noi evidenziamo come il sistema imprenditoriale italiano reagisca positivamente a tutte le riforme in una dimensione di genere. Quindi, in parte preserva l'occupazione femminile per alcuni aspetti, e in parte rilancia l'occupazione maschile per altri.*

Key words: Labour market reforms. Employment. Gender inequality.

Caterina Marini
University of Bari Aldo Moro, e-mail: caterina.marini@uniba.it

Vittorio Nicolardi
University of Bari Aldo Moro, e-mail: vittorio.nicolardi@uniba.it

1 Introduction

In the last two decades, the juridical framework of the Italian labour market has undergone several major changes to support the employment of all segments of the active population on the one hand, and the global competitiveness and productivity of the national enterprises and the productive activities on the other hand. The introduction of reforms was considered fundamental and no further procrastinating over a period that started at the beginning of the new millennium and saw the national economies of most countries, and Italy was part of them, deeply affected by the most various worldwide economic recessions. In particular, since the end of the 20th century Italy had an entrepreneurial system that was suffering the consequences of a rigid regulation of the job market, inappropriate in a world more flexible and competitive, and pressing for a change that could facilitate its survival. Therefore, starting from 2000 the Italian Government gradually and deeply reforms the regulatory framework of the job market to reconnect the country with the rest of the occidental economies and save its own stability. The economic context, when the reforms of the job market occur, is particularly unstable and variable because of the globalisation of markets and productive processes, and enterprises and companies face situations continuously changing where it is hard to preserve important market shares and gain new. On the other side, the subsequent social context reveals a labour force more vulnerable in the structural changes that necessarily involves the rapports between employers and employees. It is not intention of this work to report and analyse the juridical intentions of legislators in their decisions of jurisprudence to revise the system of rules for the job market to stimulate the employment and, consequentially, sustain the Italian economy. The economic statistic effects on the labour forces of the governmental interventions are under analysis, instead. In this paper, we depict some first numerical evidence of the effects of the labour market reforms on the gender employment over a period of 21 years, starting from 2004. It is well known that in Italy the woman participation to the labour activity is systematically lower than man participation, and compared with EU experience the Italian case is worthy of attention.

2 Employment in Italy and labour market reform scheme

The reforms of the Italian job market were addressed to the active population in all ages and all status, both unemployed and seeking a first employment, both young and middle-aged. At the end of the 20th century, a great part of the males started losing their jobs mainly in the manufacturing economic sectors because of the begin of the global recession and entered the payroll subsidy scheme status. Young and women are, instead, those segments of the Italian labour forces that historically struggle to find a job and, when this is the case, gain a fair income and a qualified position based on the educational level. Saltari and Travaglini [5] highlighted the above weakness of the Italian labour market underlining the spatial divergence between the two main Italian macro areas (North and South). Therefore, many are the political and juridical intervention lines that needed to be addressed: first, a shift towards a concrete flexibility in the labour market and employment opportunities; second, an adequate protection for those segments of labour forces more vulnerable through interventions ad hoc to stimulate their employment; third, an economic support to reduce the labour cost for those companies and enterprises that proceeded towards the government occupational programme; fourth, monetary incentives to enhance the stabilisation of job positions and reduce the overuse of temporary contracts. It is clear that the flexibility is the real challenge and the great change for the Italian labour market.

Many authors analysed the importance of flexibility on the employment rate and productivity [1,2]. In literature, the Italian labour market is extensively analysed but the complexity of the phenomenon as arises from the interference of the long reform phase on one hand, and the limited availability of open data on the other make the analysis of the labour forces and employment in Italy a complementary work with the others. Cirillo et al. [3] tried to describe the first effects on the employment of the last reform that was approved in Italy in 2014 (implemented in 2015), known as Jobs Act (JA, hereafter), alongside extraordinary economic and financial measures that planned monetary incentives as discounts in social contributions for a 3-year period for those companies/enterprises that employed new people with permanent contracts or transformed temporary contracts in permanent contracts for people already employed. Although the availability of data was limited and information aggregated, the authors highlighted the first evidence of the positive effects of the combined policies (JA + Monetary incentives) on employment. As shown elsewhere, see Marini and Nicolardi [4], the transition probabilities in the transition matrices that evaluated the change of job status from temporary to permanent highlighted that the JA met the requests from the entrepreneurial sector to the Government and the monetary incentives strengthened the employment of young in permanent positions. The description of the reform process that covers a period of 18 years and 4 important labour market reforms is reported in Table 1.

3 Stylised fact

The sole data source used for this preliminary analysis is I.Stat, i.e. the data warehouse of the Italian National Institute of Statistics (ISTAT, hereafter), where the time series of the section Labour Offer have been reconstructed to guarantee the homogeneity of information related the field considered since 2004 because of the reform of the Labour Force Survey that in 2021 introduced new definitions for the variables in compliance with the new Eurostat regulation. Data are quarterly time series and the period under investigation is 1st Quarter 2004 – 4th Quarter 2022. Therefore, the endurance of the last juridical framework for the labour market (that is the JA) will be analysed also over the pandemic period of SARS-Cov2.

Table 1 Italian Labour Market Reforms and Monetary Incentives Programmes

2001 – D.Law no. 368	Revision of the fixed-term contract: fewer causal restrictions
2003 - Legge Biagi D Law no. 276	- Introduction of the para-subordinated and non-standard contracts; - Revision of the existing temporary contracts, i.e. coordinated collaboration, apprenticeship, outsourcing contract, intermittent contract, project-contracts, experience contract.
2012 - Legge Fornero Law no. 92	- Promotion of the open-ended contracts; - Enhancement of the apprenticeship and limits for overuse; - Removal of the placement contract; - Revision of the fixed-term contract: fewer causal restrictions; - Reduction of the effectiveness of the Articolo 18, the worker protection in case of illicit layoff.
2014- D. Law Poletti D. Law no. 34 2014 - Jobs Act Law no. 183 2015 Stability Law 2016 Stability Law	- Promotion of the open-ended contracts; - Introduction of the new open-ended contract known as increasing protection contract; - Abrogation of the Article 18; - Revision of the fixed-term contract: no causal restrictions; - Revision of the outsourcing contract - Removal of the project contracts; - Monetary incentives as total discounts in social contributions for a 3-year period (Stability Law 2015) and 60% for a 2-year period (Stability Law 2016)

Data reflect at a glance not only the effects, when they occur, of the implementation of the job market reforms but also all the unpredictable episodes of the global crises and their effects on the stability of the Italian economic system[1]. In fact, as we can see in Figure 1, very often the effects are overlapped with the reforms and sometimes the structural breaks are unquestionable, as in the case of SARS-COV2 period (years 2020 – 2021). The analysis of total employment (Figure 1a) highlights how Legge Biagi induced a moderate increase in the number of employees, and they are both men and women (Figure 1b and 1c), until the 2nd Quarter 2008 when in the aftermath the decrease and instability of employment will persist for a long period. For the whole period considered in the analysis, the employment of women is always lower (a range of 8.7mil – 9.9mil) compared with the occupational levels of men (a range of 12.5mil – 13.9mil) and trends are differently characterised based on the NACE sectors and the consequences yielded as counterparts of reforms or general worldwide instability. The major impacts on employment of the economic – financial crisis, which started in US in 2008 and involved EU in the second half of 2008 and 2009, were devastating for men (-652thsnd) compared with women (-279thsnd) and the reason of this evidence is to be found in the Industrial sector that has been the most vulnerable of the economy during the crisis. The consequences of the economic crisis jointly with the sub-prime crisis in EU mitigated also the effects of Legge Fornero that was implemented in July 2012 to stop the haemorrhage of job roles and contain

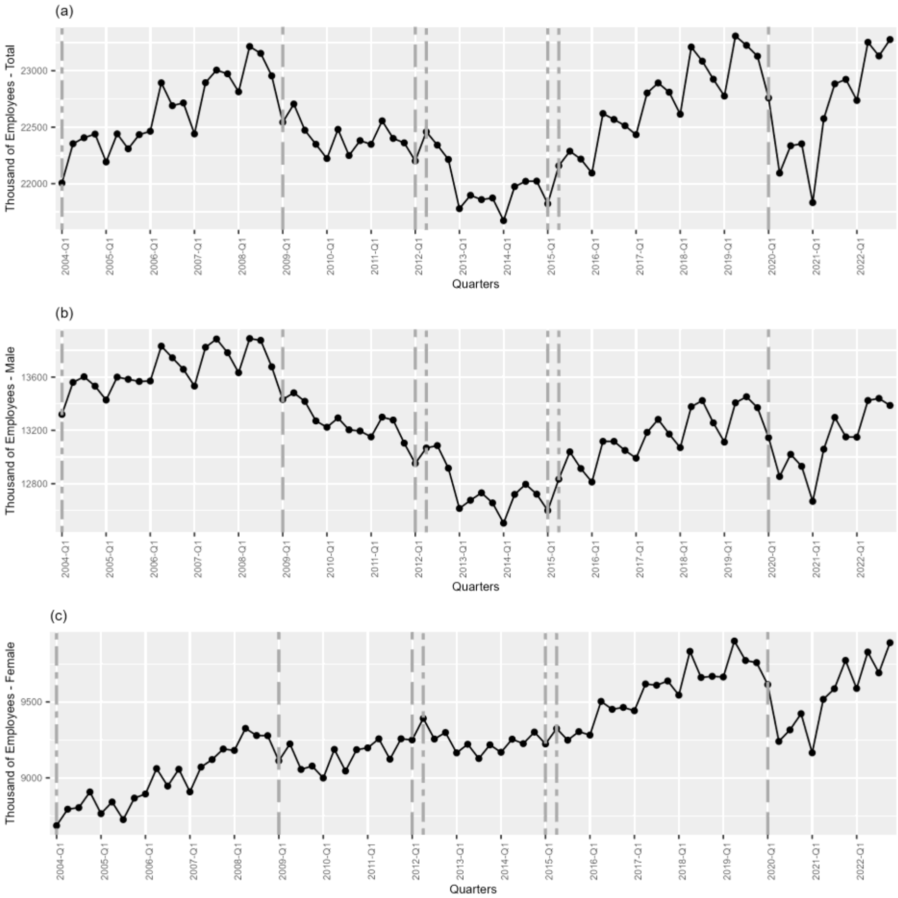
[1] In all Figures in the text the dashed vertical lines denote the time of the events of the global instability, while the dot-dashed vertical lines indicate the time of the implementation of the Italian Government reforms.

the increase of unemployment. In this case, no effects were noted on male employment, while the trend of female occupational level remained almost steady or was slightly increasing until the 4th Quarter 2015. We could affirm at a glance that on this particular period of the Italian economy (2008 – 2014) the reform Legge Fornero sustained the employment of Italian women because they are mainly employed in the Service sector (82% on average over the period 2004 – 2022) whose occupational female trend was slightly increasing for the whole period. The occupational level of men will restart sharply increasing in 2015 when the JA and the 2015 monetary incentives will be implemented. In fact, as we can see in Figure 1b, the employment of men started increasing on the 2nd Quarter 2015 when the JA was implemented (March 2015) and shows an average increase of 5.2% until the 4th Quarter 2018. The effect of JA on the female employment is less remarkable and delayed of one year. In fact, the employment of women started increasing in the 2nd Quarter 2016, when the second tranche of monetary incentives were implemented by the Italian Government. We could hazard saying that the entrepreneurial Italian system decided to save the industrial sector and male employees first, and afterwards decided to stabilise female employment. The analysis of typology of contracts, temporary vs permanent, highlights that the combined JA plus monetary incentives (2015 – 2016) yielded the effects as the legislator expected to reach. In fact, both permanent and temporary positions increased in the years 2015 – 2019 but the permanent contracts increased sharply during the period while the temporary had a variable trend. Trends were almost similar for both men and women. The effects of the monetary incentives ended in the 4th Quarter 2018 and, in fact, in 2019 trends of employment are almost steady.

4 Concluding remarks

The analysis of the Italian labour market reforms and gender inequalities we present in this paper is partially and based on the sole I.Stat data warehouse that provides a first and aggregated information on the labour forces. The gender effects of the long period of reforms that involves the job market in Italy are variegated and meet the legislator expectations partially and mainly towards the last years. They are undoubtably vast the potentialities of research that need to be investigated and require more data and an appropriate statistical method to discover correlations between employment trends and reform effects. Transition matrices and longitudinal data will help the continuation of our work.

Fig. 1 Italian employees by gender. Thousand of individuals. 1st Quarter 2004 – 4th Quarter 2020.



References

1. Battisti, M., Vallanti, G.: Flexible wage contracts, temporary jobs, and firm performance: Evidence from Italian firms. Industrial Relations: A Journal of Economy and Society, 52, 737–764 (2013)
2. Boeri, T., Garibaldi, P.: Two tier reforms of employment protection: A honeymoon effect? The Economic Journal, 117, 357–385 (2007)
3. Cirillo, V., Fana, M., Guarascio, D.: Labour market reforms in Italy: evaluating the effects of the Jobs Act. Economia Politica 34, 211–232 (2017)
4. Marini C., Nicolardi V.: Livelli e probabilità di transizione del mercato del lavoro: alcune evidenze degli effetti del JOBS ACT sull'occupazione in Italia, in Economia, Istituzioni, Etica e Territorio. Casi di studio ed esperienze (a cura di Toma E.), pp. 55-78, Franco Angeli ed. (2018)
5. Saltari, E., Travaglini, G.: Il rallentamento della produttività del lavoro e la crescita dell'occupazione. Il ruolo del progresso tecnologico e della flessibilità del lavoro. Rivista italiana degli economisti, 13, 3–38 (2008)

Intergenerational transmission of disadvanteges in the Italian labour market: evidence from AD-SILC data

Trasmissione intergenerazionale delle diseguaglianze: evidenze dai dati AD-SILC

Annalisa Busetta, Elena Fabrizi, Giancarlo Ragozini and Isabella Sulis

Abstract Family's socioeconomic conditions influence children educational choices and, indeed, their future position in the labour market. This paper aims at disentangling the role that family socioeconomic status plays in determining differences in individual educational attainment and in the labour market outcomes in terms of wages, adopting a Generalized Path Analysis. The modelling approach incorporates Heckman's selection model to take into account the presence of selection bias due to participation or not in the labour market. The study has been carried out using the innovative and rich ADSilc database, which allows us to follow up the evolution of occupational outcomes over time (at the entrance and after ten years) and provides information on an individual's family socioeconomic background.

Abstract *Le condizioni socioeconomiche della famiglia influenzano le scelte di istruzione dei propri figli e la loro posizione futura nel mercato del lavoro. Questo lavoro si propone di scorporare il ruolo diretto e indiretto che lo stato socioeconomico della famiglia svolge nel deeterminare le differenze nel livello di istruzione raggiunto dagli individui e nel loro futuro collocamento nel mercato del lavoro adottando un modello di Path Analysis generalizzato. L'approccio di modellazione incorpora il modello di selezione di Heckman e la presenza di fattori latenti non osservati che influenzano la probabilità di lavorare e quindi di percepire un salario. Lo studio è stato condotto utilizzando l'innovativo e ricco database ADSilc, che permette di seguire l'evoluzione degli esiti occupazionali nel tempo (all'ingresso e*

Annalisa Busetta
University of Palermo, e-mail: annalisa.busetta@unipa.it

Elena Fabrizi
University of Teramo, e-mail: efabrizi@unite.it

Giancarlo Ragozini
University of Naples Federico II, e-mail: giancarlo.ragozini@unina.it

Isabella Sulis
University of Cagliari, e-mail: isulis@unica.it

dopo dieci anni) e prevede informazioni sul background socioeconomico familiare dell'individuo.

Key words: AD SILC, Path Analysis, Heckman Selection Model, latent variables, intergenerational transmission inequalities.

1 Introduction

The equity of a country is measured by considering how the economic resources, the access to high educational level and to apical positions in the society is related to family socioeconomic background. It is well known that a greater family of origin socioeconomic status provides greater guidance in the building up of a strong educational background (e.g. throughout the selection of institutions during the primary and low secondary studies, support in the daily activities, addressing university choices). Nonetheless, the gap in the job market positions between individuals with similar educational backgrounds provides evidence that the family of origin influences the economic returns on the job market by ensuring direct and indirect support (e.g. financial and social support in the subsequent search for a job, allowing individuals the access to larger job markets in geographical terms and/or opening the access to those positions which requires initial financial investments) [2, 9, 3]. In this context, families adopt strategies addressed to building up credentials that individuals can spend in the job market for reaching better positions and/or offering financial support at the entrance (as happens for university mobility and the achievement of postgraduate qualifications, etc). The interactions of all these factors contribute to the amplification of inequalities [5, 7] between individuals coming from different socioeconomic backgrounds. Thus, family supports affect the future career, producing direct returns in terms of socioeconomic positions and higher earnings, but also indirect effects, manifested by the wage differences that continue to persist among workers who achieve the same educational level (i.e. wage persistence gap) [10, 2, 9, 3, 4]. Moving from this framework the main aim of this work is to test the roles played by socioeconomic conditions of the family of origin in predicting children's achievement of high level of education and subsequent outcomes in the labour market.

Namely, the main research questions that are addressed in this study are:

- RQ1 Is there a direct effect of family socioeconomic status on individual educational achievement?
- RQ2 Is there any effect (direct and indirect) of the family of origin on individual wage persistency gap along the work history pattern?

Findings rely on the AD-SILC database 2011, which has been merged with the INPS database to follow the evolution of occupational outcomes in terms of wage at the entrance and after ten years. The database provides information on individual's family socioeconomic background. To tackle such a complex system of relationships across the involved exogenous, mediators and endogenous variables, we adopt a generalized path analysis which incorporates the Heckman selection model to take into account the presence of selection bias due to the participation or not in the labour market. Our findings indicate the strong presence of intergenerational transmission of inequalities and wage persistence gap in all cohorts: being raised in an advantaged family not only positively affects the level of education achieved, but also the subsequent divergences observed in terms of wages.

2 Data

We use the AD-SILC database, which is constructed to match longitudinal information from administrative archives, held by the National Institute of Social Security (INPS), with survey data collected by the National Institute of Statistics (ISTAT). In our database, we have information on the socioeconomic conditions of the interviewed (from the 2011 IT-SILC survey) and their individual working career histories (collected in the administrative archives) up to 2018. Moreover, the 2011 IT-SILC special module on intergenerational transmission of disadvantage provides information on their parents when the individual was 14 years old (i.e., education, occupation and difficulties to make ends meet). We selected only cohorts of offspring aged respectively 30-34 years, 35-39, 40-44 and 45-49 years in 2011, excluding those who have not yet completed their education. Three dependent variables were considered: years of education (Y), the daily wage at the entrance into the job market (W) and the daily wage after 10 years (W_2). The main independent variables are those regarding the social origin (e.g. socioeconomic status of both parents). An index of family socioeconomic status has been built up (adopting a Graded Response Model) and used as a predictor in the regression setting by considering parents' education, their occupation and the family's ability to make ends meet when the individual was 14 years old. The occupational outcome W at the entrance and after 10 years was measured by the gross hourly wages paid in cash for time worked or for time not worked, such as holidays, as well as additional payments (e.g. overtime payments, benefits). The analysis is bounded to a comparison between individuals who never enter the labour market (883) and those who enter the labour market and remain at work for ten years (7,986 employees). We have excluded from the analysis those who entered the labour market discontinuously. A set of control variables have been used to take into account the differences in reproductive behaviour and other information related to the career.

3 Modelling approach

The classical path analysis model used to test intergenerational reproduction of inequalities [2, 3] has been generalized to take into account the selection bias related to the censoring of the variable wage for individuals who do not enter the labour market [6, 11]. In order to achieve this goal the path model analysis [1, 8] - which specifies three concatenated regression models to model the achieved educational levels (Y) in equation 1, the daily wage at the entrance (W_1), in equation 2, and the daily wage after 10 years (W_2) - has been generalized specifying in equation 2 a Heckman selection model. This last consists of a system of two equations which allow taking into account that wage is censored, for all those individuals who do not work at the time considered (at the entrance and after 10 years). The censored equation related to the outcome variables W (considered at the two times) has been specified by jointly modelling the level of wages (W specified in logartimic term) and the probability to work or not as a function of two external sets of predictors. The Generalized Path model is depicted in Figure 1

In equation (1) X is the set of sociodemographic covariates (i.e. the family of origin's socioeconomic status, geographical area and cohort and gender) which influences the level of education (Y). In equation 2, wages levels (W_1) are specified as a function of the individual level of education (Y), the family of origin's socioeconomic status (X) and the individual's characteristics related to professional and reproductive choices (e.g. work experience, gender, marital status) (M) and the latent variable (L). The latter influences the probability to be in the labour market, modelled with a probit model. In equation (2) the latent variable L is constrained to 1, as the value of its variance. The error terms are normally distributed, $\varepsilon_1 \sim N(0, \sigma_1^2)$, $\varepsilon_2 \sim N(0, \sigma_2^2)$ and $\varepsilon_3 \sim N(0,1)$, and ε_2 and ε_3 are allowed to be correlated with $cor(\varepsilon_2, \varepsilon_3) = \rho$, in order to take into account that there are censored information with respect to wage, as the propensity to work or not influences the observed wage. Both outcome variables related to the labour market outcomes (Selected and Daily wage) are correlated since both depend on the latent variable (L). The latter captures the association related to the existence of a common factor affecting the propensity to work and the wage, that needs to be taken into account in order to have unbiased estimates of the effect of education and family of origin's socioeconomic status on daily wage.

4 Preliminary results

Preliminary results show a clear strong (direct and indirect) effect of the family of origin not only on kids' education, as already largely confirmed in the literature, but on the labour market outcomes: the family can exert an important influence on career assuring better-payed jobs, as it is shown by the magnitude of the coefficient which describes the indirect effect that the family of origin exerts on wages. After ten years, employees coming form families with higher socioeconomic status have on average higher wages, confirming the presence of a wage persistence gap among individuals who achieved the same level of education but come from families with different socioeconomic status. If we focus on the indirect effect of the family socioeconomic status it arises that individuals whose families are located in the highest quartile of the socioeconomic status index earn on average 10% more than those coming from the most disadvantaged ones. This evidence persists over a long period. If we look at gender differences, model results confirm that men on average achieve lower educational levels than women, but on average have higher wages (+10% at the entrance) and this difference increases in the long period (+ 28%). The selection equation, which models the probability to work as function of the geographical area, the level of education, the reproductive choice before entrance into the job market, and gender, suggests that individuals residing in the South have a lower probability to work, as women with at least one child. The indirect effect of coming from a wealthy family is the strongest observed in terms of magnitude and does not change significantly between entry and the following ten years.

The latent variable L, as expected, negatively affect the initial wage and the probability to work, allowing to correct the direct and indirect effect of predictors on wage.

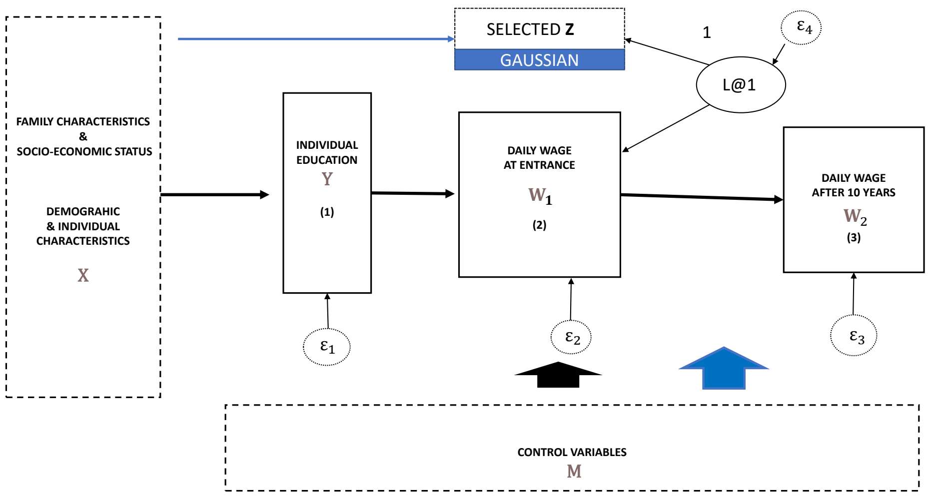


Fig. 1 Path analysis with Heckman model correction for intergenerational reproduction of inequality

References

1. Acock, A. C.: Discovering structural equation modeling using Stata. Stata Press Books, Texas (2013)
2. Bernardi, F., Ballarino, G. (Eds.): Education, occupation and social origin: A comparative analysis of the transmission of socio-economic inequalities. Edward Elgar Publishing, Cheltenham (2016)
3. Busetta, A., Fabrizi, E., Sulis, I., Ragozini, G.: Does family of origin make a difference in occupational outcomes? In Book of Short Papers SIS 2022-51th Scientific meeting of the Italian Statistical Society, 134-143. Pearson, Milano (2022)
4. Busetta, A., Fabrizi, E., Sulis, I., Ragozini, G.: Mobilità sociale delle famiglie. In Rapporto sulla popolazione. Le famiglie in Italia. Forme, ostacoli, sfide, 207-234. Il Mulino, Milano (2023)
5. Cattaneo, M., Horta, H., Malighetti, P., Meoli, M., Paleari, S.: Effects of the financial crisis on university choice by gender. Higher Education, 74, 775-798 (2017)
6. Heckman, J. J.: The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5, 475–492 (1976)
7. Impicciatore, R., Tosi, F.: Student mobility in Italy: The increasing role of family background during the expansion of higher education supply. Research in Social Stratification and Mobility, 62 (2019)
8. Kline, R. B.: Principles and practice of structural equation modeling. Guilford publications, New York (2015)
9. Raitano, M., Vona, F.: From the cradle to the grave: The influence of family background on the career path of Italian Men. Oxford Bulletin of Economics and Statistics, 80(6), 1062-1088 (2018)
10. Raitano, M., Vona, F.: Measuring the link between intergenerational occupational mobility and earnings: evidence from eight European countries. The Journal of Economic Inequality, 13, 83-102 (2015)
11. StataCorp: Structurial equation modeling reference manual release 18. College Station, Texas (2023)

Solicited Session SS10 - *Statistical methods and complexity for evaluation in finance*

Organizer and Chair: Roy Cerqueti

1. *Financial networks resilience and shocks propagation* (Cerqueti R., Cinelli M., Ferraro G. and Iovanella A.)
2. *How the choice of one parameter impacts the numerical stability of the efficient frontier* (Fassino C. and Uberti P.)
3. *Dynamic shrinkage for minimum variance combination of forecasts* (Mattera R.)
4. *Exploring the perception of the gender issue of Italian female entrepreneurs.* (Castellano R., Riccioni J. and Rinaldi A.)

Financial networks resilience and shocks propagation

Resilienza delle reti finanziarie e propagazione degli shock

Roy Cerqueti, Matteo Cinelli, Giovanna Ferraro and Antonio Iovanella

Abstract In the context of network modelling and analysis, especially as it relates to finance, the idea of resilience—i.e., the capacity of a cohesive system to absorb shocks—is highly relevant. This paper starts from this premise, and deals with a general definition of resilience measure. Moreover, we apply our methodological proposal to the resilience of a financial interbanking system. To achieve this, we build a financial network model related to the quarterly-based interbanking sector, whose weights are calibrated on high quality empirical data; next, we compute the resilience measure of the considered networks, taking into full consideration the influence of the network topology and the weights of its links in the shocks propagation. We show that our proposal is a generalization of other resilience measures. The findings are discussed in light of both the financial and network theory perspectives.

Abstract *Nel contesto della modellazione e dell'analisi delle reti, in particolare nel caso di quelle finanziarie, è di grande rilevanza il concetto di resilienza, ovvero la capacità di un sistema di assorbire shock. A partire da questa premessa, l'articolo introduce un concetto generale di resilienza. Inoltre, la metodologia proposta è applicata nel caso della resilienza del sistema finanziario interbancario. A questo fine è stato costruito un insieme di reti finanziarie su base trimestrale facenti riferimento*

Roy Cerqueti
Department of Social and Economic Sciences, Sapienza University of Rome, Italy,
e-mail: roy.cerqueti@uniroma1.it

Matteo Cinelli
Department of Computer Science, Sapienza University of Rome, Italy
e-mail: matteo.cinelli@uniroma1.it

Giovanna Ferraro
Department of Engineering and Architecture, University of Parma, Italy
e-mail: giovanna.ferraro@unipr.it

Antonio Iovanella
School of Economics, Università degli Studi Internazionali di Roma - UNINT, Italy
e-mail: antonio.iovanella@unint.eu

al settore interbancario e i cui parametri sono stati calibrati su dati empirici raccolti da dataset ufficiali; successivamente, sono stati calcolati i valori di resilienza per le suddette reti prendendo in considerazione l'influenza della loro topologia e dei parametri di propagazione degli shock. Mostriamo, inoltre, che la nostra proposta generalizza altri tipi di misure della resilienza presenti in letteratura. I risultati sono stati discussi alla luce sia della teoria delle reti che della prospettiva finanziaria.

Key words: Financial interbanking networks, Network modelling and analysis, Resilience, Shock propagation, Interconnectedness, Bank for International Settlements

1 Introduction

Assessing the ability of a system to absorb shocks is a very relevant task. Resilience in complex networks was theoretically explored starting with the seminal paper by Albert et al. [1], expanded by [8] and then taking into account the effects of various node removal strategies on the network structure [6].

For the special theme of the resilience, it is natural to capture the propagation of a shock over a system as a transition from different sites through their connecting links. In the scientific literature, the study of resilience primarily relates to the analysis of a performance metric, typically the diameter of the network, under two different stress schemes depending on a targeted attack to a specific node or considering the random failure of a node, viewed as a network error [1]. These analysis are relatively easy to conduct and this explains why the models of resilience of a system are set in the context of complex networks [6].

In this study, the resilience is evaluated by taking into account a novel measure based on shocks that arise in one of the nodes and its propagation through the network's links. Specifically, we extend the resilience measure in [3]. We assume that the propagation of a shock from a node to another one occurs over the shortest paths connecting the nodes. In so doing, we develop a model where the nodes are immunized from the subsequent effects of the shock, once they are infected for the first time by its propagation. This condition is particularly reasonable in several real-world situations and reduces the computational complexity of similar procedures, such as the one presented in [4] where all simple paths are considered. This requirement lowers the computational complexity of related procedures, like the one shown in [4] where all simple paths are taken into account, and is particularly reasonable in a number of real-world scenarios.

Due to these considerations, the proposed resilience measure outperforms [4] in terms of applicability and extend the one presented in [3], generalizing the shock propagation function. According to [7], the computation time for all simple paths increases exponentially with network size, so only small networks can be analysed in a reasonable amount of time. Importantly, we make the assumption that the shock's propagation is appropriately rescaled as the distance from the shocked nodes rises in accordance with [4] . By doing this, we offer a versatile device that incorporates either amplified or dampened effects of the distance from the shocked node, where the type of effects must be determined based on the situation being thought of in real life.

The methodological proposal is tested over the financial context, taken from the interbanking system and data are retrieved by the BIS web site database [2]. In particular, we refer to the activities associated to the consolidated banking statistics (CBS) of all the available countries which capture the worldwide consolidated positions of internationally active banking groups headquartered in reporting countries. The CBS include the business of banks' foreign affiliates but exclude intragroup positions, similarly to the consolidation approach followed by banking supervisors. Each banking group is headquartered in a country, so that banks are intuitively associated to countries.

The paradigmatic relevance of the financial context, particularly when one deals with the idea of resilience, is what drove this decision. In fact, accounting for shocks propagation in finance helps to explain a crucial aspect of financial systems, which has been put to the test by a number of paradigmatic cases, such as the financial contagion linked to the bankruptcy of Lehman Brothers [5].

2 The resilience measure

This section illustrates our resilience measure, on the basis of the case proposed by [3]. We consider a set $V = \{1, \ldots, n\}$, collecting the nodes of a complex directed and weighted network. The adjacency matrix is denoted by $\mathbf{W} = (w_{ij})_{i,j \in V}$. We assume that $w_{ij} > 0$ if and only if there exists a directed link between nodes i and j. Given two nodes $i_0, i_k \in V$, we introduce the shortest path with starting point i_0 and terminal point i_k as

$$p^{min}_{i_0,i_k} = \text{argmin} \left\{ \sum_{h=1}^{k} w_{i_{h-1}i_h} : p_{i_0,i_k} \in \bigcup_{h=1}^{\bar{k}} \mathscr{P}^{(k)} \right\}. \tag{1}$$

where $\mathscr{P}^{(k)}(i_0)$ is the collection of the k simple paths whose starting node is i_0.

The shocks are local events occurring to the nodes. The entity of the shock is captured by a scalar $\xi \in (0, +\infty)$; it grows – i.e., the effect of the shock becomes stronger – as the value of ξ increases. The shock starts from a given node $i_0 \in V$ and propagates over the shortest paths with initial value i_0. We consider the propagation on a generic k-path $p^{(k)}_{i_0}$ whose h-th link has weight $w_{i_{h-1}i_h}$, and define a discount factor $\delta \in [0, +\infty)$, that captures the effect of the distance on the propagation as follows:

$$\xi_h = \xi \cdot f(w_{i_0 i_1}, \ldots, w_{i_{h-1}i_h}, \delta) \tag{2}$$

where f is a generic function of the weights and the discount factor, ξ_h is the entity of the shock at node i_h, for each $h = 0, \ldots, k$ and $\xi_0 = \xi$.

Importantly, the dynamics in (2) represents a generalization of the specific shape of function f in [3].

We also include a propagation condition: if the shock is too weak, then the propagation motion stops. To formalize the propagation condition, we introduce a vector $\Gamma = (\gamma_1, \dots \gamma_{\bar{k}}) \in (0, +\infty)^{\bar{k}}$ and assume that the existence of $s = 0, 1, 2, \dots, h-1$ such that $\xi_s < \gamma_s$ prevents the propagation of the shock to node i_h. The validity of the propagation conditions clusters the k-paths in two classes: the ones where the shock ξ propagates and those where such a shock does not achieve the last node. The clusters depend on Γ. We denote by $\mathscr{P}^{(k)}_{\Gamma,\xi}$ the set collecting the k-paths – not necessarily the shortest k-paths – which satisfy the propagation condition for Γ and ξ.

Given the vector Γ and the shock with entity ξ, we define the $\Gamma - \xi$-resilience measure of the network $N = (V, E)$ as

$$\mu_{(\Gamma,\xi)}(N) = 1 - \sum_{k=1}^{\bar{k}} \theta_k \frac{|\mathscr{P}^{(k)}_{\Gamma,\xi} \cap \mathscr{P}_{short}|}{|\mathscr{P}^{(k)} \cap \mathscr{P}_{short}|}, \tag{3}$$

where $\theta_1, \dots, \theta_{\bar{k}}$ are selected to be nonnegative weights such that $\sum_{k=1}^{\bar{k}} \theta_k = 1$. We denote the vector of the weights by $\Theta = (\theta_1, \dots, \theta_{\bar{k}})$. The selection of Θ is implemented on the basis of the specific attention towards the different lengths of the paths when measuring the resilience of the network.

We observe that $\mu_{(\Gamma,\xi)}(N) \in [0,1]$ and when $\mu_{(\Gamma,\xi)}(N) = 1$ ($\mu_{(\Gamma,\xi)}(N) = 0$, resp.), then the shock ξ is absorbed (the shock ξ propagates over all the paths, resp.).

3 Results

We illustrate the outcomes of the resilience measure when dealing with

$$f(w_{i_0 i_1}, \dots, w_{i_{h-1} i_h}, \delta) = \sum_{s=1}^{h} w_{i_{s-1} i_s} \delta^{h-s+1}.$$

The methodological proposal is tested over the financial context, taken from the interbanking system and data are retrieved by the BIS web site database. In particular, we refer to connections weighted by the real amount of interbanking consolidated (CBS, Consolidated Banking Statistics) exposures among reporting countries. We present quarterly data, ranging from the first quarter of 2005 to the fourth quarter of 2020. Thus, we have 64 networks with number of nodes ranging from 202 to 215. The number of nodes and links varies over the periods, from 202 to 215 and from 1922 to 2882, respectively.

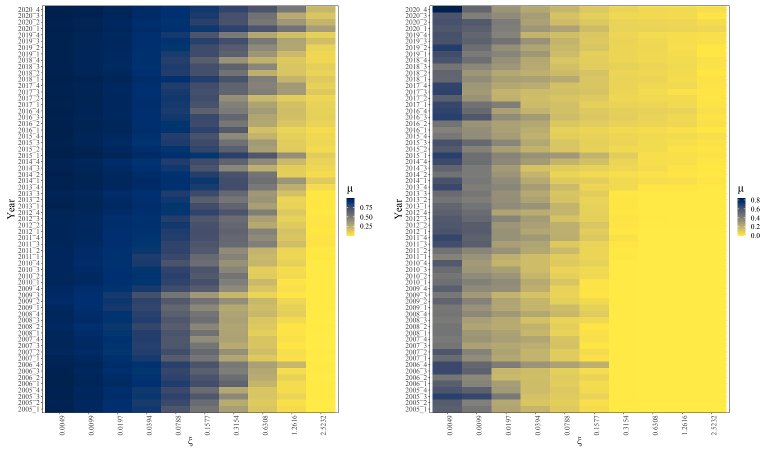


Fig. 1 Measures for the $\Gamma - \xi$-resilience $\mu_{(\Gamma,\xi)}(N)$ in the case of BIS for $\Gamma = \{\gamma_i = 1, i = 1,...,\bar{k}\}$ and $\Theta = \{\theta_i = 1/\bar{k}$, for $i = 1,\ldots,\bar{k}\}$. Different values of ξ are reported on x-axis (as index of vector $\xi(i)$), years on y-axis and values of $\mu_{(\Gamma,\xi)}(N)$ with a colour palette ranging from blue ($\mu_{(\Gamma,\xi)}(N) = 1$) to yellow ($\mu_{(\Gamma,\xi)}(N) = 0$). Left panel: $\delta = 0.25$; right panel: $\delta = 2$.

Differently from [3], we illustrate the regions of resilience over the years through a suitable heat map visualization. In doing so, we extend and complement the empirical approach proposed by [3].

Figure 1 shows the dependence of the resilience on the size of the shock restricted to the values of $\delta = \{0.25, 2\}$, where the first value captures the case of shock reduction and the second value the case of shock amplification, and $\xi_i = (w_{min} * 2^{10-i})^{-1}$ with $i = 1,\ldots,10$ and w_{min} as the lowest link weight over all the 64 instances. In the whole simulations we investigated a parameter space considering $\delta \in \{0.1, 0.25, 0.5, 0.75, 1, 2, 5, 10, 20, 50\}$

4 Discussion

From a purely financial perspective, the BIS networks' prior 2010 lack of resilience can be attributed to the financial system's idiosyncratic fragility, which became apparent after Lehman Brothers' failure. The significant difference between the network's resilience before and after the 2008 financial crisis appears to be even more obvious as δ increases, making it simpler for shocks to propagate.

Moreover, there is a collapse of the resilience around year 2014, after the period 2009-2013 of low resilience. The low resilient period 2009-2013 corresponds to the one immediately following the Lehman Brothers' bankruptcy; hence, one can argue that the lowest level of interbanking exposures among countries is associated to a low capability of the network to absorb shocks in a post-financial crisis starting event period. The financial turmoil that began in 2014, including the collapses of the Russian and Chinese stock markets and the poor performance of the global stock markets in 2018, is associated to the low levels of resilience in that time period. This result is supported by the high degree of resilience observed after 2014.

References

1. Albert, R., Jeong, H., Barabási, A-L.: Error and attack tolerance of complex networks. Nature 406, 378–382 (2000)
2. BIS database, www.bis.org.
3. Cerqueti, R., Cinelli, M., Ferraro, G., Iovanella, A.: Financial interbanking networks resilience under shocks propagation. Annals of Operations Research, 1-21 (2022)
4. Cerqueti, R., Ferraro, G., Iovanella, A.: Measuring network resilience through connection patterns. Reliab Eng Syst Safe. 188, 320–329 (2019)
5. Dumontaux, N., Pop, A.: Understanding the market reaction to shockwaves: evidence from the failure of Lehman Brothers. Journal of Financ Stab. 9(3), 269–286 (2013)
6. Ferraro, G., Iovanella, A.: Clairvoyant targeted attack on complex networks. Int J Comput Econ Ec. 8(1), 41–62 (2016)
7. Fortune S, Hopcroft JE, Wyllie J.: The directed subgraph homeomorphism problem. Theor Comput Sci.10, 111–21 (1980)
8. Gao, J., Barzel, B., Barabási, A-L.: Universal resilience patterns in complex networks. Nature 530(7590), 307–312 (2016)

How the choice of one parameter impacts the numerical stability of the efficient frontier

L'impatto della scelta di un parametro sulla stabilità numerica della frontiera efficiente

Claudia Fassino and Pierpaolo Uberti

Abstract The expected return of the portfolio is a parameter that can be arbitrarily chosen by the investor when the vectors that describe the linear restrictions of the optimization problem are linear independent. When working on real data, it often happens that these vectors are numerical almost linear dependent. In this framework, the choice of an opportune value for portfolio's expected return becomes a fundamental instrument to guarantee the stability of the optimal solution. On the basis of geometrical arguments, specifically through the use of the concepts of numerical rank and condition number of a matrix, we propose an interval of feasible values of the parameter. The effectiveness of our proposal in reducing numerical instability is supported by an extensive application on real financial data.
Abstract *Il rendimento atteso del portafoglio può essere scelto arbitrariamente quando i vettori che descrivono i vincoli lineari del problema di ottimizzazione sono linearmente indipendenti. Quando si lavora su dati reali, capita spesso che questi vettori siano numericamente vicini ad essere linearmente dipendenti. In questo contesto, la scelta del rendimento atteso del portafoglio diventa fondamentale per garantire la stabilità della soluzione ottima. Sulla base di argomentazioni di tipo geometrico-algebrico, specificamente utilizzando i concetti di rango numerico di una matrice e numero di*

Claudia Fassino
University of Genoa, Department of Mathematics, e-mail: fassino@dima.unige.it

Pierpaolo Uberti
University of Milano-Bicocca, Department of Statistics and Quantitative Methods, e-mail: Pierpaolo.uberti@unimib.it

condizionamento, proponiamo un intervallo di possibili valori per la scelta del parametro. La bontà della nostra proposta è supportata da una serie di applicazioni su dati reali.

Key words: Portfolio Optimization, Numerical Stability, Condition Number

1 Introduction

A model is considered unstable when infinitesimal variations of the inputs produce huge variations in the outputs. Markowitz model, see [10], is known not only as the referring point of modern portfolio theory but also for the numerical instability of the optimal allocation. This issue makes the implementation of the model very difficult in practice and it is probably the principal reason for its limited use by the investors.

Moreover, the numerical instability of the model is probably the reason of its poor out-of-sample performance. In a very famous paper, see [2], the authors show that the equally weighted portfolio perform better than the optimal mean variance portfolio. This fact is caused by model's uncertainty that depends on the parameters estimation, see [11].

In the literature many different approaches for model's instability have been proposed. Alternative estimation techniques for the covariance matrix are suggested; for example, see [5] for a Bayesian approach, see [9] for the so-called shrinkage estimation, see [1] for lasso techniques. A parallel branch of research focuses on robust optimization, see [12], [3], [8], [13] for a detailed review of the existing literature.

Our proposal differs from the standard approaches enumerated above because we focus on the numerical instability depending on the linear restrictions of the optimization problem, see [4]. We show, through a numerical example on real financial data, that it exists an interval of values for portfolio expected return where the condition number of the coefficients matrix of the linear restrictions is not smaller than the condition number of the augmented matrix. If the value of portfolio expected return is chosen in this interval, the correspondent optimal allocation is numerical stable.

2 The main idea

We assume that $n \geq 2$ is the number of risky assets on a given market; a portfolio is represented by a column vector $x \in R^n$ where the entry x_i is the percentage of initial wealth invested in asset i. The column vector $\mu \in R^n$ contains the expected returns of the n assets, $u \in R^n$ is the vector with unitary entries and μ_p is portfolio's expected return.
We refer to the following portfolio optimization problem,

$$\begin{aligned} &\text{Minimize}_x \; x^\top I_n x \\ &\text{subject to } \mu^\top x = \mu_p \end{aligned}$$

$$u^\top x = 1$$
$$x_i \geq 0 \text{ for } i = 1, \dots, n.$$

We substitute the covariance matrix of the standard mean-variance formulation with the identity matrix I_n in order to highlight the role of the linear restrictions of the optimization problem as the principal cause of numerical instability in the model. Moreover, we note that the objective function $x^\top I_n x$ is equal to the so-called Herfindhal Hirschman index, $HHI = \sum_n x^2$, a well-known measure of concentration, see [6,7]. The equality restrictions of the optimization problem can be written as $Bx = \left[\mu_p 1\right]^\top$, where $B = [\mu\ u]^\top$. We define $x^*(\mu_p)$ the solution of the optimization problem for a given μ_p; then the efficient frontier (F) is the set containing the solution $x^*(\mu_p)$ of the optimization problem when μ_p varies:

$$F = \left\{x^*\left(\mu_p\right) \in \mathrm{R}^{\mathrm{n}}, \forall \mu_{\mathrm{p}} \in [\min(\mu), \max(\mu)]\right\}.$$

In order to measure how the choice of μ_p impacts the stability of the solution we introduce portfolio TurnOver (*TO*), a measure of the variation of the optimal solution $x^*(\mu_p)$ with respect to μ_p

$$TO_\varepsilon\left(\mu_p\right) = \frac{\sum_{i=1}^{n}\left|x_i^*\left(\mu_p\right) - x_i^*(\mu_p + \varepsilon)\right|}{n},$$

where $\varepsilon > 0$ is a small arbitrary positive number describing the variation of μ_p. We underline that quantity *TO* has a natural financial interpretation as the cost that is needed to adjust portfolio's allocation in practice. For this reason, the *TO* provides a significant approximation of the transaction costs to implement the investment strategy. Moreover, since the optimization problem described above needs to be solved through a numerical approach, we consider the number of iterations that are required to MatLab quadprog function to converge as a further measure of instability.

As explained in the abstract, when working on real data it can happen that the vectors μ and u are numerical almost collinear. Numerical collinearity can be measured using the concept of numerical rank of a matrix: given a positive arbitrary parameter δ, the numerical rank of a matrix is given by the number of its singular values that are above δ. When μ and u are close to be linear dependent, the *B* matrix is a numerical rank deficient matrix.

Using a convenient notation, *B* can be rewritten as

$$B = \begin{bmatrix} \beta u^\top + \rho_\mu^\top \\ u^\top \end{bmatrix} \text{ with } \beta = \frac{\mu^\top u}{n}.$$

We also define the matrix $\hat{B}$ as

$$\hat{B} = \begin{bmatrix} \hat{\mu}^\top \\ \hat{u}^\top \end{bmatrix} \text{ with } \hat{\mu}^\top = \left(\mu^\top, \mu_p\right) \in R^{(n+1)}, \hat{u}^\top = (u^\top, 1) \in R^{(n+1)}.$$

Given $\delta > 0$, the idea is to choose a value for μ_p such that the numerical rank of *B* and $\hat{B}$

are equal. In particular, we want to choose μ_p in order to avoid the situation in which the numerical rank of $\hat{B}$ is greater than the numerical rank of B. As it will be clear in the following example, this choice of μ_p has interesting consequences on the stability of the solution of the optimization problem.

The following example on real financial data describes our idea. The real data application is performed on a database, called *S*&*P*500*sectors* that is composed by the daily returns from 1/3/2000 to 9/17/2020 of the $n = 10$ sectors portfolios of the S&P500 index obtained using the Global Industry Classification Standard (GICS): Energy, Material, Industrials, Consumer-Discretionary, Consumer-Staples, Healthcare, Financials, Information-Technology, Telecommunications and Utilities.

The matrix B is

$$B = \begin{bmatrix} \mu^\top \\ u^\top \end{bmatrix}$$

with $\mu^\top = 10^{-4}(5.21,4.08,3.07,1.9,3.68,3.32,4.06,3.44,1.43,2.79)$.

Then $\beta = 3.2980 \cdot 10^{-4}$ and $\|\rho_\mu\|^2 = 3.2815 \cdot 10^{-4}$. Choosing $\delta = 3.4 \cdot 10^{-4}$, B is numerical rank deficient since $\|\rho_\mu\|^2 < \delta^2(1+\beta^2)$. Furthermore, since $\frac{\delta^2}{n} = 1.156 \cdot 10^{-8} \ll 1$, we obtain

$$R \approx \sqrt{\frac{n+1}{n}} \sqrt{\delta^2(1+\beta^2) - \|\rho_\mu\|_2^2} = 9.3317 \cdot 10^{-5}.$$

We propose to restrict μ_p to the interval $[\beta - R, \beta + R] = 10^{-4} \cdot [2.365,4.231]$, so that the matrix

$$\hat{B} = \begin{bmatrix} \mu^\top & \mu_p \\ u^\top & 1 \end{bmatrix}$$

has numerical rank equals to 1.

Table 2 also highlights that $K_2(\hat{B})$ is greater than $K_2(B) = 9.636 \cdot 10^3$ and that both matrices are ill-conditioned, where K_2 is the condition number.

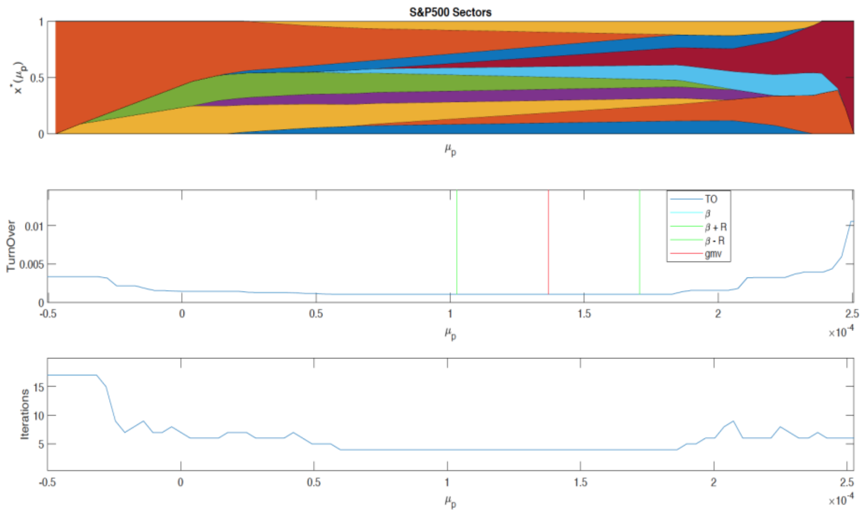
$\mu_p = 10^{-4}\cdot$	2.365	2.676	2.987	3.298	3.609	3.920	4.231
$\|\rho_{\hat{\mu}}\| = 10^{-4}$	3.400	3.335	3.295	3.282	3.295	3.335	3.400
$K_2(\hat{B}) = 10^3\cdot$	9.755	9.946	10.066	10.107	10.066	9.946	9.755

Table 1 Different values of μ_p and the associated $\|\rho_{\hat{\mu}}\|^2$ and $K_2(\hat{B})$.

The bold entries in Table 2 correspond to the case $\mu_p = \beta$. In Figure 1 we report a graphical representation of the composition of the optimal portfolio $x^*(\mu_p)$ as function of μ_p, of the *TO* and of the number of iterations that quadprog needs to converge to the optimal solution.

The results depicted in Figure 1 are clear: when the value of μ_p is close to β, the solution of the optimal allocation problem is numerical stable. The composition of the optimal portfolio shows small variations, the *TO* is low and the number of iterations for the algorithm to converge is limited. On the opposite, when the values of μ_p is far from β, the optimal portfolio shows significant numerical instability. This evidence is also supported by a growing level of *TO* and of the number of iterations needed to converge to the optimal solution.

Fig. 1 Composition of the optimal portfolio $x^*(\mu_p)$, *TO* and number of iterations as functions of μ_p.



3 Conclusions

The empirical evidence of the example on real financial data shows how the choice of the parameter μ_p strongly impacts the numerical stability of the solution of the optimal allocation model. In future research we will investigate how to restrict the choice of μ_p in a suitable interval in order to obtain a numerical stable model.

References

1. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., Loris, I.: Sparse and stable Markowitz portfolios. Proceedings of the National Academy of Sciences 106, 12267–12272 (2009)
2. DeMiguel, V., Garlappi, L., Uppal, R.: Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The review of Financial studies 22, 1915–1953 (2009)
3. Fabozzi, F.J., Huang, D., Zhou, G.: Robust portfolios: contributions from operations research and finance. Annals of operations research 176, 191–220 (2010)
4. Fassino, C., Torrente, M.L., Uberti, P.: A singular value decomposition based approach to handle ill-conditioning in optimization problems with applications to portfolio theory. Chaos, Solitons & Fractals 165, 112746 (2022)
5. Frost, P.A., Savarino, J.E.: An empirical bayes approach to efficient portfolio selection. Jour- nal of Financial and Quantitative Analysis 21, 293–305 (1986)
6. Herfindahl, O.: Concentration in the U.S. steel industry, Unpublished Doctoral Dissertation, Columbia University, (1950)
7. Hirschman, A.O.: The paternity of an index, The American Economic Review 54, 761–762 (1964)
8. Kim, J.H., Kim, W.C., Fabozzi, F.J.: Recent developments in robust portfolios with a worst-case approach. Journal of Optimization Theory and Applications 161, 103–121 (2014)
9. Ledoit, O., Wolf, M.: Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of empirical finance 10, 603–621 (2003)
10. Markowitz, H.: Portfolio selection. Journal of Finance 7, 77–91 (1952)
11. Pflug, G.C., Pichler, A., Wozabal, D.: The 1/n investment strategy is optimal under high model ambiguity. Journal of Banking & Finance 36, 410–417 (2012)
12. Pflug, G.C., Pohl, M.: A review on ambiguity in stochastic portfolio optimization. Set-Valued and Variational Analysis 26, 733–757 (2018)
13. Xidonas, P., Steuer, R., Hassapis, C.: Robust portfolio optimization: a categorized biblio- graphic review. Annals of Operations Research 292, 533–552 (2020)

Dynamic shrinkage for minimum variance combination of forecasts

Shrinkage dinamico per la combinazione di previsioni a varianza minima

Raffaele Mattera

Abstract Forecasting accuracy can be improved by combining predictions obtained with different methods. The minimum-variance approach assigns a specific weight to each competing forecasting method with the aim of reducing the forecasting error variance. This approach assumes, however, that the combination weights are fixed over the time. We propose to shrink combination weights dynamically towards a target. The proposed approach is used for predicting volatility of the US stock market.
Abstract *E' possibile migliorare l'accuratezza delle previsioni combinando quelle ottenute da metodi diversi. L'approccio della minima varianza assegna un peso specifico a ciascun metodo di previsione con l'obiettivo di ridurre la varianza dell'errore di previsione. Questo approccio presuppone, tuttavia, che i pesi della combinazione siano fissi nel tempo. In questo articolo, proponiamo l'utilizzo dello shrinkage dinamico dei pesi. L'approccio proposto è utilizzato per prevedere la volatilità del mercato azionario statunitense.*

Key words: Forecasting ensemble, financial forecasting, risk prediction, shrinkage estimation

1 Introduction

Forecasters are aware that the combination of predictions from different methods allows for considerable improvements in forecasting accuracy. There is huge empirical evidence documenting the benefit of combination in many applicative domains, but there are also valid theoretical arguments in favor of this practice. As argued by [14], forecasting combination is attractive because it offers a sort of diversification

Raffaele Mattera
Department of Social and Economic Sciences, Sapienza University of Rome, e-mail: raffaele.mattera@uniroma1.it

to the absence of any a priori knowledge about the best forecasting approach to use. This uncertainty can be hedged by combining forecasts.

At a first view, considering the diversification argument, forecasting combination is very close to portfolio selection in finance. In a seminal work, [1] proposed a minimum-variance combination method to find the so-called ”optimal” combination weights. [1] discussed only combinations of pairs of forecasts, but [11] extended the minimum-variance framework to combinations of more than two forecasts. In this minimum-variance framework, the forecasts are combined with weights obtained by variance minimization of the resulting combined forecasting errors.

Classical approaches adopt static combination weights. The main problem with using static weights is that we do not consider instabilities over time. Combination weights should be updated, instead, when some shocks occur in the system [6]. The intuition behind dynamic weights is that forecasting models performing better in some periods – having larger weights in the combination – can perform worse in other time periods. Different forecasting combination approaches using time-varying weights have been proposed [15].

In this paper, we focus on an alternative approach for dynamic combination of forecasts, that employs shrinkage. Shrinkage has been proven to be useful in the context at stake [12], but in a static framework only. In this paper, following [3], we apply shrinkage estimation to combination weights rather than covariances – by shrinking optimal minimum-variance weights towards equal weights –, and allow the shrinkage intensity to evolve over time. The proposed approach is used for predicting the volatility of the US stock market. We empirically demonstrate that the combined forecasts obtained with the dynamic shrinkage are statistically more accurate than many standard combination schemes proposed in literature.

The rest of the paper is structured as follows. Section 2 discusses the methodological proposal, while Section 3 shows the empirical experiment, discussing the data used and the forecasting results. Section 4 concludes with final remarks.

2 Optimal combination weights with dynamic shrinkage

Let us suppose to observe a set of N alternative forecasts, called $\hat{\mathbf{Y}}$, obtained with alternative statistical models. Minimum-variance forecast combination approaches seek to find a set of weights ω, such that the following forecast combination

$$\hat{Y} = \omega' \hat{\mathbf{Y}} \tag{1}$$

has the minimum variance possible. This goal turns the combination problem into the following minimization, subject to the constraint that the weights add up to one

$$\begin{aligned} &\min \omega' \Sigma \omega \\ &\text{s.t. } \omega' \iota = 1. \end{aligned}$$

with Σ be the covariance matrix of the errors associated with $\hat{\mathbf{Y}}$ forecasts and $\iota = (1,1,\ldots,1)'$ a vector of ones. By solving the first order condition and assuming that Σ is invertible, we get the optimal solution [11]

$$\omega^* = \left(\iota'\Sigma^{-1}\iota\right)^{-1}\Sigma^{-1}\iota. \tag{2}$$

In practice, the elements of the covariance matrix Σ need to be estimated, so ω is estimated as well. Thus optimal weights depend by the choice of the estimator for the covariance matrix, say the sample covariance $\hat{\mathbf{S}}$. In this case we can write the estimated combination weights as $\hat{\omega}_S$.

By adopting static shrinkage, we can shrink the minimum-variance weights towards a target, for example the equal weights $\omega_e = 1/N$, thus obtaining

$$\hat{\omega} = \delta\,\hat{\omega}_S + (1-\delta)\,\omega_e, \tag{3}$$

with δ be shrinkage intensity. Notice that under (3) the employed set of weights is however static, which can be inconvenient. Let us assume, instead, that weights are static within a given window and, then, are rebalanced at each time t according to

$$\hat{\omega}_t = \delta_t\,\hat{\omega}_{S,t} + (1-\delta_t)\,\hat{\omega}_{t-1} \tag{4}$$

given that at time $t = 1$ we use the (3). According to (4) we let combination weights to depend by the weights computed at a given lag, say one without loss of generality, with shrinkage intesities time-varying as well. In light of (4), a coherent minimum-variance criterion for the calculation of the optimal shrinkage intensity δ_t^* is

$$\min_{\delta_t} \omega_t'\Sigma\omega_t \tag{5}$$

which solution is given by [3]

$$\delta_t^* = \frac{\hat{\omega}_t'\Sigma\left(\hat{\omega}_t - \hat{\omega}_{S,t}\right)}{\left(\hat{\omega}_t - \hat{\omega}_{S,t}\right)'\Sigma\left(\hat{\omega}_t - \hat{\omega}_{S,t}\right)}.$$

The *bona fide* estimator for $\hat{\delta}_1$ and the iterative algorithm for computing shrinkage intensities $\hat{\delta}_t$ for each $t = 1,\ldots,T$ is discussed in detail in [3].

3 Empirical experiment: forecasting US stock market volatility

To evaluate the performance of the proposed dynamic combination approach, we provide an empirical experiment based on the US stock market volatility forecasting.

We consider the daily time series of the S&P500 Index in the last 10 years, i.e. from the 1st January 2013 to 31th December 2022, therefore we have a time series of length $T = 2518$. The daily returns are shown in Fig. 1.

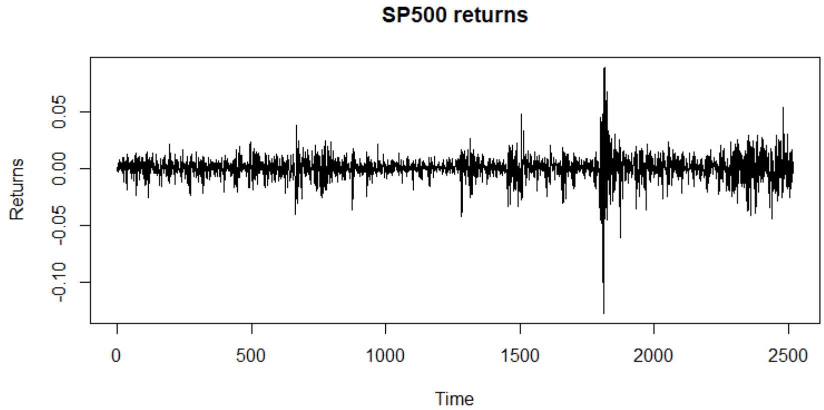


Fig. 1 S&P500 Index returns.

In what follows we are interested in predicting future volatility. Therefore, we split the full sample (Fig. 1) into train and testing set. We leave out the last year of observations for the out-of-sample experiment and consider it as the testing set, which is used to evaluate the accuracy of the proposed dynamic forecasting combination approach. Squared returns are considered as proxy of volatility.

To forecast volatility, we use alternative GARCH models. Following previous studies [16, 4], we compare the out-of-sample volatility forecasts of GARCH models under alternative distributional assumptions. The analyzed $N = 6$ alternatives are: Gaussian, t-student, Generalized Error Distribution (GED), Skew Normal, Skew-t and Skew GED. Alternative distributions are considered to account for possible non-Gaussian specifications, which have been proved to be useful in forecasting volatility [16, 4]. Indeed, is it known that allowance of a leptokurtic and skewed return distribution significantly improves out-of-sample forecasting performances. Considering alternative distributions and combining their forecast is particularly relevant. Indeed, is not known ex-ante what is the best distribution to specify. Hence, combining forecasts obtained by alternative non-Gaussian GARCH model would be beneficial in terms of out-of-sample forecasting accuracy if there is not a priori knowledge about the best model to use in out-of-sample.

To obtain out-of-sample forecasts we rely on a rolling-window procedure. The initial training set of length $\tilde{T} = 2266$ is used to estimate the parameters of the employed statistical model and to obtain the prediction for one step ahead. Then, the procedure is repeated by leaving the oldest observation and including the realization of the data at $t = \tilde{T} + 1$. The procedure stops when no new observation is available.

We compare the dynamic shrinkage-based combination approach with the following alternatives: equal weights, minimum-variance [1, 11] with sample covariance, minimum-variance with shrinkage estimator [10], time-varying weights based on inverse MSE [2], DCC-based modelling of the covariance matrix [7, 8]. For the dynamic shrinkage procedure we allow a fixed rebalance scheme, where shrinkage intensities and combination weights are updated each two months. The forecasting accuracy is evaluated in terms of both RMSE and MAE losses. Results are shown in Tab. 1.

Table 1 Predictive accuracy: RMSE and MAE. The best model is highlighted with bold font.

Forecasting methods	RMSE	MAE
GARCH	0.014370	0.118170
Skew-GARCH	0.014238	0.117724
GED-GARCH	0.014665	0.119384
SGED-GARCH	0.014549	0.118971
GARCH-t	0.015125	0.121278
Skew-t GARCH	0.014888	0.120372
Simple average (equal weights)	0.014642	0.119323
Minimum-variance [1, 11]	0.035304	0.160632
Minimum-variance with shrinkage [10]	0.015612	0.119035
Time-varying MSE [2]	0.014625	0.119266
Minimum-variance with DCC [7, 8]	0.016381	0.126994
Dynamic shrikage	**0.011955**	**0.109265**

Tab. 1 shows that the proposed approach provides more accurate forecasts than the benchmark considering both losses. Notice that in this case simple average does not perform better than the most accurate model, that is the Skew-GARCH assuming Skew Normal distribution. Minimum-variance combination with sample covariance [11] provides the worse performances. Surprisingly, this combination scheme leads to less accurate forecasts than the worse baseline model. Confirming the "forecasting combination puzzle" none of the sophisticate models perform better than the simple average. The only model outperforming this benchmark is the proposed dynamic shrinkage combination scheme, which performs better than all the models considered in this experiment. This improvement is achieved because the diversification between minimum-variance weights and the equally weighted scheme. Considering the results of Model Confidence Set procedure [9], we find that the only model belonging to the superior set is the one based on the dynamic shrinkage.

4 Final remarks

Forecasting combination is a widespread practice in many applicative domains of forecasting. It is attractive especially in the absence of any a priori knowledge about the best forecasting approach. The main problem of forecasting combination lies in the computation of appropriate weights, reflecting how forecasts obtained from alternative methods have to be aggregated. In this paper, we study the problem of minimum-variance combination of forecasts.

The main problem of classical minimum-variance combination lies in the fact that the obtained weights are static rather than dynamic. The use of dynamic weights is relevant on the light of possible instabilities over the time, arising because combination weights should be updated as far as some shocks occur.

In this paper, we propose a dynamic shrinkage approach for modelling dynamic conditional weights. The proposed dynamic minimum-variance combination is used for predicting the volatility in the US stock market. We empirically demonstrate that the combined forecasts obtained with the proposed approach are statistically more accurate than both alternative standard combination schemes as well as base forecasting models.

References

1. J. M. Bates and C. W. Granger. The combination of forecasts. *Journal of the operational research society*, 20(4):451–468 (1969)
2. C. Baumeister and L. Kilian. Forecasting the real price of oil in a changing world: a forecast combination approach. *Journal of Business & Economic Statistics*, 33(3):338–351 (2015)
3. T. Bodnar, N. Parolya, and E. Thorsen. Dynamic shrinkage estimation of the high-dimensional minimum-variance portfolio. *arXiv preprint arXiv:2106.02131* (2021)
4. R. Cerqueti, M. Giacalone, and R. Mattera. Skewed non-gaussian garch models for cryptocurrencies volatility modelling. *Information Sciences*, 527:1–26 (2020)
5. V. De Miguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953 (2007)
6. F. X. Diebold and P. Pauly. Structural change and the combination of forecasts. *Journal of Forecasting*, 6(1):21–40 (1987)
7. R. Engle. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350 (2002)
8. R. F. Engle, C. W. Granger, and D. Kraft. Combining competing forecasts of inflation using a bivariate arch model. *Journal of economic dynamics and control*, 8(2):151–165 (1984)
9. P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2):453–497 (2011)
10. O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411 (2004)
11. P. Newbold and C. W. Granger. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)*, 137(2):131–146 (1974)
12. F. Roccazzella, P. Gambetti, and F. Vrins. Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, 38(1):97–116 (2022)
13. J. Smith and K. F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355 (2009)
14. A. Timmermann. Forecast combinations. *Handbook of economic forecasting*, 1:135–196 (2006)

15. X. Wang, R. J. Hyndman, F. Li, and Y. Kang. Forecast combinations: an over 50-year review. *International Journal of Forecasting* (2022)
16. A. Wilhelmsson. Garch forecasting performance under different distribution assumptions. *Journal of Forecasting*, 25(8):561–578 (2006)

Exploring the perception of the gender issue of Italian female entrepreneurs

Analisi statistica della percezione della questione di genere nelle imprenditrici italiane

Rosella Castellano, Jessica Riccioni and Azzurra Rinaldi

Abstract Our study aims at exploring the data set collected through questionnaires filled by 109 entrepreneurs. The survey is divided into four section grounded on their job and personal satisfaction, their perception of gender issues and difficulties in accessing to credit market. The analysis of the dataset is performed through mixed effect models to identify the correlations between the variables for deriving policy implications.
Abstract *Lo scopo di questo lavoro di ricerca è quello di studiare il livello di soddisfazione, la percezione della questione di genere e le difficoltà di accesso al mercato del credito attraverso un questionario somministrato a 109 imprenditrici italiane. Sui dati è stata effettuata un'analisi attraverso un modello ad effetti misti per investigare il legame tra le variabili analizzate.*

Key words: Italian female entrepreneurs, gender issues, financial inclusion, mixed effect models

1 Introduction

Italian female entrepreneurship is increasingly moving away from the European average in terms of employment and growth. Following the pandemic, the situation

---

Rosella Castellano
Unitelma Sapienza University of Rome, Piazza Sassari, 4 – 00161 – Rome, e-mail: rosella.castellano@unitelmasapienza.it

Jessica Riccioni
Roma Tre University, Via Silvio D'Amico, 77 – 00145 – Rome e-mail: jessica.riccioni@uniroma3.it

Azzurra Rinaldi
Unitelma Sapienza University of Rome, Piazza Sassari, 4 – 00161 – Rome, e-mail: azzurra.rinaldi@unitelmasapienza.it

has worsened, with an increase in the unemployment rate from 49.9% in 2019 to 52.4% in 2020, mainly due to the unpaid family care activity. The situation worsens in the South of Italy where female unemployment reaches 67.1%. It will then be no coincidence that women, together with young people and migrants, are the categories that demonstrate greater entrepreneurial dynamism. Infact, with the employment rate so low and the difficulty in remaining in the workplace after the children (in turn caused by the absence of basic facilities, such as nurseries), many Italian women decide to start and run their own businesses. Before Covid-19, the number of the female companies in Italy was 1,336,227, almost 22% of the total number of companies. Less than 154,000 are the businesses led by young women. And it is young entrepreneurs who have suffered the most from the consequences of the pandemic. Although, in general, after six years of constant growth, the Unioncamere and InfoCamere Women's Entrepreneurship Observatory recorded its first setback in 2020. We must in fact consider that the sectors in which traditionally women are most present with their businesses are those of commerce, tourism, assistance to people and, among manufacturing activities, textiles. Precisely the activities most affected by the pandemic, according to what we read in the Report published by the Observatory. Despite that, with respect to 2019, in 2020 the volume of the female companies decreased by just 0.29%. Although, in general, after six years of constant growth, the Unioncamere and InfoCamere Women's Entrepreneurship Observatory recorded its first setback in 2020. We must in fact consider that the sectors in which traditionally women are most present with their businesses are those of commerce, tourism, assistance to people and, among manufacturing activities, textiles. Precisely the activities most affected by the pandemic, according to what we read in the Report published by the Observatory. Despite that, with respect to 2019, in 2020 the volume of the female companies decreased by just 0.29%. Certainly, even before the Covid crisis there were some persistent problems in Italian female entrepreneurship. For example, the low percentage of female enterprises in the STEM sector or among the innovative startups. According to data from the Ministry of Economic Development, indeed, in the last quarter of 2020 only 13,3% of innovative startups have a prevalence of women, compared to 22% of companies. It is also necessary to remember the issue of the credit crunch: according to the 2020 data of the Unioncamere Observatory, the use of bank credit is low for female businesses: only 20% of them use it. It must be said that the use of credit is not encouraged. Not only because the banking system asks female businesses for greater real guarantees, from third parties, of financial strength and economic growth. But also due to the fact that female entrepreneurs receive rejection from the bank in 8% of cases, against 4% of male businesses.

2 Literature Review

The untapped power of women as financial actors has the potential to dramatically expand the market for financial products and services while improving the lives of women and their communities. Increasing women's access to financial products and services will help transform the lives and futures of hundreds of millions of women in both developed and developing countries, as well as unlock billions of dollars in potential market opportunities . Evidence shows that when women control financial assets, they are often more likely than men to invest in the health, education and well-being of their families — suggesting the significant benefits of financial inclusion to society as a whole and future generations [2]. Despite important advances in expanding access to formal financial services in recent years, a significant access gap remains between men and women. This is illustrated through a basic measure of financial inclusion: account ownership. Globally, only 58% of women hold an account in a formal financial institution, compared to 65% of men [4]. This gender gap is even more pronounced between men and women in developing markets, with the largest gap, 18 percentage points, observed in South Asia [4]. The research fits into this context, with the goal of analyzing the degree of financial inclusion of Italian female entrepreneurs and managers and the impact that exclusion exerts on the value of national production. The extensive research on women-led firms continues to highlight persistent gaps that affect not only the overall number of female entrepreneurs but also the performance of this firms [5], [9]. Female-led firms tend to operate smaller businesses with fewer employees, are often located in services and lower-innovation industries. Most of the traditional sectors in which many women's businesses operate are characterized by low barriers to entry, high competition, low productivity and low profit margins – where enterprises tend to stay small and be low value-added enterprises with lower risk propensity than male-led firms [3]. This implies a more difficult access to credit, which limits the possibilities to expand and internationalize one's business and also to build stable network ties. The crisis derived from COVID-19 has also exacerbated the gender divide at home since the burden of additional domestic responsibilities has disproportionately fallen upon women business leaders. The increase of domestic and care work has decreased the time available for women entrepreneurs to dedicate to their business. The fact that women's businesses are generally smaller is not only a limitation but also an opportunity, especially when faced with sudden change. The greater flexibility of small businesses in changing, adapting or reconverting products, services and strategies, to be a significant element of competitiveness [10]. Unlike larger firms, small firms do not usually implement planned management schemes [8]. Small businesses with "do it yourself" methods try to mobilize all available resources and constantly change their plans in order to capture current opportunities [7]. In recent years, therefore, the connection between innovation and gender has attracted increasing interest among researchers [1, 5, 6]. The remaining part of the paper is organized as follows. In Section 3 we describe the data set and the models. Section 4 is devoted to results and discussion.

3 Mixed Effect Models

3.1 Data

The data set consists of the variables obtained from the questionnaires filled by 109 Italian female entrepreneurs. Sampling is random and the questionnaire was voluntarily completed online. The administration was made through two Italian associations: GammaDonna and ManagerItalia.

In the choice of variables included in the models we estimate, the available data are divided into four groups: personal and general information of the the companies and the female entrepreneurs, personal satisfaction, perception of the gender issue and the difficulty of accessing to credit market. In the mixed effects models we have decided to relate some dependent variables with a set of independent variables to look for the best combinations that allow us to describe and interpret the available data. In the mixed effects models we have decided to relate some dependent variables with a set of independent variables to look for the best combinations that allow us to describe and interpret the available data. Not all models will be shown and not all collected factors will be used for analysis purposes.

3.2 The Mixed effect models with random intercept

We show here, an example of mixed effect model. For each model we considered a random intercept model by selecting the most suitable variables each time to describe the dependent variable. In this type of model the intercept is random and the coefficient of predictors are fixed.

Results to describe the job satisfaction.

Table 1 Estimates for random intercept model considering the job satisfaction. Significative codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Fixed effects	Value	Std.Error	DF	t-value	p-value	Signif.code
(Intercept)	-1.0234162	0.7304515	94	-1.401073	0.1645	
Age_classes	0.0249788	0.0713487	94	0.350094	0.7271	
Educational	0.1019706	0.0586486	94	1.738670	0.0454	*
Company_uptime	0.2336221	0.0881009	94	2.651756	0.0094	**
Location(North)	0.4061754	0.4241583	94	1.257603	0.0307	*
Location(Center)	0.3649493	0.4395871	94	0.830209	0.4085	
Location(South)	0.1040130	0.4605381	94	0.225851	0.8218	
Development	0.1424813	0.1055272	94	1.550185	0.0182	*
Working_life_satisfaction	0.7196634	0.0795668	94	9.044773	0.0000	***
Satisfaction_welfare_maternity	0.0434855	0.0819744	94	1.530477	0.0290	*
Satisfaction_welfare_daycare	0.1335262	0.1211828	94	1.101858	0.2733	
Satisfaction_welfare_nursery	-0.0698468	0.0882711	94	-0.791276	0.4308	
Satisfaction_welfare_eldersassistance	-0.0369099	0.1163264	94	-0.317296	0.7517	

Number of Observations: 109
Number of Groups: 109

4 Results and discussion

In general, the models were selected and all non-significant variables collected were excluded.

Here, we discuss the results for female entrepreneurs. Northern women who belong to higher age groups and have a high school qualification, are generally more satisfied with their life as a whole satisfaction. For them, job satisfaction, the satisfaction of free time and that of friendships have the greatest impact. They too are more sensitive to the welfare service of maternity. The female entrepreneurs do not appreciate other services of the Italian welfare system such as daycare, nursery school and assistance for the elderly people.

What emerges from this model also underlines how women who work in the south are less satisfied with the Italian welfare system and with the other satisfactions that influence their lives.

Female entrepreneurs who work in companies located in the north are more satisfied with their job. It has great influence for both the work they carry out, as well as the growth and stability prospects of the company.

Younger companies are more sensitive to gender issues and apply gender equality policies that limit any discrimination in employment among women employed. On the contrary, companies that have been active for several years have not adapted over time and are strongly affected by the old mentality by not guaranteeing gender balance among their employees.

And this is in line with the existing literature, as well as the fact that, in large companies with greater competition between employees, the gender pay gap impacts more than in medium-small companies.

Female entrepreneurs link the satisfaction of Italian welfare services and the number of children through an inverse relationship. In fact, the rights for women who run their own company are less guaranteed than for women managers. For them, in fact, both the period of maternity (also due to the responsibilities and concerns they have in their own company) and the income received decreases. We remind you that in Italy for women employees are guaranteed 5 months of maternity leave at 100% of their salary and then they can take advantage of various rights including optional maternity, breastfeeding permits and parental leave.

We now turn to companies and their access to credit.

For entrepreneurs, the need to expand sources of financing is stronger in young companies with prospects for development and growth in their future. In this model, women who have adequate financial skills have a greater impact.

There are no gender problems in the requests for financing made over the years. The presentation of guarantees (especially personal guarantees, surities and the presentation of business plans for managers and real warranties for the entrepreneurs), have a positive impact. The highest evaluation of the financing request was given by the most financially skilled women.

The latest model always concerns the subsample, but only for female entrepreneurs. This is the model that studies access to funding sources by women's companies.

Entrepreneurs prefer ordinary credit as a funding channel (with which they deal with gender discrimination problems linked to access to the credit market) and the guarantees presented to obtain the requested loans are fundamental. The entrepreneurs with the most adequate financial skills are also more inclined to positively evaluate and ask for the financial resources necessary for their company.

References

1. Block, J.H., Fisch, C.O., van Praag, M.: The Schumpeterian entrepreneur: a review of the empirical evidence on the antecedents, behaviour and consequences of innovative entrepreneurship. Industry and Innovation, 24(1), 61-95 (2017) DOI: 10.1080/13662716.2016.1216397
2. BNY Mellon and United Nations Foundation: Powering Potential: Increasing women's access to financial products and services (2018)
3. Carranza, E., Dhakal, C., Love, I.: Female Entrepreneurs : How and Why Are They Different?. Jobs Working Paper No. 20. World Bank, Washington, DC (2018)
4. Demirguc-Kunt, A., Klapper, L., Singer, D., Van Oudheusden, P.: The Global Findex Database 2014: Measuring Financial Inclusion around the World. Policy Research Working Paper, No. 7255. World Bank, Washington, DC (2015)
5. Jennings, J. E., Brush, C. G.: Research on women entrepreneurs: Challenges to (and from) the broader entrepreneurship literature? The Academy of Management Annals, 7(1), 663–715 (2013). https://doi.org/10.1080/19416520.2013.782190
6. Johansson, A.W., Lindberg, M.: Making a Case for Gender-inclusive Innovation through the Concept of Creative Imitation. Annals of Innovation and Entrepreneurship, 2(2), 1-13 (2011)
7. Koltai, L., Geambasu, R., Bakacsi-Saffer, Z., Barna-Petróczi, A., and Zsár, V.: COVID-19 and Female Entrepreneurs Throughout Europe. Budapest: Hetfa Research Institute Ltd (2020)
8. Mallak, L.A.: Measuring Resilience in Healthcare Provider Organizations. Health Manpower Management, 24, 148-152 (1998). https://doi.org/10.1108/09552069810215755
9. McAdam, M. Harrison, R.T., Leitch, C.M.: Stories from the field: women's networking as gender capital in entrepreneurial ecosystems. Small Business Economics, 53(2), 459-474 (2019)
10. Smallbone D, Deakins D, Battisti M, Kitching J.: Small business responses to a major economic downturn: Empirical perspectives from New Zealand and the United Kingdom. International Small Business Journal. 30(7), 754-777 (2012). doi:10.1177/0266242612448077

Solicited Session SS11 - *Networks data analysis: new perspectives and applications*

Organizer and Chair: Annalina Sarra

1. *Describing Italian mobility trajectories in higher education* (Genova V.G., Giordano G., Ragozini G. and Vitale M.P.)
2. *Collaboration networks: methodological issues and updated empirical evidence on Italian statisticians* (De Stefano D., Fabbrucci Barbagli A.G., Santelli F. and Zaccarin S.)
3. *Mapping Ashtma Complexity with Graph Theory: an Integrative Approach* (Cucco A., Simpson A., Murray C., Fontanella S. and Custovic A.)
4. *Investigating the patterns of Italian internal mobility: a network analysis at provincial level* (Sarra A., D'Ingiullo D., Evangelista A., Nissi E., Quaglione D. and Di Battista T.)

Describing Italian mobility trajectories in higher education

L'analisi delle traiettorie della mobilità studentesca universitaria in Italia

Vincenzo G. Genova, Giuseppe Giordano, Giancarlo Ragozini and Maria Prosperina Vitale

Abstract The intra-national mobility in higher education represents a relevant issue to be further investigated in order to discover the universities chosen by students to attend specific bachelor's and master's degree programmes. The aim of the present contribution is to describe the main trajectories of mobility flows among Italian geographical areas by highlighting also the usefulness of network analysis approach in dealing with different data structures. Data on four student cohorts are extracted from MOBYSU.IT database. The main findings confirm the attractiveness of Northern universities as well as the South-to-North and the North-to-North student mobility trajectories.

Abstract *La mobilità studentesca rappresenta un tema di interesse per il sistema universitario per individuare le principali destinazioni che gli studenti scelgono dopo il diploma. L'obiettivo del presente contributo è descrivere le traiettorie dei flussi di mobilità tra le aree geografiche italiane. I dati di quattro coorti di studenti iscritti ad un corso di laurea triennale sono estratti dal database MOBYSU.IT. I principali risultati confermano l'attrattività delle università del Nord, evidenziando le traiettorie della mobilità studentesca dal Sud al Nord ma anche tra le regioni del Nord.*

Vincenzo G. Genova
Dept. of Economics, Business, and Statistics, University of Palermo, Viale delle Scienze Ed. 13, Palermo, Italy, e-mail: vincenzogiuseppe.genova@unipa.it

Giuseppe Giordano
Dept. of Political and Social Studies, University of Salerno, Via Giovanni Paolo II, Fisciano (Salerno), Italy, e-mail: ggiordano@unisa.it

Giancarlo Ragozini
Dept. of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, Naples, Italy, e-mail: giragoz@unina.it

Maria Prosperina Vitale
Dept. of Political and Social Studies, University of Salerno, Via Giovanni Paolo II, Fisciano (Salerno), Italy, e-mail: mvitale@unisa.it

Key words: Student mobility, Network data structures, Community detection

1 Introduction

Along with the persistence of Italian inter-regional economic disparities between the South and the Centre-North, it is underlined the relevance of investigating intra-national mobility flows [7] due to the peculiar characteristics of the Italian university system [5, 18, 4, 11, 19]. This phenomenon follows the traditional South-to-North migration chain [12] but also the recent North-to-North mobility [17]. In line with related international literature, student mobility choices seem to be related to several factors; among the others, the promotion of individual social mobility pathways influenced by parental backgrounds [14], the attractiveness of universities with high quality and reputation [3], and the possibility to move in a context with a better quality of life and a local labour market offering good job opportunities after graduation [13, 6]. This phenomenon entails negative effects in terms of an imbalance in government funds to universities, a relevant brain drain of highly educated, and welt young people also in presence of a demographic imbalance [18].

Within this scenario, the aim of the present contribution is to describe the main trajectories of mobility flows among Italian geographical areas by highlighting also the usefulness of network analysis approaches in dealing with different data structures. Data on four student cohorts are extracted from MOBYSU.IT database[1] [15]. From this archive, a very wide spectrum of data structures can be derived and further investigated by network analysis tools [5, 18, 4, 16, 12], from the simplest one in which unipartite weighted directed networks report geographical areas or universities as nodes and students' flows represent the links' weights, to complex data structures in which *students*, *regions*, *provinces*, *universities*, and *educational programmes* can be considered all together. These latter configurations, called multimode networks [8, 9], can be analyzed to enrich the interpretation of the phenomenon under analysis.

In the remainder of the contribution, Section 2 describes the students' cohorts under analysis and the main trajectories of mobility flows among geographical areas. Section 2.1 offers some findings in applying clustering methods on simplified multimode networks derived for the students' cohorts.

[1] The database includes the information on students' demographic characteristics, their high school background and their bachelor and master program choices at university. The availability of data at the individual level allows us to reconstruct the whole student career by considering socio-demographic characteristics together with educational achievements, from high school to university degree programmes.

2 Student mobility trajectories

In the following, four student cohorts enrolled in Italian universities in the academic years 2008-2009, 2011-2012, 2014-2015, and 2017-2018 are investigated considering information about the regions and provinces of origin, the university of destination, and educational programmes.[2] Table 1 shows the 107 Italian provinces, around 80 public and private universities, and the 10 ISCED fields to describe Italian student mobility flows in higher education over time. The number of students enrolled in the Italian university system for the four academic years varies from around 220,000 to 250,000 per cohort. Considering the distribution of students according to gender and mobility status, that is students who are enrolled in a university out of their region of residence (*movers*) or the ones studying in the region of residence (*stayers*), it is worth to note that the percentage of movers is increased over time especially for females, from 15.8% to 20.9% (Table 2).

The propensity to move for attending a degree programme can be affected by commuting mobility characterized by students enrolling in a university neighbouring their region of residence, which does not describe in a proper way mobility phenomenon. Indeed, it is logical to assume that these students commute daily to these neighbouring universities to attend classes and return home at the end of the day. To overcome such masking effect, an adjusted version of the outflow/inflow rates is considered in line with Attanasio et al. [1]. The authors define the raw regional outflow/inflow rate as the ratio of the number of outgoing/incoming students of/in the $i-th$ region out of the total number of students in the $i-th$ region. The adjusted version of rates purge from movements that occurred between neighbouring regions. Thus, evaluating students' mobility controlling for commuting mobility, Italy seems divided into two parts: the North has an adjusted outflow rate of around 10% and the South with outflow rate of around 40% (Figure 1, upper panel). The adjusted regional students inflow rate (Figure 1, bottom panel) confirms the well-known structure of the Italian student mobility, from South-to-North. Central-Northern Italian regions have been the ones attracting student flows' since 2008. Regardless of the North-North mobility, student mobility from the Southern to the Central and Northern regions has increased steadily in the last 10 years. This phenomenon has created inequalities as well as cultural and socio-economic losses for the South that do not appear to be slowing down.

[2] The educational programmes are measured by considering the 10 fields included in the International Standard Classification of Education (ISCED-F 2013): *Education, Arts and Humanities, Social Sciences, Journalism and Information, Business, Administration and Law, Natural Sciences, Mathematics and Statistics, Information and Communication Technologies Engineering, Manufacturing and Construction, Agriculture, Forestry, Fisheries and Veterinary,Health and Welfare, Services*.

Table 1 Student cohorts description

	a.y. 2008-2009	a.y. 2011-2012	a.y. 2014-2015	a.y. 2017-2018
Provinces	107	107	107	107
Universities	79	79	80	80
ISCED fields	10	10	10	10
N. of students	243,331	229,085	223,479	247,811

Table 2 Student cohorts by gender and mobility status (%)

Cohort	Gender	Total	Mover status	
			Stayers%	Movers%
a.y. 2008-2009	F	136,381	84.2	15.8
	M	106,950	82.8	17.2
a.y. 2011-2012	F	126,606	81.7	18.3
	M	102,479	80.9	19.1
a.y. 2014-2015	F	121,121	80.5	19.5
	M	102,358	80.4	19.6
a.y. 2017-2018	F	134,315	79.1	20.9
	M	113,496	79.8	20.2

2.1 Insights from network data analysis

The student mobility flows described in Section 2 can be further analyzed by defining different network configurations. Here we exploit the information of the geographical areas of origin and destination with respect to the chosen universities and educational programmes. These data structures lead to complex networks [8] that can be simplified and analyzed through the analytic strategy proposed by Genova et al. [10]. The derived bipartite weighted networks are then explored by means of the Network Analysis tools, for instance to reveal the presence of communities. Specifically, we present the results obtained by applying a flow-based community detection algorithm [2].

The main findings reveal preferential attractiveness routes of Northern universities and a dichotomy between scientific and humanistic fields. For the a.y. 2008-2009, two main clusters of attractive universities in North Italy are obtained, representing the main trajectories of the Italian student mobility. The a.y. 2017-2018 cluster solution identifies four main groups, highlighting the dichotomy between scientific and humanistic fields attracting students from the South to the North and within the Northern regions. The partitioning solution mixes up the different types of units enriching the interpretation of the phenomenon over time, bringing to light the main characteristics of Italian student mobility flows.

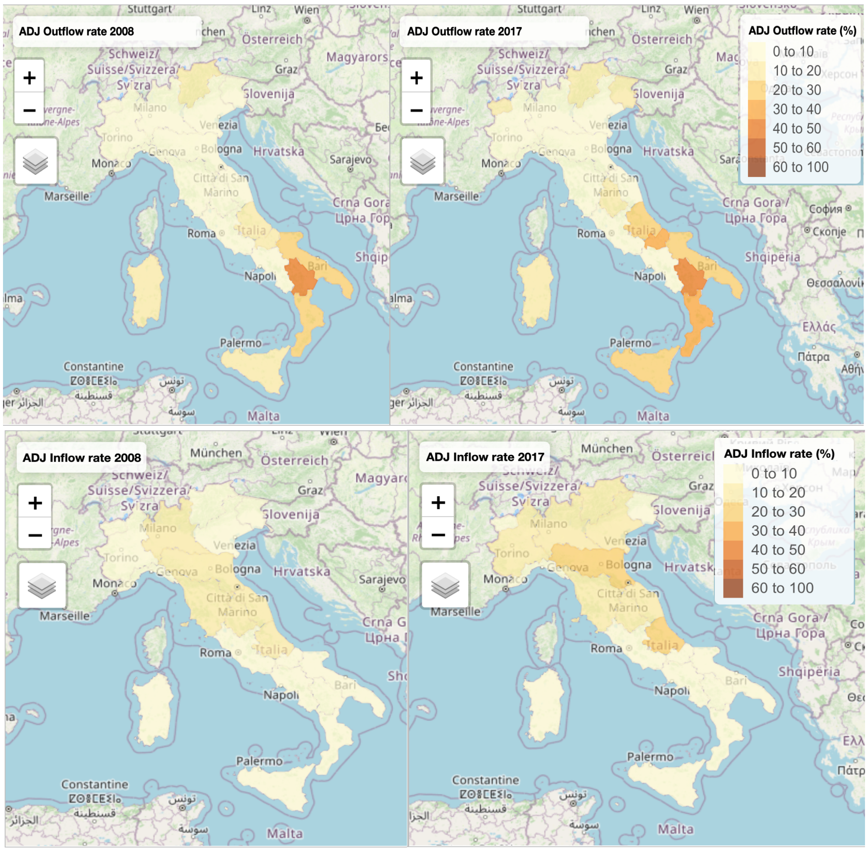


Fig. 1 Cartogram of the adjusted regional: i) outflow rate–ratio of the number of outgoing students of the $i-th$ region that move to a non-neighbouring region out of the total number of students in the $i-th$ region; and ii) inflow rate–ratio of the number of incoming students in the $i-th$ region from a non-neighbouring region out of the total number of students enrolled in the $i-th$ region.

Funding

The contribution has been supported from Italian Ministerial grant PRIN 2017 “From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide”, n. 2017 HBTK5P - CUP B78D19000180001.

References

1. Attanasio, M., Priulla, A.: Chi rimane e chi se ne va? Un'analisi statistica della mobilità universitaria dal Mezzogiorno d'Italia. FrancoAngeli (2020)
2. Blöcker, C., Rosvall, M.: Mapping flows on bipartite networks. Physical Review E 102, 052,305 (2020)
3. Ciriaci, D.: Does university quality influence the interregional mobility of students and graduates? the case of italy. Regional Studies 48, 1592–1608 (2014)
4. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Geography of italian student mobility: a network analysis approach. Socio-Economic Planning Sciences 73, 100,918 (2021)
5. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Correction to: Analysing the determinants of italian university student mobility pathways. Genus 78, 1–1 (2022)
6. Dotti, N.F., Fratesi, U., Lenzi, C., Percoco, M.: Local labour markets and the interregional mobility of italian university students. Spatial Economic Analysis 8, 443–468 (2013)
7. Dotti, N.F., Fratesi, U., Lenzi, C., Percoco, M.: Local labour market conditions and the spatial mobility of science and technology university students: evidence from italy. Review of Regional Research: Jahrbuch für Regionalwissenschaft 34, 119–137 (2014)
8. Everett, M.G., Borgatti, S.P.: Partitioning multimode networks. Advances in network clustering and blockmodeling , 251–265 (2019)
9. Fararo, T.J., Doreian, P.: Tripartite structural analysis: Generalizing the breiger-wilson formalism. Social Networks 6, 141–175 (1984)
10. Genova, V.G., Giordano, G., Ragozini, G., Vitale, M.P.: Clustering student mobility data in 3-way networks. In: Book of Abstracts IFCS 2022, 17th Conference of the International Federation of Classification Societies "Classification and Data Science in the Digital Age", p. 56. Instituto Nacional de Estatística (2022)
11. Genova, V.G., Tumminello, M., Aiello, F., Attanasio, M.: A network analysis of student mobility patterns from high school to master's. Statistical Methods & Applications 30, 1445–1464 (2021)
12. Genova, V.G., Tumminello, M., Enea, M., Aiello, F., Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electronic Journal of Applied Statistical Analysis 12, 774–800 (2019)
13. Giambona, F., Porcu, M., Sulis, I.: Students mobility: Assessing the determinants of attractiveness across competing territorial areas. Social indicators research 133, 1105–1132 (2017)
14. Impicciatore, R., Tosi, F.: Student mobility in italy: The increasing role of family background during the expansion of higher education supply. Research in Social Stratification and Mobility 62, 100,409 (2019)
15. MOBYSU.IT: Database MOBYSU.IT, *Mobilità degli studi universitari italiani*, Protocollo di ricerca MIUR - Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II, Fonte dei dati ANS-MIUR/CINECA (2016)
16. Primerano, I., Santelli, F., Usala, C.: A multiplex network approach to study italian students' mobility. Book of short Papers SIS 2021, 473–478 (2021)
17. Rizzi, L., Grassetti, L., Attanasio, M.: Moving from north to north: how are the students' university flows? Genus 77, 1–22 (2021)
18. Santelli, F., Ragozini, G., Vitale, M.P.: Assessing the effects of local contexts on the mobility choices of university students in campania region in italy. Genus 78, 1–25 (2022)
19. Santelli, F., Scolorato, C., Ragozini, G.: On the determinants of student mobility in an interregional perspective: A focus on campania region. Statistica Applicata-Italian Journal of Applied Statistics , 119–142 (2019)

Collaboration networks: methodological issues and updated empirical evidence on Italian statisticians

Reti di collaborazione: questioni metodologiche e nuove evidenze empiriche sugli statistici italiani

Domenico De Stefano, Amin Gino Fabbrucci Barbagli, Francesco Santelli and Susanna Zaccarin

Abstract This paper aims to report new results on the co-authorship networks of Italian academic statisticians, a community of more than 700 scholars distributed in 5 distinct scientific subfields. Data on scientific production is obtained from Scival platform using author Scopus IDs as scholar's unique identifier that allows to overcome misspellings, ambiguous names, duplicates, etc. that may affect network construction. Co-authorship structures and productivity over time will be analyzed and compared in two time periods: from 2012 to 2015 and from 2016 to 2019.
Abstract *L'obiettivo di questo lavoro è quello di presentare nuovi risultati sulle reti di collaborazione degli statistici italiani, una comunità di oltre 700 studiosi distribuiti in 5 distinti sottosettori scientifici. Le informazioni sulla produzione scientifica sono ottenute dalla piattaforma Scival utilizzando lo Scopus ID degli autori come identificativo univoco che consente di evitare problemi di identificazione degli autori e eventuali duplicazioni che impattano negativamente sulla costruzione della rete. La struttura delle collaborazioni e la produttività verranno analizzati e comparati in due periodi temporali: da 2012 a 2015 e da 2016 a 2019.*

Key words: Italian academic statisticians, co-authorship, networks analysis

---

Domenico De Stefano
University of Trieste, e-mail: `ddestefano@units.it`

Amin Gino Fabbrucci Barbagli
University of Trieste, e-mail: `amingino.fabbruccibarbagli@phd.units.it`

Francesco Santelli
University of Trieste, e-mail: `fsantelli@units.it`

Susanna Zaccarin
University of Trieste, e-mail: `susanna.zaccarin@deams.units.it`

1 Introduction

In the last years, the attention to scientific collaboration has increased, since it is recognized as a fundamental process that allows researchers to share ideas, knowledge, and techniques. Thanks to different expertise [4], it is also a great opportunity in research progress and innovation. In addition, this process is supported by government policies at the national and international level with programs enhancing the collaboration among researchers [17, 7].

The common way to study scientific collaboration is by analyzing co-authorship, assuming that if two or more scholars co-authored the same publication, it is highly likely that they have collaborated[14].
Different methodologies can be used to investigate this collaboration process. In this work, we will use the Social Network Analysis (SNA) tools as a methodological approach to analyze co-authorship behaviour, and to explore the topological properties of the network - that is to find out the presence of a small world pattern [16] and/or a scale-free configuration [13][1]. In addition, by means of community detection or generalized blockmodeling methods, clusters of authors can be found.

Lately, many scholars' communities have been studied and it is possible to distinguish two approaches for choosing the target and tools for the networks: 1) by focusing on collaboration among scholars of a specific scientific field or discipline, such as Economics [9], Sociology [12]; 2) by focusing on the scholars within an institution in one country, tied to a specific field [8][15][6].Building the co-authorship networks requires retrieving information from different databases and sources. Thanks to the increasing availability of electronic archives it is easier to obtain a large quantity of information by querying international bibliographic archives that store scientific products (in journals, books, reviews, collective volume) indexed by thematic area (e.g., ISI-Web of Science or Scopus or Medline, Econlit, Current Index to Statistics)[11][5][3][16].
In this work, we will focus on the scholars hired in Italian universities in a specific discipline -Statistics- collecting information from the MUR, Scopus and Scival databases. The networks have been built using R and the library (igraph) with the aim of detecting the evolution of collaboration patterns among Italian statisticians over time.

2 Target and aim

This work focuses on the scientific collaboration of the 760 statisticians in Italy – as recorded in the MUR on 31/12/2019 in permanent (FP, PA, RU) or temporary positions (Assistant Prof/Researcher)- classified into five different academic subfields named SSDs ("Settore Scientifico Disciplinare"): SECS-S/01 Statistics (438 scholars), SECS-S/02 Statistics for Experimental and Technological Research (21), SECS-S/03 Economic Statistics (156), SECS-S/04 Demography (70), SECS-S/05 Social Statistics (75). Using multiple data sources may imply errors such as misspellings, ambiguous names, and duplicates that can affect the resulting final dataset. For this reason, the full list of statisticians was retrieved from the MUR database and afterwards, the list of names was used to collect the Scopus ID from Scopus, in order to have a unique identifier for every author. Then, we obtained from Scival's database a dataset containing all the publications where at least one of our target scholars was present. In order to bring out possible evolution in the collaboration over time, two periods have been considered: 1) from 2012 to 2015, with 3329 papers, 2) from 2016 to 2019, with 4098 papers.

In general, the academic statistician community collaborates in several different subject areas (Fig.1): mathematics (40%), social science (19.8%), and medicine (18.1%). In particular, statistics (-S/01) tends to have a higher number of publications in mathematics (51.0%) and in decision science (27.6%); economic statistics (-S/03) in economics, econometrics, and finance (31.2%) and social science (30.4%); demography (-S/04) in social science (64.9%) and medicine (17.3%); social statistics (-S/05) in social science (37.0%), mathematics (25.0%).

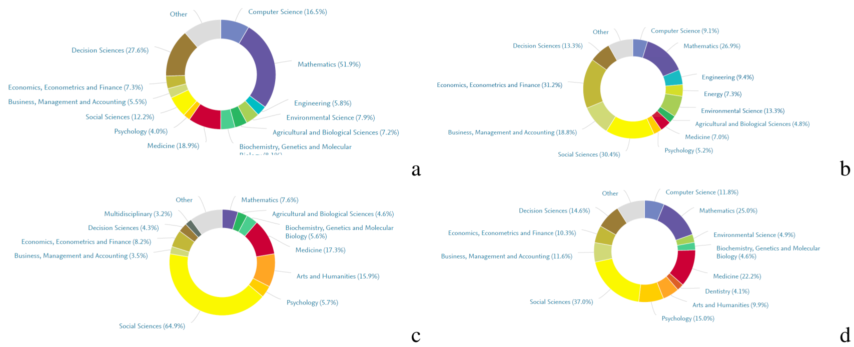


Fig. 1: Publications by subject area from Scival database for Statistics -S/01 (a), Economic Statistics -S/03(b), Demography -S/04 (c), Social Statistics -S/05 (d)

3 Co-authorship Networks

For this work, we used the SNA tools to investigate the properties and characteristics of the scholars' network by analyzing the topological structure and centralization scores of the entire network. In particular, the following statistics have been considered:

- Density, defined as the sum of the ties divided by the sum of possible ties. A high value of density means that the nodes of the network are very cohesive.
- Transitivity, defined as the extent to which a given node links to its neighbors, and measured by the clustering coefficient [2]. A high value of the clustering

Table 1: Network characteristics in the two time periods

	2012-2015	2016-2019
Nodes	1722	8130
statisticians (nodes)	372	619
Links	5881	38513
Average degree	6.83	9.48
Average papers per collaboration	1.13	1.28
Network statistics		
Transitivity	0.858	0.506
Density	0.004	0.012
Betwenness centralizaton	2.24e-09	5.91e-09
Degree centralization	0.030	0.059

coefficient means that the network is strongly connected, and it is a prerequisite for a small-world mechanism of tie formation.

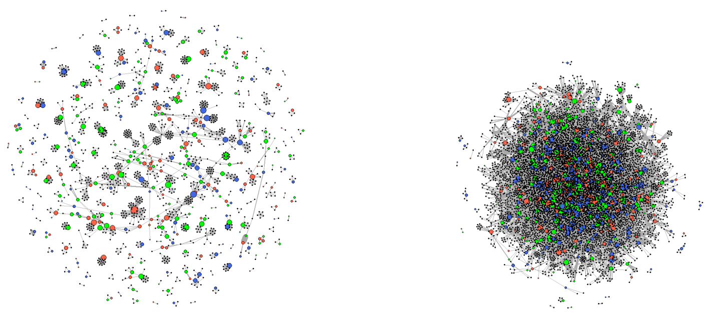
- Betweenness centralization, obtained by comparing the betweenness centrality of the nodes, and measuring if there is a concentration of nodes in the network that serve as connectors among the other authors.
- Degree centralization, obtained by comparing the degree centrality of the nodes, and representing the concentration of ties around a few authors. A high value of the index is consistent with the network representation as a star-shaped scenario [10].

These scores give a first interpretation of our networks. Betweenness and degree centralization are useful metrics to describe the overall configuration of the collaboration network in terms of importance of few authors in the network structure. Transitivity allows us to determine the propensity to collaborate with other peers in tightly connected subgroups. Therefore the interplay of clustering coefficient and the different types of centralization allows to detect if there are substantial changes in the collaboration patterns.

Table 2: Distribution of the academic roles of statisticians (Assistant Professors, Associate Professor, Full Professors) and co-authors from other fields over the two periods

Period	Assistant Professors	Associate Professor	Full Professors	No statisticians
2012-2015	40.3%	30.4%	29.3%	1350
2016-2019	33.4%	39.1%	27.5%	7511

The different network statistics allow a preliminary comparison of the two periods. In the first one, the number of nodes is 1722, of which 372 (about 22%) are statisticians, connected by 5881 links and with a low tendency to the degree centralization (see Table 2). Nodes appear to be connected among each other without the presence of a star-like graph. With respect to the second period, nodes increase to 8130, where 619 (about 8%) are statisticians and 7511 belong to other fields with more than 38000 links (see Table 1), representing an increased number of collaborations also outside the statistician's community, still with a low value of degree centralization, so with the absence of more central nodes and a star-like graph. Also, the average degree (average number of edges per node) rises, showing a higher level of connection in the network, with density also increased. In the time frame 2016-2019 scholars collaborated with a higher number of co-authors (higher average degree) but with a lower propensity to work in tightly connected groups with respect to the first period (lower transitivity). In this period, the percentage of full professors and assistant professors decreased, on the contrary, the percentage of associate professors and the number of external scholars from other fields increased (see Table 2 and Fig. 2). Also, it emerges a slightly higher propensity to centralization with respect to the 2012-2015 time frame; and a higher average weight, representing an increased number of publication for each collaboration tie. Therefore, some main structural changes occurred over time and deserve further investigation.



(a) (b)

Fig. 2: Networks in the two periods: (a) 2012-2015, (b) 2016-2019.Legend: Red nodes: full professors; Blue nodes: associate professors; Green nodes: assistant; Black nodes: not belonging to the academic statisticians community.

4 Final remarks

In this contribution, we briefly reported the preliminary results of the co-authorship networks of Italian academic statisticians retrieved through different online platforms: MUR, Scival, and Scopus databases. Collaboration in the field has increased in the period 2016-2019, with a tendency to publish more multidisciplinary works, increasing also the average number of co-authors. By considering the topology and the author's position, as future developments, we will compare the two networks over time, to better understand what are the differences in collaboration patterns as well as the impact on author performance. Thus, specific ego networks will be retrieved from the overall picture, deepening peculiar scholars' collaboration structures. We will also perform a robust network community detection able to disentangle network complexity and compare the emerging communities over time.

References

1. Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. In: Physica A: Statistical Mechanics and its Applications, Vol.311, pp.590-614 (2002)
2. Barabási, A.L., Pósfai, M.: Network science. Cambridge University Press (2016)
3. Bellotti, E., Kronegger, L., Guadalupi, L.: The evolution of research collaboration within and across disciplines in Italian Academia. In: Scientometrics, Vol. 109, pp. 783-811 (2016)
4. De Stefano, D., Vitale, M.P, Zaccarin, S.: The scientific collaboration network of Italian academic statisticians. In: 45th Scientific Meeting of the Italian Statistical Society. Book of Short Papers (2010)
5. De Stefano, D. , Fuccella, V., Vitale, M.P., Zaccarin, S.: The use of different data sources in the analysis of co-authorship networks and scientific performance. In: Social Networks, Vol. 35, pp. 370-381 (2013)
6. De Stefano, D., Fuccella, V., Vitale, M.P., Zaccarin, S.: Quality issues in co-authorship data of a national scientific community. In: Network Science, Vol. 11, pp. 98-112 (2023)
7. Defazio, D., Lockett, A., Wright, M.: Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. In: Research Policy, Vol. 38, pp. 293-305 (2009)
8. Digiampietri, L., Rego, L., Souza, F., Ospina, R., Mena-Chalco, J.: Brazilian Network of PhDs Working with Probability and Statistics. In: Brazilian Journal of Probability and Statistics, Vol. 32, pp. 755-782 (2017)
9. Goyal, S. , van der Leij, M. J., Moraga-González, J. L.: Economics: An Emerging Small World. In: Journal of Political Economy, Vol. 114, pp.403-412 (2006)
10. Hanneman, R.A., Riddle, M.: Introduction to social network methods. Riverside, CA: University of California, Riverside (2005). Available online: http://www.faculty.ucr.edu/ hanneman/
11. Kronegger, L., Ferligoj, A., Doreian, P.: On the dynamics of national scientific systems. In: Quality & Quantity, Vol.45, pp.989–1015 (2011)
12. Moody, J.: The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. In: American Sociological Review, Vol. 69, pp. 213-238 (2004)
13. Newman, M. E. J.: Scientific collaboration networks. I. Network construction and fundamental results. In: Phys. Rev. E, Vol. 64 (2001)
14. Ponomariov, B., Boardman, C.: What is co-authorship?. In: Scientometrics, Vol. 109, pp. 1939-1963 (2016)
15. Sciabolazza, V.L., Vacca, R., Okraku, T.K., McCarty, C.: Detecting and analyzing research communities in longitudinal scientific networks. In: PLoS One, Vol. 12 (2017)
16. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. In: Nature, vol.393, n.6684, pp. 440–442 (1998)
17. Wuchty, S., Jones, B., Uzzi, B.: The Increasing Dominance of Teams in Production of Knowledge. In: Science, pp.1036-1039, Vol.316 (2007)

Mapping Asthma Complexity with Graph Theory: An Integrative Approach

Esplorare la Complessità dell'Asma attraverso la Teoria dei Grafi: Un Approccio Integrativo

Alex Cucco, Angela Simpson, Clare Murray, Sara Fontanella* and Adnan Custovic*

Abstract The relationship between allergic sensitisation and asthma is intricate and characterised by conflicting data regarding its strength. Limited information exists on the longitudinal patterns of IgE responses to allergenic proteins. In this study, we aim to fill this knowledge gap by constructing networks that investigate the links between immunoglobulin E (IgE) antibodies targeting specific allergenic molecules (components). Our objective is to analyse the temporal patterns of allergic sensitisation from infancy to adolescence and assess their association with allergic diseases. To achieve this, we utilise network embedding techniques to analyse data obtained from component-resolved diagnostics from the Manchester Asthma and Allergy birth cohort. Through this approach, we aim to uncover valuable insights into the pathogenesis of asthma, early detection strategies, and personalised interventions.

Abstract *La relazione tra sensibilizzazione allergica e asma è complessa e caratterizzata da dati contrastanti riguardo alla sua intensità. Le informazioni sullo evoluzione temporalei delle risposte IgE alle proteine allergeniche sono limitate. In questo studio, ci proponiamo di colmare questa lacuna costruendo reti che investigano i leegami tra gli anticorpi immunoglobulina E (IgE) diretti contro specifiche*

Alex Cucco
National Heart and Lung Institute, Imperial College London, UK e-mail: a.cucco20@imperial.ac.uk

Angela Simpson
Division of Infection, Immunity and Respiratory Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, UK

Clare Murray
Division of Infection, Immunity and Respiratory Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, UK

Sara Fontanella
National Heart and Lung Institute, Imperial College London, UK

Adnan Custovic
National Heart and Lung Institute, Imperial College London, UK

*Joint senior author

molecole allergeniche (componenti). Il nostro obiettivo è analizzare i modelli temporali di sensibilizzazione allergica dall'infanzia all'adolescenza e valutare la loro associazione con le malattie allergiche. Per raggiungere questo obiettivo, utilizziamo tecniche di network embedding per analizzare i dati ottenuti da test diagnostici utilizzati nel Manchester Asthma and Allergy Study. Attraverso questo approccio, miriamo a scoprire importanti informazioni sulla patogenesi dell'asma, sulle strategie di rilevazione precoce e sulle terapie personalizzate.

Key words: Biological networks, graph embedding, dimensionality reduction, allergic diseases

1 Introduction

Allergy sensitisation increases the likelihood of developing asthma[1, 2, 3], albeit the strength of this link varies [4, 5]. A patient is often deemed sensitised if they have a positive skin prick test (SPT) reaction or specific IgE antibodies against common inhalation and food allergens are identified in their blood [6]. Nevertheless, in a clinical setting, the affirmation of allergic sensitization through standard diagnostic tests does not automatically imply that the patient's symptoms are a result of an allergic reaction [1]. Our previous research suggests that the inconsistencies in the relationship between allergic sensitization and asthma may be attributed to the fact that "allergic sensitisation" encompasses different types of sensitisation with varying associations with asthma and other allergic conditions Recent evidence indicates that sensitisation to specific components, rather than the complete range of allergenic proteins, plays a significant role in allergic disease development. Advanced molecular diagnostics, exemplified by the ImmunoCAP ISAC, allow for the simultaneous measurement of specific IgE antibodies to numerous allergen components, enabling a more precise examination of allergic sensitisation. In our previous studies, we utilised machine learning (ML) techniques to analyse these complex data and demonstrated the discriminatory capacity of component-specific IgE responses in asthma and rhino-conjunctivitis [7]. Latent variable modeling revealed distinct clusters of IgE responses among school-age children, each associated with specific clinical symptoms [8, 9, 10]. Utilising nested latent class probabilistic modeling, we also observed varying clinical outcomes based on longitudinal sensitisation trajectories to different allergens during childhood. These findings emphasise the critical role of time of onset and specific patterns of IgE response in the manifestation of allergic diseases[11]. By employing network analytics, w have also highlighted the significance of non-linear interactions among IgE antibodies in influencing clinical outcomes such as asthma [13]. These studies emphasize the importance of using a statistical framework to depict IgE antibody interactions as a biological network. Nonetheless, effectively capturing the diverse longitudinal responses to multiple allergen components from various sources poses a challenge. To explore these complex dynamics, we employ advanced statistical methodologies, focusing on network embedding techniques to investigate similarities between networks. Additionally, dimensionality reduction methods aid in generating hypotheses regarding key network characteristics underlying observed differences. By transforming network data into lower-dimensional representations using network embedding techniques, we can effectively capture structural and functional similarities between networks. Our primary objective is to conduct an exploratory analysis of the evolution of the IgE network over time without imposing temporal assumptions or specific models. By adopting a flexible and unbiased approach, we gain insights into the changes and patterns in the IgE network over time, leading to the discovery of novel insights and potential trends.

2 Graph embedding, dimensionality reduction

A graph $\mathscr{G}$ is represented by a pair $\mathscr{G}(V,E)$, where V is a set of n finite vertices v_i within $\mathscr{G}$, and E is a set of edges $e_{i,j}$ that connect pairs of vertices v_i and v_j from V. Nodes connected by an edge are referred to as adjacent or neighboring nodes. The degree of a vertex v_i, denoted as $\deg(v_i)$, is determined by the sum of all the edges in its neighborhood. To capture this information, a degree matrix $\mathbf{D}$ is constructed, with each diagonal element d_{ii} representing the degree of node v_i ($\deg(v_i)$). A graph is considered unweighted if the elements $e_{\cdot,\cdot}$ in the edge set E can only be 0 or 1, indicating the absence or presence of an edge, respectively. Conversely, a graph is classified as weighted if the edge values are continuous. If e_{ij} is equal to e_{ji} for all pairs of vertices i and j in V, the graph is referred to as undirected. On the other hand, if there exists at least one pair of vertices i and j in V where e_{ij} is not equal to e_{ji}, the graph is said to be directed.

For graph representation, an adjacency matrix $\mathbf{A}$ of size $n \times n$ can be defined. This matrix contains the edge information between each pair of vertices, where the elements $a_{i,j}$ are equal to $e_{i,j}$. The graph Laplacian, denoted as $\mathbf{L}$, is defined as the difference between the degree matrix $\mathbf{D}$ and the adjacency matrix $\mathbf{A}$, i.e.,

$$\mathbf{L} = \mathbf{D} - \mathbf{A}.$$

An interesting property of the graph Laplacian is its positive semidefinite nature. This implies that given a set of m graphs $[\mathscr{G}_1, \mathscr{G}_2, ..., \mathscr{G}_m]$ with a shared set of vertices V, each corresponding Laplacian $\mathbf{L}$ can be represented as a point within a hypercone. Specifically, the Laplacian space associated with a given set of m Laplacians $[\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_m]$ forms a closed convex subset of the cone, creating a manifold with corners.

When considering the correspondence between a graph $\mathscr{G}$ and its associated Laplacian matrix L, we can define the space of networks as the Laplacian space $\mathscr{L}_n$. It is important to note that $\mathscr{L}_n$ is not a Euclidean space; rather, it is a manifold with corners of dimension $n\frac{n-1}{2}$ [14]. Consequently, statistical techniques must be adapted to account for the unique geometric properties of this space.

To conduct statistical analysis on manifolds, it is common to work within the tangent space of the original space, using a suitable metric. By employing the extrinsic metric in $\mathscr{L}_n$, specifically the Procrustes distance, which involves optimizing over an orthogonal matrix $\mathbf{R}$ to achieve the Procrustes match between $\mathbf{X}$ and $\mathbf{Y}$, we can map each original element to a tangent space T_v at v [15]. This mapping can be accomplished by obtaining the coordinates on the tangent plane T_v, denoted as $\pi v^{-1}(\mathbf{X})$, for any elements $L_k \in \mathscr{L}_n$, where $k = 1, ..., m$, according to the following expression:

$$\pi^{-1}(\mathbf{X}) = \text{vec}\left\{\mathbf{H}(\mathbf{X}\hat{\mathbf{R}} - v)\mathbf{H}^T\right\}.$$

Here, vec represents the vectorize operator obtained by stacking the columns of a matrix, and $\mathbf{R}$ is the Procrustes match of $\mathbf{X}$ to $\mathbf{v}$.

Once the network data is projected onto a suitable Euclidean space, the principal components in the tangent space can be obtained by performing eigendecomposition on the empirical covariance matrix $S = \frac{1}{n}\sum_{k=1}^{n} v_k v_k^T$.

3 Understanding Allergic Sensitisation Dynamics through Graph Theory

3.1 Study population, IgE measurements and definition of clinical outcome

The Manchester Asthma and Allergy Study (MAAS), a population-based birth cohort study established in 1995 in Manchester (UK), includes a mixed urban-rural population within a 50-square-mile area of South Manchester and Cheshire[16]. Pregnant women were screened for eligibility during antenatal visits. Of the 1499 eligible couples, 288 declined participation, and 27 were lost to follow-up. A total of 1184 children were born into the study between 1996 and 1998. These children have been prospectively followed for 20 years, attending clinics for assessments such as lung function measurements, skin prick testing, biological sample collection, and questionnaires. The study received ethical approval from the North West – Greater Manchester East Research Ethics Committee.

We conducted measurements of IgE to 112 components derived from 51 sources using the ImmunoCAP ISAC (Thermo Fisher) at various ages: 1, 3, 5, 8, 11, and 16. The levels of component-specific IgE antibodies were reported in standardised units provided by the ISAC protocol. To simplify the IgE data, we discretised it using a binary threshold (positive ≥ 0.30 ISAC standardized units, according to manifacture's guidelines). Current asthma is defined as any 2 of the following 3 features: (1) current wheeze (positive answer to the question "*Has your child had wheezing or whistling in the chest in the last 12 months*?"); (2) *current use of asthma medication*; (3) *physician-diagnosed asthma ever*.

3.2 Results

The network data were defined considering the co-occurences of positive sIgEs responses. According to the methodology in section 2, we embedded the network data in a 2-dimensional Euclidea space (Figure 3.2). The analysis indicates that the temporal information, which potentially represents the progression of sensitizations, is effectively captured by the first principal component. Notably, when examining the differentiation between individuals with asthma (represented in red) and those without asthma (represented in black) at each time point analysed (ranging from age 5 to 16), the disparities between positive and negative asthma statuses are prominently highlighted within the Euclidean subspace. This observation highlights the valuable role of network analysis in investigating the heterogeneity of sensitisations and asthma.

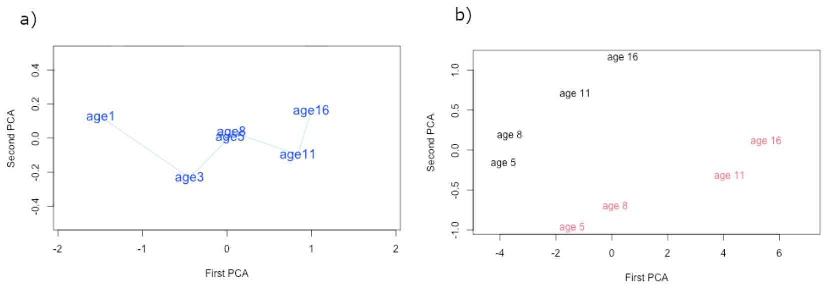


Fig. 1 Lower-dimensional representation of the sIgEs IgE network over time (a) and stratified by asthma status (b).

4 Conclusions

In conclusion, our study employed network analysis and statistical methodologies to explore the complex dynamics of IgE sensitisations and their relationship to asthma heterogeneity. By utilising network embedding techniques and dimensionality reduction methods, we were able to investigate the similarities and differences between IgE networks over time without imposing specific temporal models. The analysis revealed that the first principal component effectively captured the temporal information associated with sensitization development. Additionally, the differentiation between individuals with and without asthma was prominently highlighted in the Euclidean subspace, emphasising the potential of network analysis in understanding asthma heterogeneity. Our findings highlight the importance of considering the temporal evolution of IgE networks and suggest that network-based approaches can provide valuable insights into sensitsations and their associations with clinical outcomes. This research contributes to our understanding of the intricate interactions within the immune system and lays the foundation for further investigations into the underlying mechanisms of allergic diseases.

References

1. Custovic A., Lazic N., Simpson A.: Pediatric asthma and development of atopy. Curr Opin Allergy Clin Immunol 13:173-80 (2013)
2. Simpson B.M., Custovic A., Simpson A., Hallam C.L., Walsh D., Marolia H., et al.: NAC Manchester Asthma and Allergy Study (NACMAAS): risk factors for asthma and allergic disorders in adults. Clin Exp Allergy 31:391-9 (2001)
3. Sly P.D., Boner A.L., Bjorksten B., Bush A., Custovic A., Eigenmann P.A., et al.: Early identification of atopy in the prediction of persistent asthma in children. Lancet 372:1100-6 (2008)
4. Oksel C., Custovic A.: Development of allergic sensitization and its relevance to paediatric asthma. Current Opinion in Allergy and Clinical Immunology 18(2):p 109-116 (2018)
5. Pearce N., Pekkanen J. and Beasley R.: How much asthma is really attributable to atopy? Thorax. 54(3):268-72 (1999)
6. Dreborg S. and Frew A.: Position paper—allergen standardization and skin-tests. Allergy 48:49-82 (1993)
7. Prosperi M.C., Belgrave D., Buchan I., Simpson A. and Custovic A.: Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach. Pediatr Allergy Immunol. 25(1):71-9 (2014)
8. Simpson A., Tan V.Y., Winn J., Svensén M., Bishop C.M., Heckerman D.E., Buchan I. and Custovic A.: Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. Am J Respir Crit Care Med. 181(11):1200-6 (2010)
9. Simpson A., Lazic N., Belgrave D.C., Johnson P., Bishop C., Mills C. and Custovic A.: Patterns of IgE responses to multiple allergen components and clinical symptoms at age 11 years. J Allergy Clin Immunol. 136(5):1224-31 (2015)
10. Lazic N., Roberts G., Custovic A., Belgrave D., Bishop C.M., Winn J., Curtin J.A., Hasan Arshad S. and Simpson A.: Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. Allergy. 68(6):764-70 (2013)
11. Custovic A., Sonntag H.J., Buchan I.E., Belgrave D., Simpson A. and Prosperi M.C.F.: Evolution pathways of IgE responses to grass and mite allergens throughout childhood. Journal of Allergy and Clinical Immunology, 136(6):1645-1652 (2015)
12. Bishop C.M.: Model-based machine learning. Philos Trans A Math Phys Eng Sci. 2012 Dec 31;371:20120222 (1984)
13. Fontanella S., Frainay C., Murray C.S., Simpson A. and Custovic A.: Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A cross-sectional analysis within a population-based birth cohort. PLoS medicine 15.11: e1002691 (2018)
14. Ginestet C. E., Li J., Balachandran P., Rosenberg S. and Kolaczyk E. D.: Hypothesis testing for network data in functional neuroimaging. The Annals of Applied Statistics, pages 725–750, (2017)
15. Severn K.E., Dryden I.L. and Preston S.P.: Manifold valued data analysis of samples of networks, with applications in corpus linguistics.The Annals of Applied Statistics. 16 (1) 368 - 390 (2022). https://doi.org/10.1214/21-AOAS148
16. Custovic A., Simpson B.M., Murray C.S., Lowe L. and Woodcock A.: NAC Manchester Asthma and Allergy Study Group. The National Asthma Campaign Manchester Asthma and Allergy Study. Pediatr Allergy Immunol. 13(s15):32-7 (2002)

Investigating the patterns of Italian internal mobility: a network analysis at provincial level

La mobilità interna italiana: un'analisi di rete a livello provinciale

Annalina Sarra, Dario D'Ingiullo, Adelia Evangelista, Eugenia Nissi, Davide Quaglione and Tonio Di Battista

Abstract This study employs network analysis to investigate internal migration patterns in Italy from 2002 to 2018. Using filtering methods, the observed dense network was reduced to its essential components. The extracted backbone's characteristics were then examined, including measures of structural balance and community detection. Our findings suggest that the majority of signed links in the network exhibit reciprocity, and that positive relationships between geographically close-by cities and their shared locations at greater distances contribute to a structural balance. Additionally, a community analysis revealed that geography play a crucial role in defining communities within the network.

Abstract *Questa ricerca utilizza l'analisi delle reti per indagare i modelli di migrazione interna in Italia dal 2002 al 2018. Utilizzando metodi di filtraggio, la fitta rete osservata è stata ridotta alle sue componenti essenziali. Le caratteristiche del backbone estratto sono state quindi esaminate, inclusi i parametri di equilibrio strutturale e la rilevazione delle comunità. I nostri risultati suggeriscono che la maggior parte dei collegamenti con segno nella rete mostrano reciprocità, e che le relazioni positive tra le città geograficamente vicine e le loro destinazioni comuni a distanze maggiori contribuiscono ad un equilibrio strutturale. Inoltre,*

Annalina Sarra
University "G.d'Annunzio" of Chieti-Pescara, e-mail: annalina.sarra@unich.it

Dario D'Ingiullo
University "G.d'Annunzio" of Chieti-Pescara, e-mail: dario.dingiullo@unich.it

Adelia Evangelista
University "G.d'Annunzio" of Chieti-Pescara, e-mail: adelia.evangelista@unich.it

Eugenia Nissi
University "G.d'Annunzio" of Chieti-Pescara,e-mail: eugenia.nissi@unich.it

Davide Quaglione
University "G.d'Annunzio" e-mail: davide.quaglione@unich.it

Tonio Di Battista
University "G.d'Annunzio" e-mail: tonio.dibattista@unich.it

un'analisi delle comunità ha indicato che la geografia svolge un ruolo cruciale nella definizione delle comunità all'interno della rete.

Key words: Internal migration, Italy, Network analysis, Filtering methods, Community detection.

1 Introduction

Internal migration – that is the share of individuals who permanently transfer their place of residence within the national boundaries – affecting the spatial redistribution of people and their characteristics, is recognized to be an important mechanism able to influence several socio-economic outcomes [4]. Hence, investigating the structure and dynamics of the internal migrations should occupy a central stage on academic and policy debate in order to develop and put in place sound policies for a given economy. Studying internal migration patterns is crucial in countries like Italy, our area of focus, where significant territorial imbalances exist. It allows for a better understanding of the dynamics of population movements within the country's economy, highlighting areas of growth and areas that may require additional support. The literature strand on the determinants of Italian migration has demonstrated, in fact, how better socio-economic and institutional characteristics of the North induce a population redistribution towards this area, which in turn can further improve its already better structural characteristics (if these flows are endowed with a high level of education) [5]. By considering the consequences, in fact, the growing age and skill selectivity that characterizes the internal flows, contributing to spread skills and knowledge across regions, makes internal migration a crucial growth enhancing mechanism in the host economy [3, 1]. Therefore, this selectivity in migration contributes to feed a vicious circle to the detriment of the already backward regions of the *Mezzogiorno*. In this frame, an important contribution stems from an analytical tool largely adopted in several economic and non-economic fields: the network analysis. This methodology is particularly useful to provide an immediate representation of the relationships and functional linkages that exist among regions of a given economy, by combining measures of network connectivity with measures of network shape [2]. Building upon this, we extend the commonly used network analysis methods by incorporating a filtering technique that extracts both positive and negative migration links (i.e., a signed network) and by utilizing machine learning methods to analyze the low-dimensional representation of the migration system. This approach, as proposed by [7], allows for a more comprehensive analysis of the migration patterns and dynamics. In doing this, the present article makes use of a unique database that consists of bilateral migration flows among Italian regions (NUTS-3) collected by the Italian National Institute of Statistics (ISTAT) in the Migratory movements of resident population – registrations and cancellations to the registry office. The reminder of the article is structured as follows: Section 2 illustrates the data collection and the methodology. In Section 3 the results are presented and discussed.

2 Data and methods

2.1 Data

The bilateral migration flows among the 103 Italian provinces for the years 2002-2018 are collected by the ISTAT in the Migratory movements of resident population, registrations and cancellations to the registry office. For each year, by considering all the possible migration flows among the 103 Italian provinces, the resulting bilateral matrix of interprovincial movements is a square matrix that has a dimension of 103 x 103. Obviously, by excluding the main diagonal, which contains only zeros by definition (i.e. we do not consider the intra-provincial mobility), the total number of observations for a single year is equal to 10,506.

2.2 Methodology: Network analysis

Based on the approach outlined by [6], we implemented a four-step methodology to analyze the structure and dynamics of internal migration within Italy at the provincial level. Our first step involved constructing a directed and weighted network to model migration in-flows and out-flows between cities. Specifically, we created a distinct network for each year, in which links represented the movement of individuals from an origin city to a destination city, with the weight of each link indicating the number of people migrating. Formally, we represented the directed and weighted network as $G(V,E,W)$, where $|V|$ is the set of nodes denoted by $i, j, ...$, $|E|$ is the set of directed edges, and W is the weighted adjacency matrix. The network consisted of n nodes and m edges, with W_{ij} representing the number of migrants moving from node i to node j, and W_{ii} being equal to 0. In the second step, we aimed to extract relevant and significant information from the initial networks by simplifying and reducing them into a sparser format that preserves sufficient structural information for efficient analysis. This resulted in the creation of a meaningful signed network, denoted by $\hat{G} := (\hat{V}, \hat{A}, \hat{W})$, where the sparse adjacency matrix $\hat{A}$ typically only contains values of -1, 0, or 1. The sparse adjacency matrix is used to represent the links between nodes in the network, with a value of -1 indicating a negative link (an inhibitory relationship), a value of 1 indicating a positive link (an excitatory relationship), and a value of 0 indicating the absence of a link. The corresponding weight matrix, $\hat{W}$, is optional and only assigned to non-zero values in $\hat{A}$. If $\hat{A}_{ij} = 0$, then $\hat{W}_{i,j}$ is also 0, otherwise, $\hat{W}_{i,j}$ has the same sign as $\hat{A}_{i,j}$. This process involves information filtering, also known as “backbone extraction” or “network sparsification," which entails removing unnecessary or weak links in a network to extract its most important and informative connections. Various methods, including filtering techniques, can be used to achieve this "sparsification". We utilized a recent dense network filtering method proposed by [6] in 2021 to extract both positive and negative links. This approach relies on a null model that considers the in-strength and out-strength of nodes to estimate expected link weights and identify links with significantly different weights as positive or negative. The method can extract the network's signed backbone at a desired level of sparsity, based on the statistical significance of the links (via its significance filter) or the intensity of the links (via its vigor filter). The vigor filter is a lift-based measure that ranges from -1 to 1 and can be used to determine the strength of the links. By using this method, researchers can analyze important connections in a network and gain insights into its structure and function. Our subsequent objective was to create a concise representation of nodes by eliminating irrelevant information. This involved estimating an embedding matrix $\mathbf{Z}$ with dimensions $(n \times d)$, where n is the number of nodes, d is the embedding size, and z_i represents the embedding vector of node i. Our goal was to find a relation between $f(i, j)$ and $g(z_i, zj)$, where f is a similarity function in the observed network and g is a similarity function in the latent space. We chose cosine similarity as the similarity function, given that the input data's similarity range (i.e., edge weights in the signed backbone) is between -1 and 1. To complement the analysis of the migration system using lower dimensions, we implemented established clustering methods to identify community structures in the migration networks.

3 Results and discussion

In this section, we present the results of our study on migration patterns. In terms of the total number of migrants, we record a slow but often steady decrease in the number of people who have moved. The examination of the migration flows within the original network indicates that exists a reduced number of links that possess significantly greater weights. Additionally, the distribution of node strengths is characterized by a high degree of heterogeneity, with a few cities accounting for a disproportionate volume of migration. It is worth noting that, in the case of Italy's internal migration, there are migration flows between all pairs of cities in both directions, resulting in a network that is inherently dense. This highlights the need to filter out insignificant links. To achieve this, we employed information filtering methods that took into account the network's multiscale nature, thereby enabling the extraction of a sparse backbone. After analyzing the backbone's characteristics, it becomes apparent that most signed links in the network demonstrate reciprocity, as they are reciprocated by links bearing the same sign. Even when the size of the backbone is considerably large, conflicting links are rare. We also evaluated whether the extracted network exhibits a structural balance or weak structural balance (Fig.1). Our results indicate the presence of a structural balance among the nodes in our networks, which may be attributed to positive relationships between geographically close-by cities and their shared destinations at greater distances. Fig. 2 displays the spatial network of Italy's internal migration backbone in 2018.

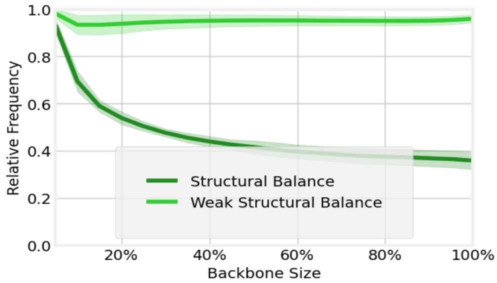


Fig. 1 Characteristics of migration backbones: structural balance

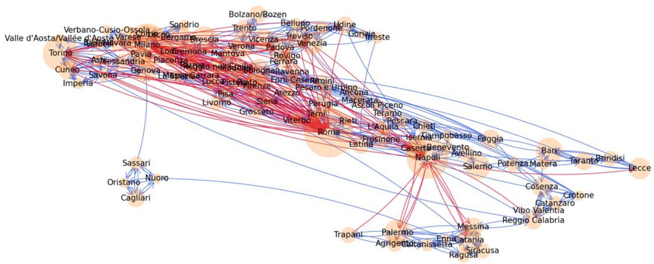


Fig. 2 Spatial network of migration backbone in 2018

Most of the positive links that extend beyond the local level in the network are oriented in a South-North direction. One possible explanation for this trend is that long-distance migrations typically occur between areas with significant economic hubs and activity centers that in Italy are located in the northern regions. We conclude our investigation with a community analysis. Figure 3 displays the results of the density-based clustering for both 2002 and 2018, using different colors for each cluster. Our analysis indicates that geographical proximity plays a major role in defining communities. The results suggest that areas that are closer to each other tend to be clustered together, regardless of other factors such as population density or economic activity. This finding has important implications for understanding patterns of regional development and social interaction.

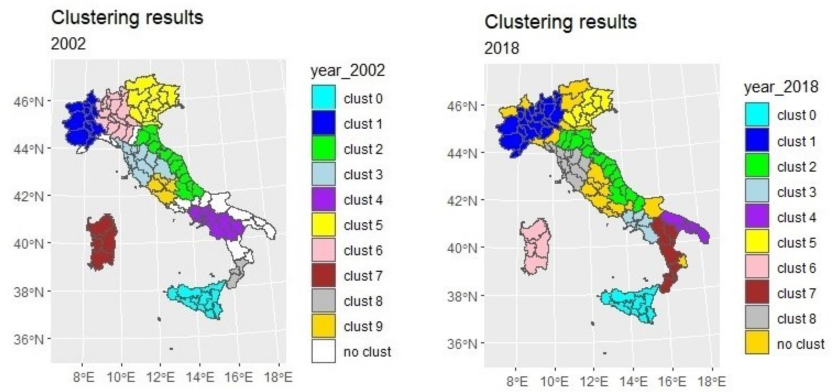


Fig. 3 Density-based clusters

References

1. Basile, R., Girardi, A., Mantuano, M., Russo, G.: Interregional Migration of Human Capital and Unemployment Dynamics: Evidence from Italian Provinces. Ger. Econ. Rev., 1-30 (2018) doi:10.1111/geer.12172
2. Carvalho, R. C. de, Charles-Edwards, E.: The evolution of spatial networks of migration in Brazil between 1980 and 2010. Popul. Space Place, 26(7) e2332 (2020) doi:10.1002/psp.2332
3. Dotti, N. F., Fratesi, U., Lenzi, C., Percoco, M.: Local labour market conditions and the spatial mobility of science and technology university students: evidence from Italy. Rev.Reg. Res. 34(2), 119-137 (2014) doi: 10.1007/s10037-014-0088-y
4. Etzo, I.:Internal migration: a review of the literature. MPRA Paper Number. 8783 (2008)
5. Etzo, I.:The determinants of the recent interregional migration flows in Italy: A panel data analysis. J. Reg. Sci. 51(5), 948-966 (2011) doi:10.1111/j.1467-9787.2011.00730.x
6. Gürsoy, F., Badur, B.: Extracting the signed backbone of intrinsically dense weighted networks. J.Complex Netw. 9(5):cnab019.(2021) doi: 10. 1093/ comnet/ cnab0 19
7. Gürsoy, F., Badur, B.: Investigating internal migration with network analysis and latent space representations: an application to Turkey. Soc. Netw. Anal. Min. 12(1), 1–16 (2022) doi:10.1007/s13278-022-00974-w

Solicited Session SS12 - *Innovations and challenges in official statistics*

Organizer and Chair: Matteo Mazziotta

1. *Formal and informal networks of care for the elderly: regional profiles compared* (Sicuro L., Tucci D. and Coniglio R.)
2. *Gender Gap: a multidimensional approach* (Acampora C., Fusco D., Liguori M.A. and Pagliuca M.M.)
3. *Using Whatsapp in Official Statistics: a New tool for managing the Agriculture Census* (Fabi C.)

Formal and informal networks of care for the elderly: regional profiles compared

Le reti formali e informali di assistenza alle persone anziane: modelli regionali a confronto

Lorella Sicuro, Domenico Tucci and Rosalia Coniglio

Abstract In recent decades the elderly population has been growing gradually over time without interruption. The purpose of this study is to analyse the needs of Italian frail elderly and the care received, taking into account the primary care, demographic and social context thanks to the use of 24 indicators from ISTAT and Ministry of Health sources. The MPI (Mazziotta-Pareto Index) method was applied to elaborate six synthesis indices. In the South there are high values of potential needs and informal care. The North is characterized by a greater presence of residential and territorial supply.
Abstract *Negli ultimi decenni la popolazione anziana è cresciuta gradualmente nel tempo. Si analizzano, a livello regionale, i bisogni degli anziani fragili e l'assistenza ricevuta, tenendo conto dell'assistenza primaria e del contesto socio-demografico attraverso 24 indicatori di fonte ISTAT e Ministero della Salute. Si è applicato il metodo MPI (Mazziotta-Pareto Index) per elaborare sei indici di sintesi. Al Sud ci sono alti valori di bisogni e di assistenza informale. Il Nord è caratterizzato da una maggiore presenza di offerta residenziale e territoriale.*

Key words: elderly, home care, residential care, aging, primary care, chronic conditions, disability, informal aid

---

Lorella Sicuro
Istat; e-mail: sicuro@istat.it

Domenico Tucci
Istat; e-mail: dotucci@istat.it

Rosalia Coniglio
Istat; e-mail: coniglio@istat.it

1 Background

In recent decades, the elderly population (over 65) has been growing gradually over time without interruption. According to forecasts, in two decades about 33 percent of the population in Italy will be elderly [6]. The increasing aging and advances in survival express improved living conditions and medical developments, but pose the problem of the quality of life years gained [7]. Therefore, the elderly, as their years of life increase, see the conditions that characterize a good quality of life put at risk [1]. Data referring to 2019 shows that in Italy the over-65s with severe chronic diseases and multi-morbidity are 33.3 percent; among the over-85s it grows to 47.7 percent. There are 3.8 million elderly people with severe reduction of autonomy in daily activities or in domestic life and one million (6.9%) with need for assistance [9]. Although demographic changes in family structure reduce the help that they can give as potential caregivers, the family is the main support for the elderly [2]. However, care pathways consisting of an organizational set of health and social treatments are necessary [3]. To achieve this goal, territorial care, also known as primary care, must be strengthened. Forms of home hospitalization allow to elderly to remain in their usual living environment, that is essential to stabilize the clinical picture and limit functional decline, besides it can avoid access and stay at hospital or residential facilities (dangerous, for example, in an epidemic context, as the covid-19 epidemic has shown) [5]. The functions of monitoring and management of care are carried out by the Ministry of Health through the Regions. The purpose of this study is to analyse, at the regional level, the needs of Italian frail elderly and the care received, both informal and formal, taking into account the presence of primary care and social factors in the area.

2 Materials and Methods

Twenty-four indicators from ISTAT and Ministry of Health surveys related to the health status of the elderly population and health care supply were elaborated. Since this is a complex and multidimensional phenomenon, it was decided to synthesize the information through the MPI (Mazziotta-Pareto Index) method [8]. Specifically, six synthetic indices - outlined below - related to the following areas were constructed: potential needs, demo-social-health context, residential supply, complementary and alternative territorial supply, hospital inappropriateness and informal care.

Potential need: Severe chronicity, Multi-morbidity, Severe difficulties in at least one personal care activity, Severe difficulties in household activities for at least one activity, Difficulties in mobility or functional limitations;

Demo-social-health context: Lonely elderly, Health perception: bad and very bad, Over 80 years of age out of total population;

Residential supply: Guests of the residential socio-medical institutions for the self-sufficient elderly, For the non self-sufficient elderly, Severe difficulties in personal care activities due to aids deemed insufficient or missing;

Complementary and alternative territorial supply: **S**ocial-medical home care, Users of day care centres, Users of vouchers, Receiving telemedicine and tele assistance, Users of community/social centres, Home care, Total home care hours per elderly person;
Hospital inappropriateness: Hospitalizations with over threshold stay with medical DRG of patients aged 65 years and older, Hospitalization rate for heart failure (age >= 65 years), Hospitalization rate for influenza in the elderly;
Informal care: Severe difficulties in personal care activities for aid received from family members, severe difficulties in personal care activities due to aid received from paid persons.
MPI method makes it possible to construct a non-compensatory synthetic measure based on the property of non-substitutability (complete or partial) of the components. It consists of aggregating, through the arithmetic mean, the elementary indicators transformed by the method of standardized deviations, corrected by a penalty coefficient that takes into account the variability of the indicators: it penalizes the regions that, with the same mean value, have a greater imbalance between the elementary indicators. We have divided the indices into tertiles and refigured through cartograms.

3 Results

In the South, there are high values of potential needs and informal care. The North is characterized by a greater presence of residential and territorial supply. (Table 1).

3.1 Regional profiles

For the North, in Piemonte, Liguria and Valle d'Aosta, the needs of the elderly are not very high. The demo socio-health context registers criticalities, with a strong presence of residential socio-medical and socio-health supply and with, except to Piemonte, territorial health offerings. Lombardia, Veneto, Friuli-Venezia Giulia and Trentino Alto Adige have the lowest needs, also with a non-critical socio-health demo context, and with high residential, territorial and hospital supply. Emilia Romagna is similar to that of Centre of Italy: the needs of the elderly as well as the demo-social characteristics are of medium intensity, which are responded to with all types of care (informal, formal and hospital).

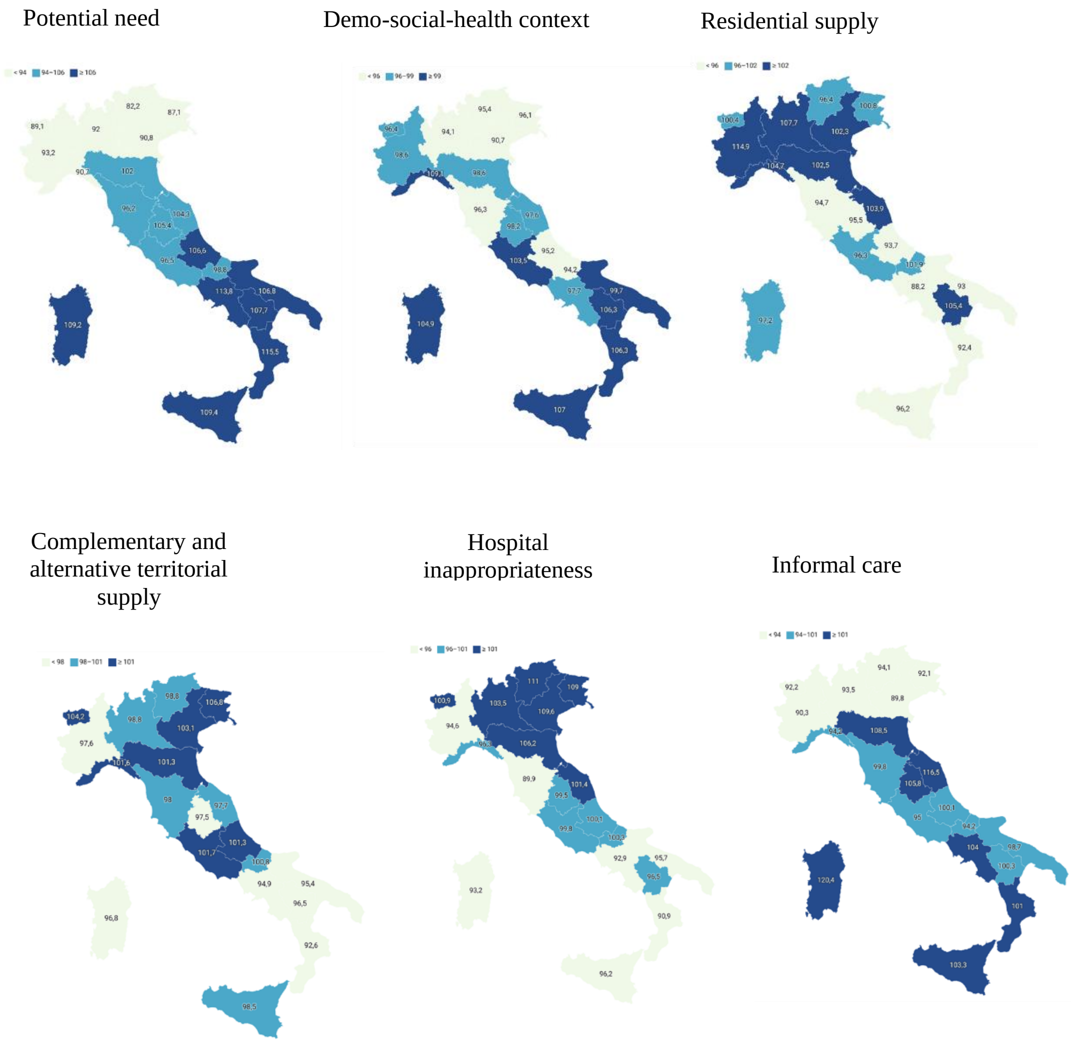
Table 1 Mazziotta-Pareto Synthesis indices related to health, formal and informal care, hospital inappropriateness, at regional level, 2019, Italy.

Regions	***Potential need***	***Demo-social-health context***	***Residential supply***	***Complementary and alternative territorial supply***	***Hospital inappropriateness***	***Informal care***
Piemonte	93,2	98,6	114,9	97,6	94,6	90,3
Valle d'Aosta/Vallée d'Aoste	89,1	96,4	100,4	104,2	100,9	92,2
Lombardia	92	94,1	107,7	98,8	103,5	93,5
Trentino Alto Adige/Sudtirol	82,2	95,4	96,4	98,8	111	94,1
Veneto	90,8	90,7	102,3	103,1	109,6	89,8
Friuli-Venezia Giulia	87,1	96,1	100,8	106,8	109	92,1
Liguria	90,7	109,1	104,7	101,6	96,3	94,2
Emilia-Romagna	102	98,6	102,5	101,3	106,2	108,5
Toscana	96,2	96,3	94,7	98	89,9	99,8
Umbria	105,4	98,2	95,5	97,5	99,5	105,8
Marche	104,3	97,6	103,9	97,7	101,4	116,5
Lazio	96,5	103,5	96,3	101,7	99,8	95
Abruzzo	106,6	95,2	93,7	101,3	100,1	100,1
Molise	98,8	94,2	101,9	100,8	100,3	94,2
Campania	113,8	97,7	88,2	94,9	92,9	104
Puglia	106,8	99,7	93	95,4	95,7	98,7
Basilicata	107,7	106,3	105,4	96,5	96,5	100,3
Calabria	115,5	106,3	92,4	92,6	90,9	101
Sicilia	109,4	107	96,2	98,5	96,2	103,3
Sardegna	109,2	104,9	97,2	96,8	93,2	120,4

For the Centre of Italy, Toscana presents medium potential needs and a prevalence of territorial supply and informal care. Marche is characterized by average needs and social demo context, which is met by high reliance on all help networks, formal and informal, including hospital inappropriateness. Lazio, one of the largest regions in

Italy, shows high levels of needs, also critical the social situation, which, however, is met with all networks of help, especially territorial care. (Figure 1)
For the South of Italy, Abruzzo meets the needs with informal care, territorial supply and hospital inappropriateness. Residential supply is rather lacking. Campania and Sardegna assume very similar behaviours, responding to the needs mostly with the greater presence of informal care (Sardegna records high values of residential care). Calabria, Sicilia and Puglia respond to the needs with informal care, Sicilia also takes high values of territorial supply. Basilicata is the southern region with the highest residential supply, followed by informal care.

Fig. 1 Territorial distribution by region of the six synthesis indices.



4 Conclusion

The alternative supply (territorial care) that allows to the elderly to remain with family and at home, is still scarcely present, in particular Central and Southern Italy record the greatest criticalities. In these geographical repartitions, the primary care is "compensated" by informal care, but it does not still allow to be provided with specific, health-related help. Therefore, there should be a strengthening of territorial supply and greater coordination between hospital and territory for the taking care of frail individual. In addition, there are forms of social non-self-sufficiency (typical of Northern Italy and large cities), determined by loneliness and the weakness of the solidarity network: interventions aimed at activating processes of integration and inclusion are needed, offering the elderly the possibility of still feeling like a "resource," an active part of life and citizenship. There are critical issues in the analysis at the regional level: although the regions are primarily responsible for planning and managing the health care system, there may be differences in health care resources and efficiency of care within the local health units of the same region. Besides, a single regional territory may contain different geographic areas (such as mountainous areas and areas of high or low population density), and the continuity of people's social and health integration is managed locally. The need emerges for targeted, structured, multidisciplinary interventions that ensure continuity of care centered on the needs of the patient as well as family, social, and geographic context [4]. It will be interesting to assess the trend and intra-regional variability identifying virtuous regions requires not only high levels of care for the frail elderly, but also uniformity within them.

References

1. Burgio, A., Battisti, A., Solipaca, A., Colosimo, S. C., Sicuro, L., Damiani, G., Baldassarre, G., Milan, G., Tamburrano, T., Crialesi, R., Ricciardi, W.: La relazione tra offerta di servizi di Long Term Care ed i bisogni assistenziali dell'anziano. Contributi Istat, n.4/2010.
2. Censis: L'Italia e le dinamiche demografiche. Scenari e strumenti per affrontare il futuro (2021)
3. Cricelli, C., Brignoli, O., Medea, G., Parretti, D., Lomabardo, F.P., Lora Aprile P., lapi, F., Cricelli, I.: Impatto epidemiologico delle cronicità in medicina generale. Rapporto Osservasalute 2020
4. Damiani, G., Colosimo, S. C., Sicuro, L., Burgio A., Battisti, A., Solipaca, A., Baldassarre, G., Crialesi, R., Milan G., Tamburrano T., Ricciardi W.: An ecological study on the relationship between supply of beds in long-term care institutions in Italy and potential care needs for the elderly. BMC Health Services Research 2009, 9:174 doi:10.1186/1472-6963-9-174 (2009)
5. Damiani, G., Michelazzo, M.B.: Assistenza territoriale, Rapporto Osservasalute 2020 (2020)
6. Istat: Invecchiamento attivo e condizioni di vita degli anziani in Italia. ISBN 978-88-458-2028-1 (2020)
7. Livi Bacci, M.: Introduzione alla demografia. Loescher editore (1999)
8. Mazziotta, M., Pareto, A.: Un indice sintetico non compensativo per la misura della dotazione infrastrutturale: un'applicazione in ambito sanitario. Istat. Rivista di statistica ufficiale n.1/2011 (2011)
9. Statistiche Report Istat: Le condizioni di salute della popolazione anziana in Italia. Anno 2019 (2019)

Gender Gap: a multidimensional approach

Divario di genere: un approccio multidimensionale

Cira Acampora, Daniela Fusco, Maria A. Liguori and Margherita M. Pagliuca

Abstract At international level, there are numerous indices built for measure gender gap. In 2013, European Institute for Gender Equalilty (EIGE) released the Gender Equality Index (GEI), created to assess the levels of gender equality in Europe based on EU policies. This index is currently based on seven domains: work and money, knowledge, time, power, health, trust and safety, quality and life satisfaction. Starting from the domains proposed by the GEI, integrated with some of the indicators of the Sustainable Development Goals (SDGs), a regional measure for calculating gender differences in Italy has been identified. Eight dimensions were identified and a ninth of a more specific nature was added, called "Gender-based violence".

Abstract *A livello internazionale esistono numerosi indici costruiti per la misurazione del divario di genere. Nel 2013, l'Istituto europeo per l'uguaglianza di genere (EIGE) ha rilasciato il Gender Equality Index (GEI), allo scopo di valutare le disparità di genere a livello europeo. Questo indice si basa attualmente su sette domini: lavoro e denaro, conoscenza, tempo, potere, salute, fiducia e sicurezza, qualità e soddisfazione della vita. Partendo da tali domini, integrati con alcuni indicatori dei Sustainable Development Goals (SDGs), è stata identificata una misura regionale delle disparità di genere in Italia. Sono state individuate otto dimensioni più una specifica, chiamata "Violenza di genere".*

Key words: Social, Well-being, Gender

Cira Acampora
Istat, e-mail: acampora@istat.it

Daniela Fusco
Istat, e-mail: dafusco@istat.it

Maria Antonietta Liguori
Istat, e-mail: liguori@istat.it

Margherita Maria Pagliuca
University of Naples Parthenope; e-mail: margherita.pagliuca@uniparthenope.it

1 Introduction

The EU Gender Equality Strategy delivers on the von der Leyen Commission's commitment to achieving a Union of Equality. The Strategy presents policy objectives and actions to make significant progress by 2025 towards a gender-equal Europe [3]. Despite the progress of recent years, with 63.8 out of 100 points, Italy ranks 14th in the EU on the Gender Equality Index. Its score is 4.2 points below the EU's score. Since 2010, Italy's score has increased by 10.5 points, raising its ranking by seven places. Since 2018, Italy's score has shown no change (only + 0.3 points) and its ranking has dropped by one place.

It is important to underline how much gender inequalities depend on territorial contexts: milieu, logistics and local opportunities can aggravate inequalities. From this point of view, the historic gap between the South and the North of Italy remains profound also with regard to gender gap [7]. The COVID-19 pandemic has negatively affected the fragile gains made since 2010. For the first time in a decade, gender inequalities in employment (full-time equivalent employment rate (FTE) and duration of working life), education (tertiary graduation and participation in formal or informal education and training), health status and access to health services have grown [4]. In order to measure the changes in gender inequalities with the arrival of the pandemic and how they differ in the Italian regions, this work intends to identify a measurement of the phenomenon from 2018 to 2021.

2 Methodology and some results

At international level, there are numerous indices built for measure gender inequalities. Already in 1995, the UNDP developed the Gender-related Development Index (GDI), which considered three domains: life expectancy and health, knowledge and standard of living. In 2013, EIGE released the Gender Equality Index (GEI), created to assess the levels of gender equality in Europe based on EU policies. This index is currently based on seven domains: work and money, knowledge, time, power, health, trust and safety, quality and life satisfaction.

Starting from the domains proposed by the GEI, integrated with some of the indicators of the Sustainable Development Goals (SDGs), a regional measure for calculating gender differences in Italy has been identified.

The second phase concerned the identification of inequality measurement tools. For each domain, a research of the sources was carried out in order to identify the most appropriate indicators to represent the phenomenon. At the end, 14 sources have been chosen, statistical and administrative, for a total of 52 indicators divided in eight domains plus one (table 1).

Table 1 Number of indicators for each domain

Domains	*Number of indicators*
Knowledge	7
Work and money	11
Power	4
Health	4
Well-being	5
Trust and safety	4
Quality and life satisfaction.	5
Time	6
Gender-based violence	6
Total	52

Each indicator is calculated for male, female and total, while the ratio female on male gives the gap. Therefore, the gap is plus then one if it is in favour of female, minus then one if it is in favour of male.

To reduce the complex and multidimensional nature of these phenomena, it has been applied a methodology known as composite indicators or composite index [6]. A composite indicator is formed when individual indicators are synthesized into a single index on the basis of an underlying model. Composite indicators are much like mathematical or computational models. As such, their construction owes more to the craftsmanship of the modeler than to universally accepted scientific rules for encoding. For this research is applied the Adjusted Mazziotta–Pareto Index (AMPI). The AMPI is a non-compensatory (or partially compensatory) composite index that allows comparability of the data across units and over time [5]. It is a variant of the Mazziotta–Pareto Index (MPI), based on a re-scaling of the individual indicators by a Min–Max transformation, in contrast with the classic MPI where all the indicators are normalized by a linear combination of z-scores [2].

3 Final remarks

Over the last decade, the EU has made progress towards gender equality, albeit at a snail's pace. While the EU is considered to be a global leader in gender equality, this obscures the diversity of achievements at national level [1].

In Italy the gap is evident also at regional level. The Italian gender equality strategy, approved by The National Recovery and Resilience Plan (NRRP) emphasises the need for further advances in tackling gender gaps and unequal labour-market participation. For this reason, regional monitoring of the phenomenon is necessary.

The choice of the 52 indicators was a complex choice, weighted considering:

- Data availability. Only regional indicators classifiable by gender have been taken into consideration;
- Feasibility. The availability of obtaining and processing updated data in a simple way has been taken into account;
- Timeliness of the data to ensure an adequate time comparison;
- Thematic appropriateness.

The idea is the construction of a composite index for each domain because the loss of information deriving from the calculation of a single index, which further summarizes the domains, would lead to the exclusion of the possibility of undertaking this choice. However, the construction of a regional ranking, alongside the reading of individual domains, could represent an important information input for monitoring the phenomenon over time. The application of the methodology, beyond the merely empirical value, would be useful for the political decision-maker, considering the programmatic choices made by our nation in the context of the national strategy for gender equality.

References

1. Bisello, M., Caisl, J., Maftei, A., Peruffo, E., Barbieri, D., Reingarde, J., Mascherini, M.: Promoting social cohesion and convergence Upward convergence in gender equality: How close is the Union of equality? Luxembourg: Publications Office of the European Union. European Institute for Gender Equality (2016)
2. De Muro, P., Mazziotta, M., Pareto, A.: Composite indices of development and poverty: An application to MDGs. Social Indicators Research, 104, 1–18 (2011)
3. European Commission: Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions A Union of Equality: Gender Equality Strategy 2020-2025 Brussels, 5.3.2020 COM152 final (2020)
4. European Institute for Gender Equality: Gender Equality Index 2022: The COVID-19 pandemic and care. Publications Office of the European Union. Luxembourg (2022)
5. Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. Social Indicators Research, 127, 983–1003 (2016)
6. OECD: Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing (2008)
7. Pastorelli, E., Stocchiero, A.: Le disuguaglianze in Italia. ENGIM Internazionale/FOCSIV (2019)

Using Whatsapp in Official Statistics: a New tool for managing the Agriculture Census

L'uso di Whatsapp nella Statistica Ufficiale: uno strumento di gestione del Censimento dell'Agricoltura

Claudia Fabi

Abstract The data collection of the Italian 7th General Census of Agriculture took place from January to July 2021, through a synchronous multi-technique (CAPI, CATI, CAWI) design. During the survey, innovative contact tools were offered to respondents, alongside traditional communication channels. Through SMS and Whatsapp it was possible to collect the requests of over 2,000 farms, facilitating subsequent contacts with the interviewers. The following work intends to show the results obtained, the propensity to use innovative contact channels and proposes a model that allows to correlate this use to some representative indicators of the territory characteristics, demographics and the diffusion of the so-called "digital culture".

Abstract *La raccolta dati del 7° Censimento Generale dell'Agricoltura si è svolta da gennaio a luglio 2021, attraverso un disegno che prevedesse l'utilizzo della multitecnica sincrona (CAPI, CATI, CAWI). Nel corso della rilevazione sono stati offerti ai rispondenti strumenti innovativi di contatto, in affiancamento ai tradizionali canali di comunicazione. Attraverso SMS e Whatsapp è stato possibile raccogliere le istanze di oltre 2.000 aziende agricole, agevolando i successivi contatti con gli intervistatori. Il seguente lavoro intende mostrare i risultati ottenuti, la propensione all'utilizzo di canali di contatto innovativi e proporre un modello che permetta di correlare tale utilizzo ad alcuni indicatori rappresentativi del territorio, demografici e della diffusione della cosiddetta "cultura digitale".*

Key words: Census, Agriculture, Whatsapp, SMS, respondent support

---

Claudia Fabi

ISTAT – Istituto Nazionale di Statistica; Via Cesare Balbo 16 – 00184 Rome (Italy); e-mail: claudia.fabi@istat.it

1 The reference context

Between January and July 2021, the survey of the seventh General Census of Agriculture took place in Italy, the last one before the advent, also in this sector, of annual Permanent Censuses, on sample basis.

For the first time, in an Italian census survey, a synchronous multi-technique design (CAWI, CAPI, CATI) was adopted, which allowed respondents a large amount of flexibility in choosing how to fill in the questionnaire [1,2].

Important technological innovations, necessary to ensure the fluidity of data collection operations conducted by different data collection networks, have been supported by new strategies of contact with respondents. In fact, starting from May 2021, a new contact channel for respondents was introduced, through a mobile phone number to which users could send SMS or WhatsApp messages, with the aim of requesting an appointment to subsequently complete the interview by phone. The main results obtained will be illustrated below, examining in detail the diffusion and propensity to use innovative contact tools on the national territory, compared with the reference population of the Census.

2 Characteristics of the SMS and WhatsApp service in supporting respondents

The communication and assistance channels dedicated to the respondents of a survey, whether a census or a sample one, can be classified as synchronous or asynchronous[1]. The former, usually managed by operators are represented in most cases by free toll-free numbers, which users can contact to request general information on the survey, on how to participate or ask for assistance in completing it, on specific thematic issues.

However, even a synchronous tool such as the toll-free number has potential critical issues in its use, especially in some time slots of the day or in periods of intense inbound traffic, concentrated in the weeks immediately following the massive sending of information letters to the population involved in the survey. In this sense, SMS and WhatsApp are a hybrid channel of support for respondents, being able to guarantee both synchronous – when managed by operators – and asynchronous support.

In the case of the Agricultural Census, the mobile telephone number, available for sending of SMS and WhatsApp by the respondents, was managed by operators on the same days and times as the CATI survey[2] and remained available as an asynchronous

[1] A communication channel is defined "synchronous" when provides a contextual interaction between user and operator (e.g. Toll Free Number, Contact Center, chat), while it's defined as asynchronous the communication channel through which the user forwards a request that will be processed at a later time by the operators who supervise the service (e.g. email, forum, etc.).

[2] The telephone survey was carried out from Monday to Friday, from 9.00 to 21.00 and on Saturday from 10.00 to 19.00, except for public holidays.

channel for the rest of the time, including holidays. Users became aware of the existence of this innovative channel by calling the toll-free number or, for some of them, through the memorandum sent to the units still not responding in June 2021, by paper letter. The service was activated with the sole aim of offering to respondents a means of requesting an appointment to complete the questionnaire with CATI technique, and not as a complete information channel. It should be remembered that the number to send SMS and WhatsApp messages was activated only later, in relation to the start of the survey, and in particular only for the last three months of fieldwork: from May to July 2021.

Table 1 Channel used by respondents (SMS or WhatsApp)

Channel	***Users***	
	A.V.	***%***
SMS	195	9.3
WhatsApp	1,908	90.7
TOTAL	2,103	100.0

Table 2 SMS and WhatsApp users by age group[3]

Age group	***SMS or WA users***		***Census List***	
	V.A.	***%***	***V.A.***	***%***
Up to 30 years	15	0.7	27,816	1.7
From 31 to 40 years	63	3.1	82,882	5.2
From 41 to 50 years	210	10.3	177,750	11.1
From 51 to 60 years	381	18.8	320,700	20.1
From 61 to 70 years	581	28.6	370,022	23.2
From 71 to 80 years	438	21.6	339,186	21.2
Over 80 years	343	16.9	279,555	17.5
TOTAL	2,031	100.0	1,597,911	100.0

3 An explanatory model of the use of SMS and WhatsApp

Given that the use of innovative communication channels such as SMS and WhatsApp to support surveys of official statistics in Italy is still largely experimental, and that the diffusion of digitalization habits still penalizes some geographical areas and some segments of the population[4], it seemed reasonable to deep the analysis through the comparison between a set of indicators selected to represent the transversal dimension of "*smart farmers*" and the use of SMS and WhatsApp, in the census survey.

[3] Missing any other source of information, the distribution by age was obtained starting from the year of birth present in the tax ID code of the units in the census list. Where this information was not available, either in the census list or for SMS and WhatsApp users, the record was excluded from the calculation of the frequency distribution.

[4] See, for example, the results of the Survey "Aspects of Daily Life" - ISTAT.

In particular, the indicators [3] by Italian region reported in Table 4 were considered with the aim of identifying and aggregating, through a composite index, two main dimensions:

- propensity of individuals to use digital communication channels;
- development and entrepreneurship in the Agricultural Sector.

Furthermore, in the choice, were included only the indicators for which was available the annual data for 2021 or, in its absence, for 2020 at regional level, in order to represent phenomena contemporary to the data collection period of the Census.

Table 4 Indicators selected for the composite index.

Indicator	***Content***	***Year***
Ind_460	Employment rate in rural areas (15-64 years)	2021
Ind_062	Percentage of diffusion of the Internet in families	2021
Ind_250	Growth rate of agriculture	2021
Ind_031	Agricultural land productivity	2021
AVQ_1	People aged 6 and over who use the Internet	2021
AVQ_2	Possession of at least 1 mobile phone in the family	2021
AVQ_3	People who have used the Internet to request information to PA Entities in the last 12 months	2021
AVQ_4	Use of social networks in the last 3 months	2020
ISTR_CP	People with at least a high school diploma	2021

The elaboration of a composite index was computed following the guidelines suggested by Mazziotta-Pareto for the synthesis of non-compensatory composite indexes [4] using the normalization formula that follows:

$$z_{ij} = 100 \pm \frac{(x_{ij} - M_{x_j})}{S_{x_j}} 10$$

where M_{x_j} e S_{x_j} are, respectively, the mean and the standard deviation of the indicator *j* and the polarity, in the case of the subset of indicators chosen, is the positive one.

The composite index was computed following the formula:

$$MPI_i = M_{z_i} - S_{z_i} cv_{z_i}$$

where $cv_{zi} = S_{zi} / M_{zi}$ is the coefficient of variation for the unit *i*.

The following Table 5 shows MPI and the correlation with the percentage of users of SMS and WhatsApp, by region, computed respect to the total number of units in the census list. The Autonomous Provinces of Trento and Bolzano values have been computed separately. The result, somewhat surprising, is a clear absence of correlation between the propensity to use innovative channels of communication with the PA and the PMI index calculated as detailed above. The two determinants of the phenomenon, digitization, and development of the agricultural sector, do not seem to have an impact on the use of SMS or WhatsApp to communicate with institutions.

Table 5 MPI Index by Region

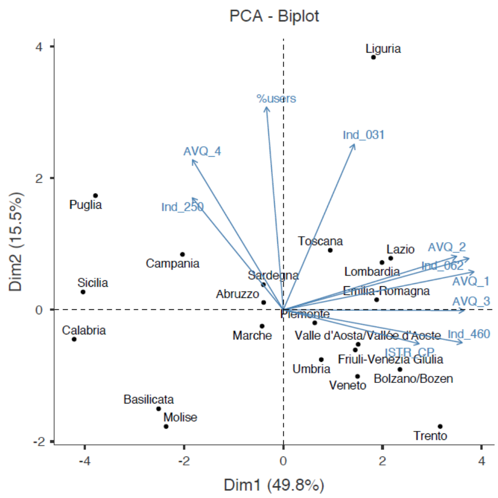
Regione	***MPI***	***% Users***
Piedmont	98.8	0.15
Valle d'Aosta	102.4	0.09
Lombardy	105.2	0.14
Trentino Alto Adige/Südtirol		
- *Bolzano/Bozen*	*106.5*	*0.03*
- *Trento*	*103.1*	*0.05*
Veneto	99.3	0.08
Friuli-Venezia Giulia	101.4	0.12
Liguria	109.1	0.21
Emilia-Romagna	104.1	0.12
Tuscany	104.2	0.12
Umbria	98.0	0.13
Marche	98.9	0.12
Lazio	105.7	0.16
Abruzzo	98.8	0.14
Molise	91.8	0.06
Campania	96.5	0.08
Puglia	93.2	0.17
Basilicata	91.1	0.09
Calabria	88.1	0.11
Sicily	89.2	0.13
Sardinia	100.0	0.08
Coefficient of correlation		*0.098*

However, this is not entirely true: in fact, by calculating the simple correlations between each indicator and the percentage of users in the census list, more sensitive correlations emerge with indicators Ind_031 and AVQ_4 (respectively the use of social networks and productivity of agricultural land), drivers also highlighted in the *biplot* among regions and the matrix of indicators (Figure 1). There is therefore a "*smart farmers*" phenomenon, as the correlations seem to suggest, even if this phenomenon does not significantly affect the propensity to use SMS and WhatsApp among the reference population.

4 Conclusions

The correlation that emerges between individual/sectoral indicators and the use of SMS and WhatsApp, albeit slight, seems to suggest an expected tendency to favor these innovative channels exactly by people who possess dynamic characters and propensity to digitalization, even in the agricultural sector that is traditionally managed by small operators of high average age.

Fig. 1 Biplot regions - indicators



Furthermore, the absence of correlation with the PMI also suggests that investing in a digital communication channel such as WhatsApp to support respondents, is not an exclusive choice, in the literal sense of the term. The widespread diffusion of this tool among the population makes it extremely interesting, right now, as an additional resource to complement the more usual contact strategies with respondents.

References

1. De Gaetano, L., Fabi, C., Triolo, V.: Tecniche integrate di data collection per il Censimento Generale dell'Agricoltura 2020. In: Poster scientifici 14° Conferenza Nazionale di Statistica (2021) https://www.istat.it/storage/14-Conferenza-nazionale-statistica/poster/022.jpg
2. De Gaetano, L., Fabi, C., Triolo, V.: 7th Agriculture Census: strategies, and methodological and technological innovations for the survey. In: Book of Abstracts of 7th Italian Conference on Survey Methodology, Perugia (2022) https://www.dropbox.com/s/a9uuajjvb7gbyd9/BoA%20_itacosm2022.pdf?dl=0
3. Indicatori territoriali per le politiche di sviluppo – ISTAT (2023) https://www.istat.it/it/archivio/16777
4. Mazziotta, M., Pareto, A.: On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phoenomena (2015) https://link.springer.com/article/10.1007/s11205-015-0998-2

Solicited Session SS13 - *Statistical methods and composite indicators for healthcare*

Organizer: Maria Gabriella Grassia and Corrado Crocetta
Chair: Paolo Mariani

1. *Longitudinal composite indicators to measure the quality of health services* (Crocetta C., Antonucci L., Cataldo R. and Mazza R.)

2. *Past and Future of Doctor-Patient Communication* (Tedesco N., Zavarrone E. and Forciniti A.)

3. *Network Analysis approach to customer satisfaction and service quality detection: an application to health-care services* (Crocetta C., Grassia M.G., Marino M., Mazza R., Simonacci V. and Stavolo A.)

4. *A project evaluation study on multiset Likert scale data* (Simonacci V., Marino M., Grassia M.G. and Gallo M.)

Longitudinal composite indicators to measure the quality of health services

Indicatori compositi longitudinali per misurare la qualità dei servizi sanitari

Corrado Crocetta, Laura Antonucci, Rosanna Cataldo and Rocco Mazza

Abstract The study of the quality of services provided in the health sector plays a fundamental role in applied statistical studies. The work focuses on using the Structural Equation Modeling Path Modeling as a valuable way to analyze longitudinal data of the satisfaction expressed by a selection of public and private operators. The aim of the study is to develop a experimental model for measuring satisfaction with the services provided by an external organization (Sanita'Service s.r.l.) at the ASL (local sanitary organization) agencies on Foggia.

Abstract *Lo studio della qualità dei servizi erogati in ambito sanitario riveste un ruolo fondamentale negli studi statistici applicati. Il lavoro si focalizza sull'utilizzo dei modelli ad equazioni strutturali come valido strumento per analizzare i dati longitudinali della soddisfazione espressa da diversi operatori pubblici e privati. L'obiettivo dello studio è quello di sviluppare un modello sperimentale per misurare la soddisfazione per i servizi forniti da un'organizzazione esterna (Sanita'Service s.r.l.) presso le ASL di Foggia.*

Key words: Health services, PLS-PM, Longitudinal study

Laura Antonucci
Department of Economics, University of Foggia, Italy e-mail: laura.antonucci@unifg.it

Rosanna Cataldo
Department of Social Sciences, University of Naples, Italy e-mail: rosanna.cataldo2@unina.it

Corrado Crocetta
Department of Economics, University of Foggia, Italy e-mail: corrado.crocetta@unifg.it

Rocco Mazza
Department of Political Science, University of Bari, Italy e-mail: rocco.mazza@uniba.it

1 Introduction

The study of the quality of services provided in the health sector plays a fundamental role in applied statistical studies [16]. In recent years, this field has also started a process of outsourcing many activities [8]. This new organizational assessment implies committing key services or operations of an organization to a provider specialized in that area [3]. The reduce maneuver costs behind the support operation made by outsourcing are the primary motivations of this organizational approach [21]. Hospitals are able to reduce costs by outsourcing secondary maintenance to third-party operators, the expenses that healthcare companies can reduce in this way include those associated with personnel, training and equipment [2]. This allows healthcare institutes to free up resources, improve efficiency and customer satisfaction, and focus on delivering high-quality core services [18]. The topic has been addressed by Yesildag et al. (2022) [23] and Guimarães (2011) [8], who defined a picture of the phenomenon by comparing various international organizations with each other. The Italian case, on the other hand, has been addressed in Mariani et al. (2014) [15]. This paper focuses on the study of the satisfaction expressed by a selection of public and private operators in order to intercept the different approaches of these actors. In particular, work on the development of a quantitative model for the evaluation of services can be found in Langabeer (2008) [13]. In this regard, it becomes very important to investigate the satisfaction perceived by the management of healthcare facilities with the third-party companies providing these services. Satisfaction itself becomes a very complex concept in a healthcare landscape such as the Italian one, where the capillary network of public healthcare companies throughout the territory offers very complex logistical and organisational planning. It therefore becomes very important to study a method for evaluating the actions taken, in terms of outsourcing especially in the cost containment and efficiency area [14]. Hence the need to construct a model capable of capturing not only the multidimensionality of satisfaction with respect to services but also capable of intervening on the organisational complexity that this context brings with it. The aim of our study is to develop a experimental model for measuring satisfaction with the services provided by an external organization (Sanita'Service s.r.l.) at the ASL (local sanitary organization) agencies on Foggia. The reference time range of this study is the years from 2019 to 2021 and the considered aspects are Auxiliarity and Cleaning. The work focuses on using the Partial Least Squares- Path Modeling (PLS-PM) as a valuable way to analyze longitudinal data in order to obtain a synthetic indicator able to synthesize these aspects and monitor them over the time. The years considered in the analysis represent the most turbulent years in recent history for Italian healthcare due to the COVID-19 pandemic. This has certainly impacted on the organization of these institutions and on the organizational dynamics that have regulated them. In the following paragraphs, we will examine in detail the method used to develop the model, the data collection tool and the selection of individuals carried out and the results obtained.

2 Method

A longitudinal study refers to an investigation where participant outcomes are collected at multiple times. A common design for longitudinal research in the social sciences is the panel or repeated measures design, in which a sample of subjects is observed at more than one point in time. If all individuals provide measurements at the same set of occasions, we have a fixed occasions design. Longitudinal studies provide important information about service incidence and satisfaction trajectories. Recent developments in statistical methods for analyzing longitudinal data provide efficient estimates of change and predictors of change over time [6]. The model on longitudinal data can be approached from several perspectives, and the model can be constructed as a Structural Equation Model (SEM). Recently, Roemer [19] has proposed using the component-based approach to SEM PLS-PM in a longitudinal study. In agreement with Roemer [19], we state that PLS-PM is highly appropriate for an analysis of development and change in constructs in longitudinal studies, since PLS-PM is suited to handle such complex models [7], typical of data longitudinal [12], and with a rather small sample size [10]. The PLS-PM approach is useful to investigate the course of effects (path coefficients between the constructs over time [1]. The path coefficient is fundamentally evaluated by two criteria: by the magnitude value (the greater the value between 0 and 1, the greater the magnitude of the effect between the constructs), and statistical significance (the 95% confidence interval) [11].

3 The PLS-PM model

In this study, nine constructs were investigated through a PLS-PM in three different time: 2019 (t_0), 2020 (t_1), and 2021 (t_2). The constructs are: Auxiliary-t_0, Cleaning-t_0, CS-t_0, Auxiliary-t_1, Cleaning-t_1, CS-t_1, Auxiliary-t_2, Cleaning-t_2, CS-t_2. The theoretical PLS-PM model proposed is represented in the Fig. 1. The model

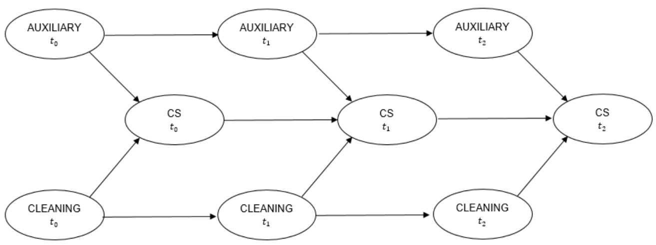


Fig. 1 The theoretical PLS-PM model

is a hierarchical model: CS, in the different time period is the hierarchical construct related to its two concrete subdimensions, measuring Auxiliary and Cleaning. Different approaches have been developed adn proposed in literature [5]. In this work the hierarchical dimensions have been estimated with the Mixed Two Step approach, according to which, first, a hierarchical model is formed by all MVs of the two dimensions and the algorithm is performed. Next, the scores for each blocks obtained in the first step of the algorithm are used as the Manifest Variables (MVs) of the hierarchical dimension and the PLS-PM algortihm is performed again [4]. The study focuses on reflective-formative approach: each of these dimensions is measured by own MVs and the relationship between them and the respective block is assumed to be reflective (every Latent Variable (LV) is the reflection of the MVs to which it is connected), with these two dimensions having an impact on the CS in a formative way. This kind of reflective-formative modelization is a particularly important and widely adopted type of model, frequently used in research in the social sciences [17], particularly in the construction of composite indicators.

4 PLS-PM results

The survey data used in this work were collected with the use of a questionnaire administered to the employees of the ASL of Foggia between 2019 and 2022. Head nurse, primaries and managers of the directorate general are interviewed, who are asked to evaluate the services offered by the provider company. The questionnaire is structured in modules, in the first part we request information from the interviewee on the district of affiliation, its role and the area of competence. In the second part there are evaluation modules on the company's services: auxiliary, cleaning, maintenance, transport, emergency, management, hospitality. Assumed the experimental and preliminary nature of the study, we decided to focus exclusively on auxiliary and cleaning services. The data relating to the 2022 questionnaire appear to have too many missing to be processed well, related to questionnaires not yet received; for this reason we have decided to collect only the years from 2019 to 2021. The analysis was performed using the "plspm" package in the R programming language [20] in order to perform PLS-PM analysis involving formative indicators. Any missing data was handled by using the NIPALS algorithm [24], while the centroid scheme was chosen [9] for the model and it was estimated with a maximum of 1,000 iterations [9]. Before applying the PLS-PM approach, an exploratory Principal Component Analysis (PCA) has been performed in order to analyse the relationship of elementary indicators for each construct. In order to evaluate the model it is important to consider some synthesis and quality measures of the model, in term of reliability and construct validity (in Table 1) and the path coefficients and the R-square (in Table 2).
Composite Reliability and AVE measures exceed the conventional acceptablity threshold of 0.7 and so they are appreciably higher for the different time periods. In Table 1 the GoF criterion is reported. Although this approach has been proposed for PLS-PM, it mainly serves a diagnostic purpose and is not used for formal testing

Table 1 Assessment Measures

Construct	Composite Reliability	AVE	Construct	Composite Reliability	AVE	GoF
AUX (t_0)	0,966	0,825	AUX (t_2)	0,960	0,798	
CLE (t_0)	0,981	0,827	CLE (t_2)	0,979	0,808	
CS (t_0)	0,982	0,964	Cs (t_2)	0,965	0,932	0,798
AUX (t_1)	0,945	0,740				
CLE (t_1)	0,971	0,759				
Cs (t_1)	0,968	0,938				

[22]. Even if the GoF is lower than the threshold of 0,90, it represents an acceptable value.

Table 2 Path coefficients linking the two dimensions to CS in different time period

LVs	**P.C.**	**S.E.**	**CI (95%)**	R^2
Auxiliary $t_0 \rightarrow$ CS t_0	0,513	0,012	[0,494;0,547]	
Cleaning $t_0 \rightarrow$ CS t_0	0,505	0,007	[0,490;0,517]	0,986
Auxiliary $t_0 \rightarrow$ Auxiliary t_1	0,349	0,08	[0,192;0,493]	
Auxiliary $t_1 \rightarrow$ CS t_1	0,533	0,020	[0,496;0,578]	
Cleaning $t_0 \rightarrow$ Cleaning t_1	0,237	0,09	[0,045;0,403]	
Cleaning $t_1 \rightarrow$ CS t_1	0,499	0,014	[0,471;0,527]	
CS $t_0 \rightarrow$ CS t_1	0,003	0,003	[0,001;0,005]	0,942
Auxiliary $t_1 \rightarrow$ Auxiliary t_2	-0,081	0,142	[-0,332;0,182]	
Auxiliary $t_2 \rightarrow$ CS t_2	0,508	0,013	[0,485;0,537]	
Cleaning $t_1 \rightarrow$ Cleaning t_2	-0,043	0,128	[-0,275;0,184]	
Cleaning $t_2 \rightarrow$ CS t_2	0,528	0,010	[0,509;0,551]	
CS $t_1 \rightarrow$ CS t_2	0,002	0,005	[-0,016;0,005]	0,869

The fitting R^2 for CS was high in all time periods. The results of path coefficients β showed that the auxiliarity and cleaning dimensions have a moderate and significant effects on satisfaction in all periods. The curious thing to note is that in the two years before the COVID-19 pandemic the dimension that has a greater impact is the Auxiliary, while after the pandemic the most important dimension is Cleaning. Therefore, it is important to point out that the carry-over effects are positive and significant for Auxiliarity and Cleaning from t_0 to t_1, while the situation is reversed from t_1 to t_2. This negative and insignificant effect, that changes from 2020 to 2021, may depend on various factors, one of which may be the COVID-19 pandemic and all its consequences. These factors will be investigated. Moreover, we are waiting for the complete survey data of 2022 and results of the 2023 survey to include them in our analysis and understand if this reversal is confirmed or the years considered have been a bit turbulent for italian healthcare.

References

1. Alwin, D. and Hauser, R.: The decomposition of effects in path analysis. American Sociological Review. pp. 37-47 (1975)
2. Buxbaum, J.: Spotlight on HIEs and EHRs. How outsourcing fits in. Health Management Technology. 32, 20-21 (2011)
3. Carr, L. and Nanni, A.: Delivering results: managing what matters. (Springer,2009)
4. Cataldo, R., Grassia, M.G., Lauro, N.C., and Marino, M.: Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators. Quality and Quantity, Springer, 51(2), 657-674 (2017)
5. Crocetta, C., Antonucci, L., Cataldo, R., Galasso, R., Grassia, M., Lauro, C. and Marino, M.: Higher-order PLS-PM approach for different types of constructs. Social Indicators Research. 15: 725-754 (2021)
6. Duarte-Guerra, L., Kortchmar, E., Maraviglia, E., Silva Costa, T., Lasmar, C., Morin, R., Grossi, I., Villares, J., Cremonesi, M., Watanuki, H. and Others: Longitudinal patterns of comorbidity between anxiety, depression and binge eating symptoms among patients with obesity: a path analysis. Journal Of Affective Disorders. 303: 255-263 (2022)
7. Fornell, C. and Cha, J.:Advanced Methods of Marketing Research, ed. RP Bagozzi, Blackwell, Cambridge. (1994)
8. Guimarães, C. and Carvalho, J.: Outsourcing in the healthcare sector-a state-of-the-art review. Supply Chain Forum: An International Journal. 12, 140-148 (2011)
9. Hair, J.F., Hult, G.T.M., Ringle, C.M. and Sarstedt, M.: A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2nd ed. Thousand Oaks. CA: Sage (2017)
10. Henseler, J., Ringle, C.M. and Sinkovics, R.R.:The use of partial least squares path modeling in international marketing. New challenges to international marketing, 277-319, Emerald Group Publishing Limited (2009)
11. Hu, L. and Bentler, P.: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 6, 1-55 (1999)
12. Jones, E., Sundaram, S.and Chin, W.:Factors leading to sales force automation use: A longitudinal analysis. Journal of personal selling and sales management, 22 (3), 145-156, Taylor and Francis. (2002)
13. Langabeer, J.: Health care operations management: a quantitative approach to business and logistics. (Jones and Bartlett Learning,2008)
14. Macinati, M.: Outsourcing in the Italian National Health Service: findings from a national survey. The International Journal Of Health Planning And Management. 23, 21-36 (2008)
15. Mariani, P., Falotico, R. and Zavanella, B.: Outsourcing services in the Italian National Health Service: The evaluation of private and public operators. Procedia Economics And Finance. 17: 256-264 (2014)
16. Ozcan, Y.: Quantitative methods in health care management: techniques and applications. (John Wiley and Sons,2005)
17. Ringle, C., Sarstedt, M. and Straub, D.: Editor's comments: a critical look at the use of PLS-SEM in" MIS Quarterly". MIS Quarterly. pp. iii-xiv (2012)
18. Roberts, J., Henderson, J., Olive, L. and Obaka, D.: A review of outsourcing of services in health care organizations. Journal Of Outs. And Org. Inf. Man. pp. 1 (2013)
19. Roemer, E.:A tutorial on the use of PLS path modeling in longitudinal studies, Industrial Management and Data Systems, 116 (9), 1901-1921,Emerald Group Publishing Limited. (2016)
20. Sanchez, G.: PLS path modeling with R. Berkeley: Trowchez Editions, 383, 551 (2013)
21. Sunseri, R.: Outsourcing on the outs. Hospitals and Health Networks. 73, 46-48 (1999)
22. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y. M., Lauro, C. N.: PLS Path Modeling. Computational Statistics and Data Analysis, 48 (1), 159 - 205 (2005)
23. Yesildag, A. and Sayar, B.: The effect of outsourcing on quality in hospitals: a systematic review. International Research In Health Sciences IV. pp. 89 (2022)
24. Wold, H.: Path models with latent variables: The NIPALS approach. In Quantitative sociology (pp. 307-357). Academic Press (1975)

Past and Future of Doctor-Patient Communication

La comunicazione tra medico e paziente tra passato e futuro

Nicola Tedesco, Emma Zavarrone and Alessia Forciniti

Abstract The health of the patient benefits from clear and effective doctor-patient communication since the patient is more likely to follow the doctor's instructions and experience less stress. The evolution of the communication forms between doctor and patients has been investigated through the Natural Language Process to identify semantic identity, regularities and capture differences over time..
Abstract *La salute del paziente beneficia dalla comunicazione chiara ed efficace con il medico poiché è più probabile che il paziente possa seguire le istruzioni del medico e sperimentare meno stress. La evoluzione delle forme di comunicazione tra medico e paziente è stata indagata attraverso il processo del linguaggio naturale al fine di identificare identità semantiche, regolarità e catturare le differenze nel tempo.*

Key words: patient-doctors communication, clinician–patient, CBOW, empathy, semantic regularity, collocation analysis

1 Background

A good doctor-patient communication contributes to the patient's health by making things easier to understand and reducing stress; moreover, it helps people admit they have health problems, know their treatment options, change their behaviour, and take their medications [4].It would also give the patient the chance to make

Nicola Tedesco
Univ. di Cagliari, e-mail: tedesco@unica.it

Emma Zavarrone
Iulm University, Milano, e-mail: emma.zavarrone@iulm.it

Alessia Forciniti
Iulm University, Milano, e-mail: alessia.forciniti@iulm.it

choices and give informed consent. Due to the complexity of the profession, doctors should theoretically exercise advanced communication skills that relate not only to the ability to collect information to make a correct diagnosis, guide patients in the right direction and give them instructions for treatment, but also to building care relationships with them. Unfortunately, these skills are disregarded by too many doctors, leading to low levels of empathy and compassion. Developing empathy in a doctor-patient relationship improves the final outcomes of care and reduces the risk of legal action and physical aggression. Notwithstanding the central role that communication plays in the doctor- patient relationship, a systematic evaluation procedure has not been implemented in the various health facilities yet, although this aspect has been largely investigated in the literature and multiple interpretative keys have been proposed. This paper deals with discovering how the communication dimensions have governed the physician- patient relationship and testing its invariance over time.

2 Methods

Under the perspective of the mapping review [3] that means to seek and to identify an interpretable pattern of similar terms, we adopt a Natural Language Process for exploring the words linked to the doctor-patient communication applying the Word2vec approach [2]. Before applying Word2vec, an exploratory textual analysis has been conducted through preliminary operations: textual pre-treatment, keyword in context (KWIC) and collocation analysis. Textual pre-treatment consists of: removal of symbols, numbers, punctuation marks, stopwords for the English language based on Smart system, stemming, TFIDF weight scheme adoption and Part Of Speech approach (POS). The KWIC and collocation analysis help outline the scenario which is not always easily understood under the bag of words approach. KWIC examines the word position inside the sentence with its occurrences and it allows to determine the deep meaning and way of use of the investigated term [5] .The collocations can be defined as a statistically significative association between co-occurring terms [1] using two measures:λ (the natural log of the odds ratio) and z score. Word2vec learns word representations and creates vocabulary from the from POS results by training a three-layered neural network. Word2vec can be applied through two models, namely Continuous Bag of Words (CBOW) and Skipgram. CBOW learns representations by making predictions about the target word based on its context words, whereas Skipgram learns representations by making predictions about each context word based on the target word. CBOW model consists of three layers: input layer, hidden layer and output layer. These layers are connected by two weight matrices $\mathbf{W}$ and $\mathbf{W}^{'}$. The input layer takes the one hot vectors of target word applying softmax function. Error between the original and predicted vectors is back propagated to update the weight matrices $\mathbf{W}$ as well as $\mathbf{W}^{'}$. Finally for each word in the vocabulary of given corpus, two vectors $\mathbf{V}_c$ and $\mathbf{V}_w$ are obtained (i.e., $\mathbf{V}_c$ is from $\mathbf{W}$ and $\mathbf{V}_w$ is from $\mathbf{W}^{'}$). A set of negative sampling weights has been assumed.

In our paper, we adopt the CBOW to study the communication physician-patients focusing on its most important dimensions.

3 Data and descripitive statistics

The data collection extends from 1958 to 2023 and consists of 64500 abstracts from the Scopus database selected through keywords: doctor-patient relationships, patient-specialist communication, medical empathy. A corpus was created and the usual text pre-processing operations were applied. The final size of the corpus was reduced to 44828 abstracts Table 1.

Table 1 Time series of selected abstracts

Years	Abstracts
1953-1993	1100
1994-2004	13506
2005-2015	8196
2016-2023	22026
Total	44828

The general description highlights the mention of the dimensions related to communication in the first ten terms Table 2.

Table 2 Distribution of TFIDF terms

Terms	TFIDF	Terms	TFIDF
Patient	126.74	Inform	108.88
Care	125.42	Doctor	104.75
Health	115.35	Nurse	102.41
Empathy	109.69	Treatment	97.21

Moreover, we add a context description through KWIC analysis and collocation analysis. These results portray an uncertain landscape where communication dimensions are not well defined. The KWIC reveals that the pool of the communication dimensions has a low level of presence, i.e. empathy has only been cited in 43 articles (0,09%) in which the phrases are very close to the aspects of communication. The collocation analysis confirms the scenario in which empathy continues to be far from the top (table 4) even though communication dimensions characterize the corpus. A similar pathway to empathy does not represent a driver of communication and leads to widely underestimated results.

Table 3 Collocation Analysis: top six results

Collocation	count	λ	z
Health care	11949	3.84	326.18
Primary care	7423	5.22	260.93
Communication skill	5380	4.47	244.61
Decision making	4938	4.26	235.67
General practioners	2812	8.01	226.16
Quality life 2014	6.01	201.75	

4 Results and final remarks

The CBOW results (Table 4) reinforce the configurated scenario, we focus on the terms communication.

Table 4 Comparison results of CBOW for Communication over time

Overall[a]	Q_1	Q_1	Q_1
Protocol [0.993]	interpersonal [0.942]	provider[0.924]	engagement [0.947]
Member [0.993]	expertise [0.923]	clinician [0.923]	empowerment [0.922]
Beneficiary [0.991]	empathy [0.920]	engagement[0.923]	skills[0.920]
organization[0.991]	empathic[0.910]	enhanced[0.918]	motivation[0.912]

[a] *In the square brackets the similarity value computed with CBOW*

These results tell us that until 2023 the soft skills devoted to the care of the patient-clinicians communication have contributed to sediment the literature on relationship but their importance does not seem to be perceived. The second research question helps us in understanding the true state of the art for the patient-clinicians communication over time: observing the investigated span of the time and we suppose a sort of halo effect. We decided to create three subpopulations of abstracts based on the three time slots corresponding to the quartile positions (Qi, for i=1,2,3). We repeated the textual procedures, and we discovered a transition from the basic medical service to the communication of the medical service. During the last time slot, from 2019 to 2013, this transition highlights that terms like engagement, empowerment, skills and motivation have a high value of similarity with communication.

References

1. Lyse, G. I., and Andersen, G.: Collocations and statistical analysis of ngrams. Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Nor-

wegian: 49 (2012)

2. Paré G., Trudel M.-C., Jaana M., Kitsiou S.: Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 2,183–199 (2015)
3. Mikolov, T., Yih, W. T. and Zweig, G.: Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human language technologies 75 (2013, June)
4. Ranjan P, Kumari A, Chakrawarty A.: How can Doctors Improve their Communication Skills? J Clin Diagn Res. 9(3) (2015)
5. Xiao, R. and McEnery, T.: Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. Applied linguistics, 1 103–129 (2006)

Network Analysis approach to customer satisfaction and service quality detection: an application to health-care services

Network Analysis per la rilevazione della customer satisfaction e della qualità del servizio: un'applicazione ai servizi sanitari

Corrado Crocetta, Maria Gabriella Grassia, Marina Marino, Rocco Mazza, Violetta Simonacci and Agostino Stavolo

Abstract In health-care systems, outsourcing refers to the practice of contracting external providers to perform certain functions or services that are traditionally performed by health system. So, health systems can often achieve cost savings by reducing the need for in-house staff and resources. It can also present challenges where patient confidentiality and data privacy are paramount. The objective of the paper is to assess the satisfaction levels of health systems with the outsourced services they utilize. To accomplish this, the paper employs Social Network Analysis (SNA)

Corrado Crocetta
University of Bari Aldo Moro; e-mail: corrado.crocetta@uniba.it

Maria Gabriella Grassia
University of Naples Federico II; e-mail: mariagabriella.grassia@unina.it

Marina Marino
University of Naples Federico II; e-mail: marina.marino@unina.it

Rocco Mazza
University of Bari Aldo Moro; e-mail: rocco.mazza@uniba.it

Violetta Simonacci
University of Naples Federico II; e-mail: violetta.simonacci@unina.it

Agostino Stavolo
University of Naples Federico II; e-mail: agostino.stavolo@unina.it

to create a methodology that considers both customer satisfaction and quality detection.
Abstract *Nei sistemi sanitari, l'outsourcing si riferisce alla pratica di stipulare contratti con fornitori esterni per svolgere determinate funzioni o servizi che tradizionalmente sono svolti dal personale del sistema sanitario. In questo modo, i sistemi sanitari possono spesso ottenere risparmi sui costi riducendo la necessità di personale e risorse interne. Essa, però, presenta criticità in quanto deve rispettare il principio di riservatezza dei pazienti e la privacy dei dati. Dunque, l'obiettivo del lavoro è definire il livello di soddisfazione che i sistemi sanitari esprimono nei servizi esterni ai quali si riferiscono. Per fare ciò è stata utilizzata la Social Network Analysis (SNA) al fine di sviluppare una strategia che consideri la customer satisfaction e la quality detection.*

Key words: social network analysis, health-care services, outsourcing, service quality detection, customer satisfaction

1 Introduction

Applied statistical studies play a crucial role in assessing the quality of services provided in the health sector [15], assessing various factors such as patient satisfaction, safety measures and efficiency, to identify areas of improvement and develop strategies for enhancing the quality of care. In recent years, this field has also started a process of outsourcing many activities [10].

Outsourcing involves delegating essential services or operations of an organization to a provider who specializes in that area [6]. One of the primary motivations behind hospital executives' decision to outsource support services is to lower operational costs [17], by reducing expenses associated with personnel, training, and equipment [5]. This approach enables healthcare institutions to free up resources, improve efficiency and customer satisfaction, and focus on delivering high-quality core services [16].

It is essential for decision-makers to ensure that outsourced services are being carried out by competent professionals, regardless of the nature of the service being provided [11], but also must ensure strict compliance with all relevant regulations, with a special focus on maintaining data privacy and security.

The phenomenon has been investigated through different perspectives and with different methodologies of analysis. Social network analysis (SNA) has been extensively used [1] to help improve the effectiveness and efficiency of decision-making processes [9] that improve patient safety and quality of care. Therefore, the objective of this study is to employ social network analysis (SNA) to assess the quality of Apulian outsourced healthcare services and develop a strategy that considers the overall satisfaction level of the services provides.

2 Methodology

Social Network Analysis (SNA) refers to a set of methods and techniques that detect the connections within a network, as well as the roles and positions that individuals play within it [12]. Networks are defined as systems in which actors, or so-called nodes, are connected to each other via links [18]. They are classified as *one*-mode networks, i.e. with a single set of nodes that have links between them, but there are also two-mode networks, or bipartite networks [13], that have two different sets of nodes and identify relationships between classes of actors [4]. Many mathematical techniques are available to measure networks [18]. Degree centrality (d_i^*) is a measure of the centrality of a node *i* in a network based on the number of connections. It is defined by $d_i^* = d_i / n - 1$ and it is normalized by dividing by the maximum number of ties possible, which in a graph of *n* nodes is *n-1*. Betweenness centrality is a measure of how often a node acts as a bridge along the shortest path between two other nodes. The betweenness of node k in an ordinary graph is defined by:

$$b_k = 1/2 \sum_{i,k}^{n} \sum_{j,k}^{i} \frac{g_{ikj}}{gij}$$

where g_{ij} is the number of geodesic paths from node *i* to node *j*, and g_{ikj} is the number of geodesic paths from *i* to *j* that pass-through *k*. Betweenness is ordinarily normalized by dividing by (*n*-1) (*n*-2).

Density is the number of existing links in a network divided by the maxim number of possible links:

$$D = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}$$

A high density indicates a higher level of interconnection between members.

Diameter is the length of the longest geodesic in the bipartite graph, within components [4]. Social Network Analysis (SNA) has found wide application in various fields and disciplines. It is often used to improve decision-making processes and organizational performance in business organizations, and to inform the implementation of change management programs [7]. The information flow develops targeted interventions aimed at improving the quality and efficiency of care delivery [8;14]. For monitoring performance reports, outsourcing management should prioritize the softer factors that underpin the organizational relationship, such as commitment, trust, respect, and shared understanding [2]. Effective relationship management can enhance outsourcing governance by enabling managers to efficiently monitor the quality of the relationships with service providers [3]. It has also proven useful in communication and collaboration among healthcare professionals, the dissemination of new practices, and knowledge sharing among clinicians [10].

3 Data

For this paper, we utilized survey data collected between 2020 and 2022 from employees of the ASL of Foggia through a questionnaire. We submitted the questionnaire to head nurses, primaries, and managers of the directorate general to assess the quality of services offered by the provider company. The questionnaire is divided into modules, with the first part seeking information on the interviewee's district of affiliation, role, and area of competence. The second part contains evaluation modules on the company's services, including auxiliary, cleaning, maintenance, transport, emergency, management, and hospitality. The third part - on which our study is based - is about the relation between ASL sanitary and manager staff, and the services offered by provider. The formal interactive process between these actors would involve several steps, defined by organizational hierarchy. The questionnaire aims to trace the informal social structure inside sanitary organization as respond to Covid emergency. In this regard we interviewed not only manager staff but also head nurse. The questionnaire was administered to the same individuals over the course of three years, so the information contained refers to the evolution of relationships over the time range considered.

4 Results

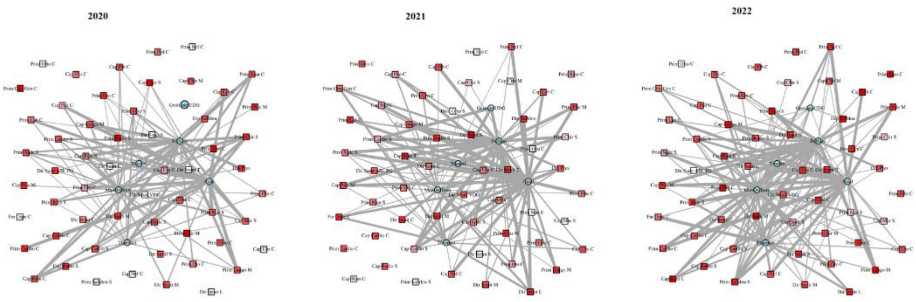


Fig. 1 Network evolution in 2020, 2021 and 2022

Figure 1 illustrates the changes in network evolution over three years, namely 2020, 2021, and 2022. Specifically, the nodes of the actors have been indicated with the rectangle symbol; the circles represent the services they refer to. The different shades of red indicate the level of satisfaction: the closer to pink, the lower the satisfaction rating. Rectangles with a white color indicate a non-response. Finally, the thickness of the link was weighted according to the frequency of monthly contact: the more double the link, the more the service was contacted by the subject.

The network consists of a two-mode network, encompassing the social relations occurring between two categories of actors. In the comments, we will refer to the dimension related to asl employees in terms of overall satisfaction with the services offered by the provider. We have also included the structural indices of the network over the time range analyzed. The mode related to services is analyzed with respect to the centrality they occupied within the network.

Table 1 Structural measures and satisfaction

Network	***Density***	***Diameter***	***Average satisfaction***
2020	0.37	3.00	8.15
2021	0.40	4.00	7.54
2022	0.50	4.00	8.26

The 2020 network is characterized by a higher number of isolated nodes, which characterize the lower level of network density (Tab. 1). Management of the Territorial Care Unit is identified as an isolated node, as can be seen in Fig1. In 2021, this service has become widely used, with thick links indicating a higher frequency of contacts. In 2022, service utilization was highest, indicating an increasing need to outsource activities outside hospitals, also demonstrated by a higher value of network density (0.51). As shown in Tab.1, the density measures over the years increased, indicating a progressive increase in the social activity carried out by the service-related actors. Despite this, the network for the first year has a smaller diameter value than the others, as networks with a smaller diameter tend to be more efficient in transmitting information. Considering the average satisfaction level, we see that it remained high in all three years, with a value of 8.15 in 2020. However, in 2021 there was a decrease to 7.5, followed by an increase in 2022 to 8.26, indicating the highest satisfaction with the services. So, we can say that satisfaction with the services offered by SanitaService appears to have high values overall, denoting another quality of the proposed activities.

Table 2 Centrality measures of networks

	2020		***2021***		***2022***	
Services	***Degree***	***Betweenness***	***Degree***	***Betweenness***	***Degree***	***Betweenness***
ManOrdImm	0,483	0,125	0,483	0,114	0,69	0,291
Tfarmaci	0,328	0,045	0,31	0,303	0,552	0,135
Tmalati	0,172	0,005	0,086	0,042	0,121	0,002
PulSani	0,466	0,096	0,483	0,101	0,569	0,14
Ausil	0,414	0,091	0,569	0,147	0,534	0,117
GestioneUDG	0	0	0,052	0	0,121	0,003

Through feedback on perceptions of the quality of services provided, possible areas for improvement can be identified. The degree centrality measure identifies the actors or services that are central in the network according to their average distance from other actors or services in the network. According to the Tab.2, we can state that in 2020 the services that were most in demand were regular maintenance of properties, cleaning and sanitization and auxiliary services. We can note that in 2021 the auxiliary service becomes the main one to be used, followed by maintenance and Cleaning.

This could be due to the emergency from COVID-19, which has increasingly led to the need for activities related to the securing of buildings through disinfestation and sanitization activities, but also to the need for external auxiliary personnel consisting of both personal care and support activities for healthcare facilities. In 2022, on the other hand, the auxiliary service is replaced by medicines transport. Indeed, the betweenness centrality of a service indicates the number of times it is on the shortest path between two other services through an actor. According to the measure, the services reported earlier are confirmed for 2020.

However, in 2021, the transportation of medicines emerges as the service with the highest value, indicating its increased importance. In the last year, however, the situation changes again, and building maintenance activities regain their importance.

References

1. Bae, S. H., Nikolaev, A., Seo, J. Y., Castner, J.: Health care provider social network analysis: a systematic review. Nursing outlook, 63(5), 566-584 (2015)
2. Beimborn, D.: Considering the relative relevance of outsourcing relationship quality. In Proceedings of the 20th European Conference on Information System (ECIS). AIS Electronic Library (AISeL) (2012)
3. Beimborn, D., Jentsch, C., Lüders, P.: Measuring Outsourcing Relationship Quality: Towards a Social Network Analysis Approach (2015)
4. Borgatti, S. P.: 2-Mode concepts in social network analysis. Encyclopedia of complexity and system science, 6, 8279-8291 (2009)
5. Buxbaum, J. L.: Spotlight on HIEs and EHRs. How Outsourcing Fits in Health Management Technology, 32(5), 20-21 (2011)
6. Carr, L. P. Nanni, A. J.: Delivering Results: Managing What Matters, New York, Springer Science Business Media, LLC (2009)
7. Creswick, N., Westbrook, J. I.: Social network analysis of medication advice-seeking interactions among staff in an Australian hospital. International journal of medical informatics, 79(6), 116-125 (2010)
8. Cross, R., Parker, A.: Charged up: Creating energy in organizations. Journal of Organizational Excellence, 23(4), 3-14 (2004)
9. Freeman, L.: The development of social network analysis. A Study in the Sociology of Science, 1(687), 159-167 (2004)
10. Guimarães, C. M., de Carvalho, J. C.: Outsourcing in the healthcare sector-a state-of-the-art review. In Supply Chain Forum: An International Journal, Vol. 12, 140-148 (2011)
11. Hazelwood, S. E., Hazelwood, A. C. Cook, E. D.: Possibilities and Pitfalls of Outsourcing, Healthcare Financial Management, 59(10), 44-48 (2005)
12. Hoppe, B., Reinelt, C.: Social network analysis and the evaluation of leadership networks. The Leadership Quarterly, 21(4), 600-619 (2010)
13. Latapy, M., Magnien, C., Del Vecchio, N.: Basic notions for the analysis of large two-mode networks. Social networks, 30(1), 31-48 (2008)
14. O'malley, A. J., Marsden, P. V.: The analysis of social networks. Health services and outcomes research methodology, 8, 222-269 (2008)
15. Ozcan, Y.A.: Quantitative methods in health care management: techniques and applications. John Wiley & Sons (2005)
16. Roberts, J. G., Henderson, J. G., Olive, L. A., Obaka, D.: A review of outsourcing of services in health care organizations. Journal of Outsourcing and Organizational Information Management (2013)
17. Sunseri, R.: Outsourcing on the Outs Hospitals & Health Networks, 73(10), 46-2 (1999)
18. Wasserman, S., Faust, K.: Social network analysis: Methods and applications (1994)
19. Young, S.: Outsourcing in the Australian health sector: the interplay of economics and politics. International Journal of Public Sector Management, 18(1), 25-36 (2005)

A project evaluation study on multiset Likert scale data

Uno studio di valutazione su dati multiset in scala Likert

Violetta Simonacci, Marina Marino, Maria Gabriella Grassia and Michele Gallo

Abstract This work is part of the evaluation proposal for the experimental phase of the ClassMate Robot project, promoted by the Protom Group. The experimentation consists in testing how a newly developed AI device for social education is received in a classroom environment. To assess usability, likability, and social impact pre- and post-trial surveys were administered to the participating students of 4 schools. The data is arranged in multi-block architectures and then summarized with IRT tools. A classic non-parametric approach is employed for testing before and after differences. Post-experimentation results are explored via PARAFAC2 to model school differences while accounting for a multiset structure.
Abstract *Questo lavoro fa parte della proposta di valutazione per la fase sperimentale del progetto ClassMate Robot, promosso dal Protom group. La sperimentazione consiste nel testare come un nuovo dispositivo AI per l'istruzione viene percepito in un contesto scolastico. Per valutarne usabilità, likability e impatto sociale sono stati somministrati questionari pre e post agli studenti partecipanti di 4 scuole. I dati sono inseriti in architetture multiblocco e riassunti con strumenti IRT. Un approccio non parametrico viene utilizzato per testare le differenze prima-dopo. I risultati post-sperimentazione saranno esplorati via PARAFAC2 per modellare le differenze tra scuole tenendo conto della struttura mulitset.*

Key words: AI for education, impact assessment, IRT, PARAFAC2, survey data

Violetta Simonacci
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: violetta.simonacci@unina.it

Marina Marino
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: marina.marino@unina.it

Maria Gabriella Grassia
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: mgrassia@unina.it

Michele Gallo
Dept. of Human and Social Sciences, University "L'Orientale", Naples, Italy e-mail: mgallo@unior.it

1 Project description and introduction

Protom Group S.p.A. is the first Italian Knowledge & Technology-Intensive (KTI) company with a cutting-edge profile in the field of digital transformation. In 2021, they set in motion a pioneering business project based on technology for education, known as ClassMate Robot (CMR). The idea behind CMR is to use Artificial Intelligence (AI), by introducing an in-house built social robot archetype, to bring upon the conventional Italian school framework innovative teaching and learning processes.

The project carried out through Protom Robotics and Scuolab includes the collaboration with the Projects of Intelligent Robotics and Advanced Cognitive System (PRISCA) Lab of the University of Naples ”Federico II” for the development of the software infrastructure and the scientific support of the Department of Social Sciences (DiSS) of the University of Naples ”Federico II”. In detail, DiSS played an active role in defining software requirements, outlining the educational framework, and implementing an experimental phase in the real context of 4 Italian schools (Junior High and High School level). DiSS is also responsible for final reporting and for carrying out a full assessment study. The experimental phase will officially close at the end of the 2022-2023 school year.

Impact assessment includes a handful of qualitative and quantitative tools. Interviews and case studies are accompanied by the development of business and social intelligence paths. A detailed plan for data harvest during the experimentation was devised together with Protom Group, specifically imagining the potential use of collected information.The goal of the quantitative assessment is not only to measure the success of the project but also to provide useful tools for the improvement of the AI device and the joint cloud platform.

The plan includes two types of data collection tools: automatic detection via device and platform; and the administration of surveys to the entire cohort of students. Questionnaires are to be submitted at two time points to allow a comparison between the final perception and the initial expectation. In the specific context of this presentation, we focus on survey data only for brevity reasons.

In detail, two data-sets are considered for each participating school: 1) Survey at t_0, student responses to the initial questionnaire, nested by school, which include 28 items on a four-point Likert scale (grouped in 4 thematic blocks) and sociodemographic items; and 2) Survey at t_1, student responses to the final questionnaire, nested by school, encompassing 8 blocks of items with the same 28 questions of the previous survey and an additional 33 items all on a four-point Likert scale. A more detailed description of the data is provided in Section 2.

The aim of this work is to implement an initial exploratory study of survey data by keeping in mind two major data characteristics: the ordinal nature of Likert scale items and the presence of a nested design, as school grouping is likely to have an impact on student responses. The analytical goal is to answer several questions concerning:

1. Survey validity: can we obtain valid summary measures for each thematic block?

2. Pre- and post-experimentation differences: are there significant changes in student responses on common blocks before and after experimentation?
3. Measure interactions: how do the detected constructs interact with each other?
4. Nesting effects: What is the imprint of school grouping?
5. Role of External variables: Are there sociodemographic group differences?

The methodological design developed to address these queries is articulated in the following manner. To start, summary measures are built for all blocks using an Item Response Theory (IRT) approach in order to properly treat Likert scale items and assess survey validity. The impact of experimentation on the 4 blocks common to Survey at t_0 and Survey at t_1 is verified via Wilcoxon signed-rank tests. To study measure interaction in the nested samples at t_1, an exploratory perspective was preferred, using the PARAFAC2 model. Section 3 outlines the methodology in more detail. Section 4 presents the project outlook.

2 Survey Data

To properly assess CMR performance it was deemed necessary to quantify the likability of the device and its validity as a teaching tool through the administration of a post-experimentation survey (at t_1). Questions were specifically developed to measure: students' general perception of the CMR (8 items on usability and likability), students' comfort level using the CMR (9 items), students' perception of CMR impact on school results (5 items), students' perception of platform likability (6 items).

To gather more information on students and classroom environment a set of general questions were also added to measure school well-being, following [8]. These Likert-scale items are divided into 3 blocks: relationship with teachers (7 items on trust, support, recognition), relationship with classmates (6 items on acceptance, trust, friendships), and sense of self-efficacy (10 items). In addition, a collection of socio-biographical information (8 questions on gender, parental education and employment, living situation, and grade) and a block on the relationship with technology on a Likert scale (10 items) were added. It was decided to submit well-being questions and the relationship with the technology block also before the experimentation (at t_0) to test the CMR social impact. All Likert scale items are on a 4-point system where the options were "1 = NOT AT ALL", "2 = A LITTLE", "3 = ENOUGH", "4 = A LOT".

Four classes, located in four different schools in Italy, were selected for the trial (Rome, Carrù, Dalmine, and Verona). A total of 96 students participated in the project.

Collected data can be easily organized in nested data structures. The responses to Survey at t_0 can be arranged in a multi-level object $\mathscr{X}^{t_0}$ subdivided in school-by-item-block tables. Formally, for the k-th school and s-th item block, we have a generic table $\mathbf{X}_{ks}^{t_0}$ holding the scores of the I_k students of school k on the J_s items of the block s as follows:

$$\mathbf{X}_{kb}^{t_0} = \begin{bmatrix} x_{1_k 1_s} & \cdots & x_{1_k j_s} & \cdots & x_{1_k J_s} \\ \vdots & \ddots & & & \vdots \\ x_{i_k 1_s} & & x_{i_k j_s} & & x_{i_k J_s} \\ \vdots & & & \ddots & \vdots \\ x_{I_k 1_s} & \cdots & x_{I_k j_s} & \cdots & x_{I_k J_s} \end{bmatrix} \tag{1}$$

where $k = 1, \ldots, 4$ and $s = 1, \ldots, 4$ for a total of 16 tables. If the tables a juxtaposed we have a 96 rows by 28 columns object:

$$\mathscr{X}^{t_0} = \begin{bmatrix} \mathbf{X}_{11}^{t_0} & \cdots & \mathbf{X}_{1s}^{t_0} & \cdots & \mathbf{X}_{1S}^{t_0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{X}_{k1}^{t_0} & & \mathbf{X}_{ks}^{t_0} & & \mathbf{X}_{kS}^{t_0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{X}_{K1}^{t_0} & \cdots & \mathbf{X}_{Ks}^{t_0} & \cdots & \mathbf{X}_{KS}^{t_0} \end{bmatrix} \tag{2}$$

Similarly, for Survey at t_1, a multilevel object $\mathscr{X}^{t_1}$ can be built. The object includes 32 school-by-item-block tables, as in the matrices 1 and 2, where the only difference is that $s = 1, \ldots, 8$.

An object $\mathscr{G}$, nested by school can also be built for student socio-demographic variables, in which student information is collected for the 8 described items.

3 Methodology

The methodological flow of this work can be summarized in the following phases:

STEP 1: Each item block in $\mathscr{X}^{t_0}$ and $\mathscr{X}^{t_1}$ is tested for consistency and reliability throughout schools using Cronbach's alpha and the Automated Item Selection Procedure (AISP) [6]. Problematic items may be considered as separate measures or sub-blocks may be formed if diagnostics suggest modification.

STEP 2: For each consistent block a measurement scale is identified throughout schools which represents a one-dimensional latent trait. To attain an interval scale, an IRT approach is used, which is based on the probabilistic relation between item difficulty and subject ability. In detail, the Partial Credit Rasch Model (PCM) [10] will be employed. The model can be described as follows:

$$\texttt{log}(P_{ilc}/P_{il(c-1)}) = U_i - V_l - F_{lc} \tag{3}$$

Here the probability P_{ilc} for the i-th subject of responding in the category c rather than in the category $(c-1)$ in reference to the l-th item is a function of the subject ability U_i, the item difficulty V_l and the rating scale structure F_{lc}. This model yields simplified versions of $\mathscr{X}^{t_0}$ and $\mathscr{X}^{t_1}$ where item blocks are replaced by summary measures. We obtain reduced-size objects, only nested by school, which can be defined as the multiset tensors $\underline{\mathbf{T}}^{t_0}(I_k \times N \times K)$ and $\underline{\mathbf{T}}^{t_1}(I_k \times M \times K)$, with generic element $t_{i_k nk}$ and $t_{i_k mk}$ respectively. They represent a collection of the K tables $\mathbf{T}_k^{t_0}(I_k \times N)$ and $\mathbf{T}_k^{t_1}(I_k \times M)$, where $n = 1, \ldots, N$ and $n_1 = 1, \ldots, M$ indicate the set of summary measured obtained at t_0 and t_1.

STEP 3: Paired Wilcoxon signed-rank tests are performed on all measures to assess if significant differences were recorded at t_1 on the well-being and technology relationship items. A non-parametric test was preferred due to the sample size.

STEP 4: To study measure interaction, $\underline{\mathbf{T}}^{t_1}$ is decomposed via PARAFAC2. The PARAFAC2 model [5, 7] can be described as a less restrictive version of standard PARAFAC which can also be applied when the data tensor is not fully-crossed (presenting a dimension of varying size).

Ordinary PARAFAC [1, 4] decomposes a fully-crossed tensor $\underline{\mathbf{T}}(I \times M \times K)$ based on the parallel proportional profiles principle [2] and the assumption of complete trilinearity. In detail, it assumes that the obtained latent variables correspond to real constructs which hold proportional patterns throughout levels. As a result, it yields only three loadings matrices $\mathbf{A}(I \times R)$, $\mathbf{B}(M \times R)$ and $\mathbf{C}(K \times R)$, where R is the number of extracted factors, one for each dimension of the tensor, in the following manner:

$$\mathbf{T}_k = \hat{\mathbf{T}}_k + \mathbf{E}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1, \cdots, K. \tag{4}$$

Here $\mathbf{T}_k$ is the generic frontal slices of $\underline{\mathbf{T}}$, i.e. an $(I \times M)$ matrix for a given occasion k. $\mathbf{D}_k$ is a diagonal matrices holding the kth row of the third-mode loading matrix $\mathbf{C}$ and lastly $\mathbf{E}_k$ is the frontal slice of the residual tensor $\underline{\mathbf{E}}$. The model is unique under mild conditions.

PARAFAC2 relaxes the assumption of trilinearity by allowing for different loading matrices across levels in one of the three dimensions (conventionally the first dimension). This is particularly useful in the case of multiset data, where there are incomparable observation units across samples. The model can thus be adjusted as follows:

$$\mathbf{T}_k = \hat{\mathbf{T}}_k + \mathbf{E}_k = \mathbf{A}_k\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1, \cdots, K \tag{5}$$

The main difference is that the model generates K loading matrices $\mathbf{A}_k$. To ensure the uniqueness advantage also to the PARAFAC2 model, a restriction is imposed that the loading matrices $\mathbf{A}_k$ only differ in terms of rotation, i.e. the cross-product (covariance or correlation) matrix ${\mathbf{A}_k}^t\mathbf{A}_k$ is constant over k.

STEP 5: PARAFAC2 results are visually studied also for assessing the behavior of different socio-demographic groups with special attention to gender differences and parental education level.

4 Project outlook

The experimental phase officially terminates in May 2023. Nonetheless, most surveys have already been collected. A first, though incomplete, analysis has been implemented which demonstrates the feasibility and effectiveness of the planned methodology described in this paper. Complete results will be available for discussion during the presentation.

Some methodological advancements are also being considered. The advantage of using a multilevel approach for the extraction of summary measures, following [3], rather than standard PCM, will be explored. Alternative methods to build summary measures will also be implemented and analyzed in a comparative fashion, such as the approach proposed in [9], where a multivariate compositional analysis is carried out to extract bipolar constructs known as log-contrasts. Lastly, in the second stage of the evaluation process, Data Analytics will be studied in detail to also evaluate technical performance and its impact on likability and overall student experience.

References

1. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35(3):283–319 (1970)
2. Cattell, R.B.: "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 9(4):267-283 (1944)
3. Doran, H., Bates, D., Bliese, P., Dowling, M.: Estimating the multilevel Rasch model: With the lme4 package. Journal of Statistical software 20:1-18 (2007)
4. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. UCLA Working Papers in Phonetics 16:1–84 (1970)
5. Harshman, R.A.: PARAFAC2: Mathematical and technical notes. UCLA Working Papers in Phonetics, 22:30-44 (1972)
6. Hemker, B.T., Sijtsma, K., Molenaar, I.W.: Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken I RT model. Applied Psychological Measurement 19(4):337-352 (1995)
7. Kiers, H.A., Ten Berge, J.M., Bro, R.: PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics: A Journal of the Chemometrics Society 13(3-4):275-294 (1999)
8. Marzocchi, G.M., Tobia, V.: QBS 8-13: Questionari per la valutazione del benessere scolastico e identificazione dei fattori di rischio. Edizioni Centro Studi Erickson (2015)
9. Simonacci, V., Gallo, M.: Statistical tools for student evaluation of academic educational quality. Quality & Quantity 51: 565-579 (2017)
10. Wright, B.D., Masters, G.N.: Rating scale analysis. MESA press (1982)

Solicited Session SS14 - *Distance and depth-based statistical learning methods for robust data analysis*

Organizer and Chair: Silvia Salini

1. *Robust distance-based predictive models* (Boj E., Grané A. and Parron D.)
2. *Data depth for mixed-type data through multidimensional scaling. An application to biological age imputation* (Cascos I., Grané A. and Qian J.)
3. *A compared protocol to improve clustering procedures* (Grané A., Riani M. and Salini S.)
4. *Robust diagnostics for Linear Mixed Models with the Forward Search* (Corbellini A., Grossi L. and Laurini F.)

Robust distance-based predictive models

Robusti modelli predittivi basati sulla distanza

Eva Boj, Aurea Grané and David Parron

Abstract In this work a robust version of Gower's distance is proposed to be used in the predictors' space of distance-based predictive models. The performance of the new proposal is compared to that of classical Gower's metric in the presence of outliers in data sets of multivariate heterogeneous data. Mean squared error, as well as other goodness of fit measures, are used to evaluate the effectiveness in the prediction of responses. Computations on real data sets are made using the dbstats package for R.

Abstract *In questo lavoro viene proposta una versione robusta della distanza di Gower da utilizzare nello spazio dei predittori dei modelli predittivi basati sulla distanza. Le prestazioni della nuova proposta vengono confrontate con quelle della metrica di Gower classica in presenza di valori anomali in insiemi di dati eterogenei multivariati. L'errore quadratico medio, così come altre misure di bontà di adattamento, vengono utilizzati per valutare l'efficacia nella previsione delle risposte. I calcoli su insiemi di dati reali vengono eseguiti utilizzando il pacchetto dbstats per R.*

Key words: dblm, dbstats, mixed-type data, outliers, robust Gower's distance, R

1 Models and distances under evaluation

Distance-based regression (DBR) [5, 6, 1, 2] is a prediction tool which can be applied to qualitative or mixed explanatory variables while keeping compatibility with

Eva Boj
Universitat de Barcelona, e-mail: evaboj@ub.edu

Aurea Grané
Department of Statistics, Universidad Carlos III de Madrid, e-mail: aurea.grane@uc3m.es

David Parron
Department of Statistics, Universidad Carlos III de Madrid, e-mail: 100409463@alumnos.uc3m.es

ordinary regression by least squares (LS), which appears as a particular case. The model projects the vector of continuous responses onto a Euclidean space obtained by metric multidimensional scaling (MDS) (see, e.g., [4]) from the observed predictors, which are nonlinearly mapped into a set of latent, i.e., non-observed, dimensions in this space.

1.1 Model description

A continuous response Y is to be predicted from a set of p predictors, $\mathbf{w}_1$, $\mathbf{w}_2$, ..., $\mathbf{w}_p$, possibly a mixture of quantitative and qualitative variables. An n-vector, $\mathbf{y}$, contains the values of Y for an n-set Ω of cases or units. Let $\delta : \Omega \times \Omega \to \mathbb{R}^+$ be a distance function acting on the $\mathbf{w}_j$-coordinates, $\delta_{ij} = \delta(\mathbf{w}_i, \mathbf{w}_j)$, and consider the $n \times n$ predictor distance matrix $\Delta = \left(\delta_{ij}^2\right)$. Matrix Δ is called Euclidean if, for some integer r, we can find $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^r$, such that $(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) = \delta_{ij}^2$. The $n \times r$ matrix X formed by stacking the n rows $\mathbf{x}_i^T$ verifies that the Gram matrix $G = XX^T$ is positive semi-definite, with rank $r = rank(G) \leq n-1$ and X is called the Euclidean configuration of Δ.

The DBR of $\mathbf{y}$ on Δ is defined as an LS regression with response $\mathbf{y}$ and matrix of predictors X, where X is a Euclidean configuration of Δ. The adjusted $\hat{\mathbf{y}}$ assuming that $\mathbf{y}$ is centered is given by:

$$\hat{\mathbf{y}} = X\hat{\beta}, \text{ where } \hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}.$$

The prediction for a new unit $\{n+1\}$ is given by:

$$\hat{\mathbf{y}}_{n+1} = \hat{\mathbf{x}}_{n+1}\hat{\beta}, \text{ where } \hat{\mathbf{x}}_{n+1} = \frac{1}{2}(\mathbf{g}-\mathbf{d})X\left(X^T X\right)^{-1}$$

is computed by Gower's interpolation formula [9, 8], and the row vector $\mathbf{d}$ contains the n squared distances from unit $\{n+1\}$ to the other units in Ω, $\mathbf{g}$ is the row vector containing the main diagonal entries in G and G^+ is the Moore-Penrose g-inverse of G. Thus,

$$\hat{\mathbf{y}}_{n+1} = \frac{1}{2}(\mathbf{g}-\mathbf{d})X\left(X^T X\right)^{-2} X^T \mathbf{y} = \frac{1}{2}(\mathbf{g}-\mathbf{d})\,G^+ \mathbf{y}.$$

1.2 Predictor distance matrices

A well-know distance for mixed type data, that is, a mixture of quantitative, qualitative, and binary variables, is Gower's distance, which was defined in [10] as $\delta_{ij}^2 = 1 - s_{ij}$, where

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - \left|w_{ih} - w_{jh}\right| / G_h\right) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

with p_1 the number of continuous variables, G_h the range of the h-th continuous variable, a and d the number of positive and negative matches, respectively, for the p_2 binary variables, and α the number of matches for the p_3 multi-state categorical variables.

Gower's distance can be defined as the Pythagorean sum of three distance measures for quantitative, binary and multi-state categorical variables, $\Delta_1 = \left(\delta_1^2(\mathbf{w}_i, \mathbf{w}_j)\right)$, $\Delta_2 = \left(\delta_2^2(\mathbf{w}_i, \mathbf{w}_j)\right)$, $\Delta_3 = \left(\delta_3^2(\mathbf{w}_i, \mathbf{w}_j)\right)$ where δ_1 is the range-normalized city-block distance, δ_2 distance is associated to Jaccard's similarity coefficient and δ_3 is the Hamming distance. However, this classical distance presents two main drawbacks: It does not take into account the correlations between quantitative variables and it is a not robust metric. In [7] a robustification of Gower's distance is proposed in the context of MDS, by taking δ_1 as a robust Mahalanobis distance, instead of the range-normalized city-block one, and δ_2 and δ_3 are left unchanged. A similar metric is used in this paper in the context of distance-based prediction, where function *cov-Rob* of the *robust* package [11] for R [12] is used to estimate the covariance matrix when computing Mahalanobis distance for quantitative variables.

2 Simulation scheme and preliminary results

The performance of robust Gower distance is evaluated and compared to that of classical Gower's in several scenarios with a given percentage of outlier contamination. In particular, two datasets of mixed-type predictors were generated with p_1 quantitative variables, p_2 binary variables and p_3 multi-state categorical ones and response variable was obtained as a random linear combination of predictors. Sample size was set to $n = 300$. Next, outliers were introduced by changing several characteristics in the explanatory variables of existing units, as follows:

1. Lowly correlated/associated mixed-type predictors with $p_1 = p_2 = p_3 = 2$.
   a) A 2% outlier contamination in the first quantitative variable.
   b) A 2% outlier contamination in the second quantitative variable.
   c) A 5% outlier contamination in all binary variables.
   d) A 5% outlier contamination in all multi-sate categorical variables.
2. Highly correlated/associated mixed-type predictors with $p_1 = p_2 = p_3 = 2$.
   a) A 2% outlier contamination in the first quantitative variable.
   b) A 2% outlier contamination in the second quantitative variable.
   c) A 5% outlier contamination in all binary variables.
   d) A 5% outlier contamination in all multi-sate categorical variables.

Distance-based predictive models under evaluation are compared in terms of mean square error (MSE) computed by leave-one-out. Models were fitted using the *dblm* function of the *dbstats* package for R [3]. Some preliminary results are shown in Table1, where it can be observed that, in general, DBR with robust Gower outperforms the classical one.

Table 1 Performance of the models: Leave-one-out estimated MSE

Dataset	Gower	Robust Gower	Dataset	Gower	Robust Gower
1	2.5489	1.0162	2	6.5305	4.4573
1 a)	2.4961	2.3559	2 a)	6.3810	5.6671
1 b)	2.5151	2.5823	2 b)	6.5656	4.4175
1 c)	2.3874	1.0341	2 c)	6.4374	4.3974
1 d)	2.3029	1.0374	2 d)	6.0043	4.4296

Acknowledgements The authors acknowledge the support of the Spanish Ministry of Science and Innovation by grants PID2021-123592OB-I00 and TED2021-129316B-I00.

References

1. Boj, E., Claramunt, M. M., Fortiana, J.: Selection of predictors in distance-based regression. Commun. Stat. Theory Methods 36, 87-98 (2007)
2. Boj, E., Delicado, P., Fortiana, J.: Local linear functional regression based on weighted distance-based regression. Comput. Stat. Data Anal. 54, 429-437 (2010)
3. Boj, E., Caballé, A., Delicado, P., Fortiana, J.: dbstats: Distance-Based Statistics (dbstats). R package version 2.0.1 (2022)
http://CRAN.R-project.org/package=dbstats
4. Borg, I., Groenen, P.: Modern multidimensional scaling: theory and applications. Springer, New York (1997)
5. Cuadras, C. M., Arenas, C.: A distance-based model for prediction with mixed data. Commun. Stat. Theory Methods 19, 2261-2279 (1990)
6. Cuadras, C. M., Arenas, C. and Fortiana, J.: Some computational aspects of a distance-based model for prediction. Commun. Stat. Simul. Comput. 25(3), 593-609 (1996)
7. Grané, A., Manzi, G., Salini, S.: Smart Visualization of Mixed Data. Stats, 4, 472-485 (2021)
8. Gower, J. C., Hand, D. J.: Biplots. Chapman and Hall, London (1996)
9. Gower, J. C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338 (1966)
10. Gower, J. C.: A general coefficient of similarity and some of its properties. Biometrika 27, 857-874 (1971)
11. Wang, J, Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., Konis, K.: robust: robust: Port of the S+ ”Robust Library”. R package version 0.7-1 (2022)
http://CRAN.R-project.org/package=robust
12. R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria (2023)
http://www.R-project.org/

Data depth for mixed-type data through multidimensional scaling. An application to biological age imputation

Profondità dei dati per dati misti attraverso la scalatura multidimensionale. Un'applicazione all'imputazione dell'età biologica

Ignacio Cascos, Aurea Grané, and Jingye Qian

Abstract For a mixed-type dataset, we propose a new procedure to assess the quality of an observation as a central tendency of the dataset. Next, we apply this technique to valuate the functional condition of the body of a human being in terms of its biological age, which is based on biomarkers, medical conditions, life habits, and sociodemographic variables. These records are of mixed-type since they are made up by numerical and categorical variables. In order to evaluate the centrality of an observation in a mixed-type dataset, we obtain a Multidimensional Scaling representation and use some classical notion of multivariate data depth in the corresponding space. The biological age of an individual is finally assessed in terms of the age of the deepest possible artificial individual sharing all other records with the given one.

Abstract *Per un dataset di tipo misto, proponiamo una nuova procedura per valutare la qualità di un'osservazione come tendenza centrale del dataset. Successivamente, applichiamo questa tecnica per valutare la condizione funzionale del corpo umano in termini di età biologica, basata sulla condizione medica, sui biomarcatori, sulle abitudini di vita e sulle variabili sociodemografiche. Questi dati sono di tipo misto poiché sono composti da variabili numeriche e categoriche. Per valutare la centralità di un'osservazione in un dataset misto, otteniamo una rappresentazione della Scalatura Multidimensionale e utilizziamo alcune nozioni classiche di profondità dei dati multivariati nello spazio corrispondente. L'età biologica di un individuo viene infine valutata in termini di età dell'individuo artificiale più profondo possibile che condivide tutti gli altri dati con quello fornito.*

Key words: Biological Age, Data Depth, Imputation, Mixed-type Data

---

Ignacio Cascos
Department of Statistics, Universidad Carlos III de Madrid, e-mail: ignacio.cascos@uc3m.es

Aurea Grané
Department of Statistics, Universidad Carlos III de Madrid, e-mail: aurea.grane@uc3m.es

Jingye Qian
Department of Statistics, Universidad Carlos III de Madrid, e-mail: 100384773@alumnos.uc3m.es

1 Introduction

The *biological age* serves as a gauge of the functional condition of a human's body. It is measured in the same time unit of the chronological age (commonly years), and an individual whose biological age is x years is a person whose body — from a functional viewpoint — and life habits resemble the ones of a standard person who is x years old (chronological age). See [10, 13] for some reviews on the sheer concept of biological age and its predictors. There is a wide variety of techniques to valuate the biological age, ranging from multiple linear regression, see [19], to deep learning, see [16], while an extense overview of them can be found in [9]. Similarly to [3], we use a variable imputation procedure based on data depth, see also [15]

In Section 2 we introduce the concept of data depth, which is extended to mixed-type data in Section 3. Finally, in Section 4 we briefly describe the imputation procedure for the biological age.

2 Data depth

In order to assess the biological age of an individual, we use information of numerical and categorial variables (mainly binary ones), and the main innovation in the current proposal is to introduce notions of data depth for mixed-type data based on a Muldimensional Scaling (MDS) representation of the original dataset, see [2] for and overview of MDS and [7, 8] for specific MDS techniques for mixed-type data. Despite of the large number of depth proposals for data in non-Euclidean spaces, mainly but not only infinite dimensional ones (see [12] for functional data, [5] for set data, [17] for fuzzy data, or [1] for text data), we are not aware of proposals handling mixed-type data that comprises of numerical and categorical variables.

In multivariate statistics, the term *data depth* refers to the degree of centrality of a point with respect to a data cloud. The more central points assume higher depth values, while depth decreases as we move away from the centre of the data cloud. For general introductions to the concept of multivariate depth and some basic applications see [4, 11, 20]. Some classical instances of multivariate depths are the Mahalanobis depth, the L_2 depth, the halfspace depth, or the expectile depth. If $x \in \mathbb{R}^d$ is a d-dimensional vector and X is a d-dimensional random vector, when evaluating the depth of x with respect to X, the former four depth notions are defined as:

- Mahalanobis depth: $\mathrm{MhD}(x;X) = \big(1 + (x - \mathbb{E}X)^\top \Sigma_X^{-1}(x - \mathbb{E}X)\big)^{-1}$, where $\mathbb{E}X$ stands for the expectation of X and Σ_X for its covariance matrix.
- L_2 depth: $\mathrm{L_2D}(x;X) = (1 + \mathbb{E}\|x - X\|)^{-1}$, where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$, see [14].
- Halfspace depth: $\mathrm{HD}(x;X) = \inf\{\Pr(X \in H) : x \in H \text{ halfspace}\}$, see [18].
- Expectile depth: $\mathrm{ED}(x;X) = \left(2 - \inf_{u \in \mathbb{S}^{d-1}} \frac{\langle \mathbb{E}X - x, u\rangle}{\mathbb{E}\langle X - x, u\rangle_+}\right)^{-1}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, $\langle\cdot,\cdot\rangle$ the inner product, and for any $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$, see [6].

The previous notions of data depth are independent of the system of coordinates (except the L_2 depth), assume their maximum value at the *centre* of the distribution of X (a multivariate median or the mean), from which they decrease radially, and vanish (converge to zero) when x tends to infity. In fact, the halfspace and expectile depths are zero out of the convex hull of the support of X, but there are corrections that can be applied to them so that they never vanish.

3 Data depth for mixed data

In order to introduce a notion of depth for mixed-type data, we obtain an MDS representation of the original dataset in a Euclidean space of some appropriate dimensionality and evaluate a classical depth there. The assessment of the depth of a new observation with respect to the existing dataset requires its projection (interpolation) on the established Euclidean space. When considering mixed-type observations $z_1,\ldots,z_n$ on p numerical and/or categorical variables, proceed as follows:

1. Take a notion of distance that can handle mixed-type data (specifically Gower's distance or a robustification).
2. Obtain the MDS representation of the original variables, represented as $x_1,\ldots,x_n$ lying on the n-dimensional Euclidean space, and keep a number d of principal components $y_i = \mathrm{proj}_d(x_i)$, where proj_d is the projection on those d components.
3. The depth of z_i with respect to the original dataset is assessed in terms of the one of y_i with respect to the dataset in the d-dimensional space, $y_1,\ldots,y_n$.
4. The depth of any other p-variate element z with the same variables as the orginal observations is obtained after projecting z on the n-dimensional Euclidean space using Gower's interpolation formula, keeping the corresponding d components, and evaluating the depth with respect to the dataset in the d-dimensional space.

4 Biological age imputation

We have a dataset with 107 variables on 911 individuals, including chronological age, sex, 3 biometric numerical variables (Body Mass Index and Diastolic/Systolic Blood Pressures), 12 binary variables about medical conditions (asthma, cancer, COPD, Diabetes Mellitus,...), approx. 20 variables about eating and life habits (some are binary, others categorical with more categories), and data on 6 genes.

Take one individual from the sample, its biological age is assessed as follows:

1. Select all other individuals whose chronological age is within ± 15 years the one of the reference individual.
2. With the selected individuals, build a balanced sample by resampling (with replacement) uniformly on the chronological age. That sample has approximately the same number of individuals with each chronological age.

3. For the balanced sample, impute the biological age of the given individual as the chronological age of the deepest possible synthetic individual with the remaining variables identical to its.

Acknowledgements The authors acknowledge the support of the Spanish Ministry of Science and Innovation by grants PID2021-123592OB-I00 and TED2021-129316B-I00.

References

1. Bolívar, S., Nieto-Reyes, A.; Rogers, H.L.: Statistical Depth for Text Data: An Application to the Classification of Healthcare Data. Mathematics 11, 228 (2023)
2. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling. Springer, Boston (2005)
3. Cabras, S., Cascos, I., D'Auria, B., Durbán, M., Guerrero, V., Ochoa, M.: Biological Age Imputation by Data Depth. In: García-Escudero, L.A. *et al.* (eds.) Building Bridges between Soft and Statistical Methodologies for Data Science. Adv. Intell. Syst. Comput. 1433, 57-74. Springer, Cham (2023)
4. Cascos, I.: Data Depth: Multivariate Statistics and Geometry. In: Kendall, W.S., Molchanov, I, (eds.) New Perspectives in Stochastic Geometry, pp. 398-423. OUP, Oxford (2009)
5. Cascos, I., Molchanov, I., Li, Q.: Depth and outliers for samples of sets and random sets distributions. Aust N. Z. J. Stat. 63, 55-82 (2021)
6. Cascos, I., Ochoa, M.: Expectile depth: Theory and computation for bivariate datasets. J. Multivariate Anal. 184, 104757 (2021)
7. Grané, A., Manzi, G., Salini, S.: Smart Visualization of Mixed Data. Stats 4, 472-485 (2021)
8. Grané, A., Salini, S., Verdolini, E.: Robust multivariate analysis for mixed-type data: Novel algorithm and its practical application in socio-economic research. Socio-Econ. Plan. Sci. 73, 100907 (2021)
9. Jia, L., Zhang, J.L., Chen, X.: Common methods of biological age estimation. Clin. Interv. Aging. 12, 759-772 (2017)
10. Jylhävä, J., Pedersen, N.L., Hägg, S.: Biological Age Predictors. EBioMedicine 21, 29-36 (2017)
11. Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann. Stat. 27, 783-858 (1999)
12. López-Pintado, S., Romo, J.: On the Concept of Depth for Functional Data. J. Am. Stat. Assoc. 104, 718-734 (2009)
13. Ludwing, F.C., Smoke, M.E.: The measurement of biological age. Exp. Aging Res. 6, 497-522 (1980)
14. Mosler, K.: Depth statistics. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather, pp. 17-34. Springer (2013)
15. Mozharovskyi, P., Josse, J., Husson, F. Nonparametric Imputation by Data Depth. J. Am. Stat. Assoc. 115, 241-253 (2020)
16. Rahman, S.A., Giacobbi, P., Pyles, L., Mullett, C., Doretto, G., Adjeroh, D.A.: Deep learning for biological age estimation, Brief. Bioinformatics 22, 1767-1781 (2021)
17. Sinova, B.: On depth-based fuzzy trimmed means and a notion of depth specifically defined for fuzzy numbers, Fuzzy Sets Syst. 443, 87-105 (2022)
18. Tukey, J.W.: Mathematics and the Picturing of Data. In: James, R.D. (ed.) Proceedings of the International Congress of Mathematicians, Vancouver, 1975, pp. 523-531 (1975)
19. Voitenko, V.P., Tokar, A.V.: The assessment of biological age and sex differences of human aging. Exp. Aging Res. 9, 239-244 (1983)
20. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Stat. 28, 461-482 (2000)

A compared protocol to improve clustering procedures

Un protocollo comparato per migliorare le procedure di clustering

Aurea Grané, Marco Riani and Silvia Salini

Abstract In this paper we study two widely used Machine Learning dimensionality reduction techniques, such as t-SNE and UMAP, in the presence of outliers and/or inliers, with the purpose to understand whether and how they can be used to improve well-known statistical clustering procedures, such as k-means or t-clust.
Abstract *In questo articolo studiamo due tecniche di riduzione della dimensionalità ampliamente utilizzate in Machine Learning, come t-SNE e UMAP, in presenza di outlier e/o inlier, con lo scopo di capire se e come possono essere utilizzate per migliorare note procedure statistiche di clustering, come k-means o t-clust.*

Key words: tSNE, tclust, UMAP, robust clustering

1 Methods under study

t-Distributed Stochastic Neighbor Embedding (***t*-SNE**) was introduced by [3] as an improvement of Stochastic Neighbor Embedding ([2]) in the context of large datasets. It is a dimensionality reduction technique that emphasizes the preservation of local and global structure in high-dimensional data. It employs a probabilistic approach to model pairwise similarities between data points and minimizes the Kullback-Leibler (KL) divergence between high-dimensional and low-dimensional similarities.

Aurea Grané
Department of Statistics, University Carlos III of Madrid, e-mail: aurea.grane@uc3m.es

Marco Riani
Department of Economics and Management, University of Parma, e-mail: marco.riani@unipr.it

Silvia Salini
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: silvia.salini@unimi.it

Uniform Manifold Approximation and Projection (**UMAP**), proposed by [4], is a dimensionality reduction technique that aims to preserve both local and global structure in high-dimensional data [5]. It constructs a weighted graph representation of the data and optimizes the embedding to reflect the pairwise distances in the original space. UMAP introduces attractive and repulsive forces to capture the manifold structure of the data.

Both *t*-SNE and UMAP are popular dimensionality reduction techniques used for visualizing high-dimensional data. While they share the goal of mapping data to a lower-dimensional space, there are some key differences between t-SNE and UMAP. Here's a comparison of their main characteristics:

1. Algorithmic Approach:
   - *t*-SNE: constructs a probabilistic model that defines a similarity measure between pairs of data points, both in the original high-dimensional space and the low-dimensional embedding. It focuses on preserving the local structure and aims to map nearby points in the original space to nearby points in the embedding.
   - UMAP constructs a weighted graph representation of the data and optimizes the embedding to preserve both local and global structure. It introduces attractive and repulsive forces to capture the pairwise distances between connected points, aiming to preserve the underlying manifold structure of the data.
2. Scalability:
   - *t*-SNE: has scalability limitations, particularly for large datasets. It involves computing pairwise similarities between all data points, resulting in a time complexity of $O(N^2)$, where N is the number of data points. This makes it computationally expensive for very large datasets.
   - UMAP introduces the concept of a navigational grid, which allows for efficient computation of the embedding. It scales better than *t*-SNE and can handle larger datasets without sacrificing much performance.
3. Preservation of Global Structure:
   - *t*-SNE: tends to preserve the local structure of the data but may not accurately preserve the global structure. It can sometimes create artificial clusters or distort the overall distribution of the data.
   - UMAP aims to preserve both local and global structure by balancing attractive and repulsive forces. It often provides a better representation of the global structure of the data and is less prone to creating artificial clusters.
4. Parameter Sensitivity:
   - *t*-SNE: has several parameters that need to be carefully tuned, such as the perplexity parameter that controls the balance between local and global aspects of the data. The choice of perplexity can significantly impact the resulting visualization.

- UMAP has fewer sensitive parameters and is more robust to different parameter settings. The main parameter in UMAP is the ”n_neighbors” parameter, which determines the number of neighbors used to construct the graph.

5. Interpretability:

- *t*-SNE: is often used for exploratory data analysis and visualization. However, interpreting the relative distances between points in *t*-SNE space is challenging due to the non-linear nature of the algorithm.
- UMAP provides a more interpretable embedding as it aims to preserve the manifold structure. The distances between points in the UMAP space more accurately reflect the pairwise distances in the original data, making it easier to interpret and reason about the data.

In summary, t-SNE and UMAP are both valuable dimensionality reduction techniques with different strengths. *t*-SNE is often suitable for visualizing local structure and revealing clusters, while UMAP focuses on preserving the global structure and providing a more interpretable embedding. The choice between the two methods depends on the specific characteristics of the dataset and the visualization goals.

Several studies [6, 11, 7, 8, 9, 10] have compared *t*-SNE and UMAP to assess their performance and characteristics.

In the final paper, we will discuss which method works best in the presence of anomalous data and which is best suited as pre-processing when the ultimate goal of the analysis is clustering.

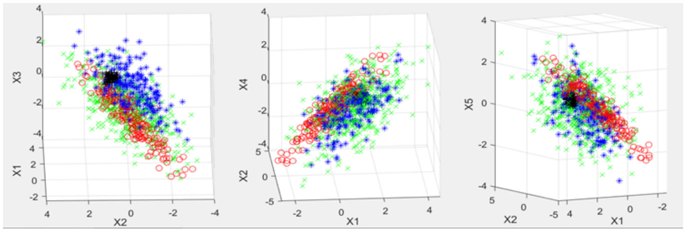
2 Simulation scenarios and preliminary results

To evaluate the procedure we generate several scenarios from multivariate normal distributions with a given percentage of pointwise outlier or inlier contamination. Figure1 contains some of them.

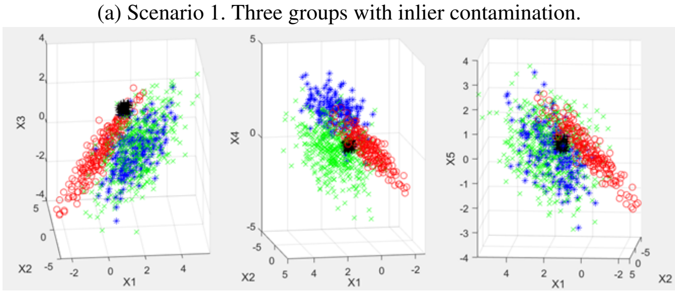
1. $n = 1000$ generated units from three unbalanced p-multivariate normal distributions accounting for 90% of the data (the bulk of data), a 10% amount of a rather disperse pointwise contamination placed at point (1,0.5,1.5,0,0), which is located inside the ellipsoides (inliers). In this case, $k = 4$.
2. $n = 1000$ generated units from three unbalanced p-multivariate normal distributions accounting for 90% of the data (the bulk of data), a 10% amount of a pointwise contamination placed at point (2.5,0,2,0,2), which is located outside the ellipsoides (outliers). In this case, $k = 4$.
3. $n = 1000$ generated units from three unbalanced p-multivariate normal distributions accounting for 90% of the data (the bulk of data), a 10% amount of a rather disperse pointwise contamination placed at point (2.5,0,2,0,2), which is located outside the ellipsoides (outliers). In this case, $k = 4$.

For each scenario we compare the following procedures to find the clusters.

Fig. 1 Scenarios of simulated data



(a) Scenario 1. Three groups with inlier contamination.



(b) Scenario 3. Three groups with outlier contamination.

- **Case A:** The number of clusters k and the amount of contamination α are known. We apply `tclust` on simulated data and on t-sne and UMAP results with different metrics.
- **Case B:** The number of clusters k and the amount of contamination α are unknown. We estimate the number of clusters with `tclustIC` either on simulated data or on t-sne and UMAP results. Next, we apply `tclust` with $\alpha = 0.01, 0.33$ on simulated data and on t-sne and UMAP results with different metrics.

Adjusted Rand index (ARI), Rand index (RI), Mirkin's index (MI) and Hubert index (HI) are used to evaluate the goodness of the partitions. Some preliminary results concerning only t-SNE are shown in Table 1–Table 2.

In the final paper we will consider other scenarios, show the results for UMAP as well, and discuss the various cases. In addition, it is our intention to apply but proposed methodology to a real-world example as well, choosing a data set in which classical robust clustering techniques struggle to identify the final groups

Acknowledgements

The authors acknowledge the support of the Spanish Ministry of Science and Innovation by grants PID2021-123592OB-I00 and TED2021-129316B-I00.

Table 1 Results for case A: k and α are known.

Scenario 1: tclust ($\alpha = 0.10$)	ARI	RI	MI	HI
Data	0.2110	0.6584	0.3416	0.3169
t-sne Euclidean	0.1633	0.6534	0.3466	0.3069
t-sne Mahalanobis	0.4025	0.7432	0.2568	0.4864
t-sne robust Mahal., bdp=0.1	0.4654	0.7665	0.2335	0.5331
t-sne robust Mahal., bdp=0.3	0.4388	0.7623	0.2377	0.5245
Scenario 2: tclust ($\alpha = 0.10$)				
Data	0.3727	0.7145	0.2855	0.4289
t-sne Euclidean	0.4092	0.7271	0.2729	0.4541
t-sne Mahalanobis	0.6954	0.8709	0.1291	0.7418
t-sne robust Mahal., bdp=0.1	0.7003	0.8729	0.1271	0.7458
t-sne robust Mahal., bdp=0.3	0.7083	0.8746	0.1254	0.7492
Scenario 3: tclust ($\alpha = 0.10$)				
Data	0.4632	0.7496	0.2504	0.4991
t-sne Euclidean	0.3891	0.7257	0.2743	0.4514
t-sne Mahalanobis	0.6410	0.8485	0.1515	0.6970
t-sne robust Mahal., bdp=0.1	0.6741	0.8624	0.1376	0.7247
t-sne robust Mahal., bdp=0.3	0.6856	0.8671	0.1329	0.7342

Table 2 Results for case B: k and α are unknown, we use tclustIC to estimate k and we fix $\alpha = 0.01$.

Scenario 1: tclust ($\alpha = 0.01$)	k	ARI	RI	MI	HI
Data	3	0.4338	0.7337	0.2663	0.4674
t-sne Euclidean	5	0.1471	0.6523	0.3477	0.3046
t-sne Mahalanobis	5	0.2999	0.7115	0.2885	0.4231
t-sne robust Mahal., bdp=0.1	5	0.3873	0.7340	0.2660	0.4680
t-sne robust Mahal., bdp=0.3	5	0.2079	0.6702	0.3298	0.3404
Scenario 2: tclust ($\alpha = 0.01$)					
Data	4	0.3751	0.7249	0.2751	0.4499
t-sne Euclidean	5	0.2525	0.6886	0.3114	0.3771
t-sne Mahalanobis	5	0.4965	0.7915	0.2085	0.5831
t-sne robust Mahal., bdp=0.1	5	0.4748	0.7848	0.2152	0.5697
t-sne robust Mahal., bdp=0.3	5	0.5130	0.7981	0.2019	0.5962
Scenario 3: tclust ($\alpha = 0.01$)					
Data	3	0.5309	0.7655	0.2345	0.5310
t-sne Euclidean	5	0.3363	0.7295	0.2705	0.4590
t-sne Mahalanobis	5	0.4740	0.7847	0.2153	0.5693
t-sne robust Mahal., bdp=0.1	5	0.5307	0.8053	0.1947	0.6106
t-sne robust Mahal., bdp=0.3	5	0.4839	0.7884	0.2116	0.5767

References

1. García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A.: A general trimming approach to robust cluster analysis, The Annals of Statistics, 36(3), 1324-1345 (2008)
2. Hinton G. and Roweis, S.: Stochastic neighbor embedding, Advances in neural information processing systems, 15 (2002)
3. Van der Maaten, L. and Hinton, G.: Visualizing data using t-SNE, Journal of machine learning research, 9, 2579-2605 (2008)
4. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint arXiv:1802.03426, https://arxiv.org/abs/1802.03426 (2018)
5. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S., Trapnell, C.: Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nature communications, 11(1), 1537 (2020)
6. Parks, D., Miri, A. L., et al. (2021). UMAP is a novel dimensionality reduction technique for large-scale datasets. Nature Communications, 12(1), 1-14. doi: 10.1038/s41467-020-19457-5
7. Becht, E., McInnes, L., Healy, J., and Dutertre, C. A.: Dimensionality reduction for visualizing singlecell data using UMAP. Nature Biotechnology, 37(1), 38-44. https://doi.org/10.1038/nbt.4314, (2019)
8. Kobak, D., Berens, P.: The art of using t-SNE for single-cell transcriptomics. Nature Communications, 10(1), 1-14. https://doi.org/10.1038/s41467-019-13056-z, (2019).
9. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods, 16(3), 243-245. (2019) https://doi.org/10.1038/s41592-019-0358-1
10. Weber, L. M., Robinson, M. D.: Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry Part A, 89(12), 1084-1096. https://doi.org/10.1002/cyto.a.23030. (2016)
11. Vieth, B., Ntranos, V., et al.: UMAP outperforms t-SNE for single-cell RNA-seq data visualization and analysis. Nature Communications, 10(1), 1-13. doi: 10.1038/s41467-019-13056-z (2019)
12. Weber, L. M., Jost, C., et al.: Comparison of t-SNE and UMAP for quality control of single-cell RNA-seq data. Communications Biology, 2(1), 1-13. doi: 10.1038/s42003-019-0744-7 (2019)

Robust diagnostics for Linear Mixed Models with the Forward Search

Analisi robusta di modelli lineari misti con la Forward Search

Aldo Corbellini, Luigi Grossi and Fabrizio Laurini

Abstract With the aid of the forward search (FS), the robustness of linear mixed models (LMM) with random effects is examined. The FS's extension to the LMM presents new computational difficulties because of the limitations imposed by the model and its estimates. The approach is demonstrated by using actual data to examine coffee exports to the European Union in order to spot outliers that could be indicative of fraud.
Abstract *Viene proposta un'applicazione della Forward Search (FS) a un modello misto lineare (LMM) con effetti casuali. L'estensione di FS a LMM offre nuove sfide computazionali, poiché sono necessarie alcune restrizioni imposte dal modello e dalle relative stime. Il metodo è illustrato da un'applicazione a dati reali in cui vengono analizzate le esportazioni di caffè verso l'Unione Europea per identificare valori anomali che potrebbero essere collegati a potenziali frodi.*

1 Introduction and background

An effective method for studying random effect models is proposed in this research. When repeated assessments of variables are gathered on a sample of people, Linear Mixed Models (LMM), a specific case of which includes random effect models, are particularly appealing.

Aldo Corbellini
Ro.S.A. and Department of Economics, University of Parma, Italy, e-mail: aldo.corbellini@unipr.it

Luigi Grossi
Ro.S.A. and Department Statistical Sciences, University of Padova, Italy, e-mail: luigi.grossi@unipd.it

Fabrizio Laurini
R.o.S.A. and Department of Economics, University of Parma, Italy, e-mail: fabrizio.laurini@unipr.it

Outliers and significant observations have a negative impact on these models' parameter estimates; robust estimators can be used to track their impact. When the actual number of outliers is higher than k, certain existing diagnostics suffer from "masking." Other diagnostics are mostly based on the leave-k-out technique.

In various multivariate contexts, an alternative "monitoring" technique known as the forward search (FS) approach has proven to be very effective in dealing with the masking effect due to its high degree of flexibility, efficiency, and robustness, despite having relatively high computational costs.

In order to track the impact of outliers on estimated coefficients and to offer a method for obtaining a robust estimator, we present a forward search approach to LMM in this study. For cross-sectional data, the correlation structure imposed by LMM models is more complex than that of regression models, which raises additional, difficult problems that we attempt to address in this study.

The paper covers a genuine data problem that has several intriguing properties and is used as an illustration. Section 3 discusses the forward search's extension to LMM as well as the key challenges involved. The forward search is then applied to real trade data, with some discussion of the results and potential future developments.

2 Data, their features and economic characteristics

The majority of coffee is produced in poor nations. Brazil is one of the major producers. The consumption of coffee, however, is widespread across all of Europe despite no European country being its producer. Because new plants take more than two years to become productive, the price of coffee is defined by a relatively low price elasticity of supply. Similar to supply, demand exhibits low price elasticity since it is relatively steady and only varies in response to significant price changes that occur independently of any increase in income. As a result, there is a significant level of variability and a high likelihood of outliers in coffee price time series.

In this study, we take into account the coffee trade between Brazil, its country of origin, and 18 European destinations. The monthly value (in Euro) and monthly volume (in tons) of coffee sold from Brazil to each destination countries are kept track of. Sales statistics for each month go back around five years. We continue to focus on the 18 nations that had monthly trade with Brazil throughout the previous five years. Therefore, in the context of longitudinal data analysis, our dataset serves as an example of a balanced design. Though modest adjustments might be needed to address some missing data, our conclusions and methodology are not sensitive to departures from that balanced structure.

For more precise notation, we have the value of coffee imported from Brazil available for each destination country (group), represented as V_i, t with $i = 1, \ldots, g$ and $t = 1, \ldots, ng$. Additionally, each country's import volume is indicated by the W_i, t symbol. The possibility of having a varied number of duplicates for each country is already taken into consideration from the aforementioned notation, as the number of replicates ng may vary for each country.

Fig. 1 Left panel: Scatterplot of the data with symbol for each destination. Largest importers are Germany and Italy. Right panel: Time series of relevant quantities, in log scale, for destination Italy

Figure 1 left panel, where each destination country is shown with a different symbol, provides a snapshot of the entire set of data. We also display the recorded time series of V and W for Brazil vs Italy in Figure 1 right panel to illustrate the transactions' monthly time evolution. Although there is an obvious correlation between the two series, a long-term trend may not be there. When the minimum and maximum yearly values occur in June and December, respectively, the seasonality is extremely obvious, but in other months it does not appear to be as strong.

3 The forward search for LMM

We consider the linear mixed model (LMM) using the notation of [4], Chapter 4. The LMM has fixed component $X\beta$ and the random effect u written in the form

$$y = X\beta + Zu + \epsilon, \tag{1}$$

EBLUP are important to obtained the fitted values of the model given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}. \tag{2}$$

The residuals from model (1), which are computed by taking the observed y and the fitted values from (2) are the building bricks to set up the forward search, a sequential procedure which is based on a set of algorithms such that we start from an outlier-free subset of size $m^\star < n$ (the Basic Subset, BSB) and, at every step, the BSB is increased by including units closer to the selected model. In general, inclusion is such that at every step we move from a subset of size m to a subset of size $m+1$, with iterations until all n data are included. When outliers or other influential observations enter the subset, through the monitoring of relevant statistics, sharp movements are recorded. For regression models, this procedure is illustrated in [1] and made precise in [3].

There are several differences from regression to LMM. We start by highlighting that it is common to have time series even for simple random effects models, thus the selection of units belonging to $m^\star$ is made accordingly. To be specific, a sensible approach is to build the initial subset by taking contiguous observations in each group (here represented by all destination countries). Using proper notation, tailored to the set of available regressors (one explanatory variable, one time index for the trend and eleven dummies for the monthly seasonality), we have that, for each destination country i, with $i = 1, \dots, g$, a coherent set of observations is given by fixing a time index $t^{(i)}$, and then selecting contiguous observations, the number of which is driven by k, the number of columns of the X matrix. Specifically, we select an arbitrary set of observations for each group $m_{t^{(i)},t^{(i)}+1,\dots,t^{(i)}+k+1}$ which ensures that the model is identifiable. A similar choice is made for all groups but, in general, $t^{(i)} \neq t^{(j)}$, with $j = 1, \dots, g$ and $i \neq j$. An initial subset M1 given by

$$\mathrm{M1} = \bigcup_{i=1}^{g} \{m_{t^{(i)},t^{(i)}+1,\dots,t^{(i)}+k+1}\},$$

is then used to fit the random effect model. After fitting the model the following algorithm is performed

1. Squared residuals $r^2_{i,t} = (y_{i,t} - \hat{y}_{i,t})^2$, for all units are computed (even for those that did not contribute to the fitting). With this notation $y_{i,t}$ and $\hat{y}_{i,t}$ denote a single element from the vector expressed in equations (1) and (2) respectively.
2. Squared residuals are sorted yielding to $r^2_{i,t}[\star_i]$, with the argument $[\star_i]$ denoting that the sorting is kept separated for each group.
3. The median of $r^2_{i,t}[\star_i]$ is computed for each i and then stored.

Steps 1 to 3 above are then repeated by changing, for each group, the time index $t^{(i)}$ and leading to sets M2, M3, ..., all having the same size. This procedure is repeated 10000 times, as suggested by [3], since choosing among all possible subsets is unfeasible for almost all practical applications.

For each group the observations that lead to the smallest median of sorted residuals are those contributing to the BSB $m^\star$. Hence, not necessarily the time index to build $m^\star$ is identical for all groups, but inside each group, the contiguity of observation is at this stage guaranteed. This last restriction, however, might be relaxed when having an X matrix with a time index as explanatory variable, since the sorting of the data is irrelevant in the fitting when the time dependence is explicitly modelled with a time trend variable.

As stated above, once fitted the model, computing residuals during the forward search is quite similar to standard regression. The fitted values (2) are obtained after getting $\hat{\beta}$, $\hat{\sigma}_u$ and $\hat{\sigma}_\epsilon$. Plugging such estimates into BLUP gives estimates of $\hat{u}$ and $\hat{y}$.

Some discussion is needed when fitted values (2) and residuals are computed for all units, i.e. also for units that did not contribute to the estimation step. Fitted values and residuals are obtained by considering all entries of X (similarly to the regression model) but using only units that contributed to the fit to extract entries of Z (unlike the regression model). This last restriction is needed to ensure the conformability of product of matrices.

The key difference with regression model, however, comes when moving from m to $m+1$ because of the presence of groups in the data (here represented by each destination). The move from m to $m+1$ requires, once again, the sorting of squared residuals.

As stated above, the inclusion of new units, i.e. the move from size m to $m+1$, is based on the squared residuals computed for all data. As for the choice of the BSB, once residuals have been computed, the sorting is made separately for every group. Therefore, the difference compared to the forward search in regression, is that the sorting is not made on the whole dataset, but split by group. In practice, there are squared residuals sorted for the first group, then squared residuals sorted for the second group, and so forth.

This implies that, the $(m+1)$-th observations joining the dataset are the ones for which the squared sorted residuals not belonging to the m-th step are smallest, disregarding group ownership; e.g. if observations belonging to the same group are well described by the model, and consequently all have the smallest residuals, then the units forming the subset at steps $m+1$, $m+2$, ..., will belong to the same group. Our forward search algorithm is quite flexible and relatively general, as it does not require any group membership balancing during the procedure, the only constraint is that, at each step of the forward search, all groups must have only the minimum number of observations, say $m_{t^{(i)},t^{(i)}+1,\ldots,t^{(i)}+k+1}$, such that the full model is identifiable. Therefore, at each step of the forward search, the size of each group belonging to the subset of size m can be, potentially, very different from the previous step. As stated above, when discussing the choice of the BSB, the time index in the inclusion is not considered, as there is an explanatory variable taking the time trend into account, so that the fit does not require, necessarily, observations to be contiguous.

The peculiar feature of the forward search is that the inclusion of outliers or the inclusion of influential observations is highlighted by sharp peaks moving from m to $m+1$. This is also true for LMM. Influential observations, hidden structures, or outliers display similar pattern in many diagnostics summaries, like plot of residuals and standardized estimates of random effects.

4 Illustration for trade coffee data

Running the forward search yields results for the coffee data, introduced and explained in Section 2, that do not reveal any significant observations in the data. Examining the estimated standardized random effects during the forward search reveals a snapshot of such a summary. It is challenging to determine whether outliers or hidden structures are present in the dataset given that there are 18 destination nations. These are shown in Figure 2 in the right panel.

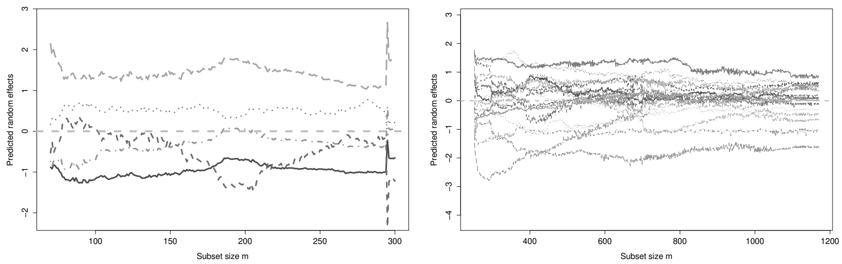


Fig. 2 Estimates of random effects on the simulated and contaminated data during the forward search (left panel). Peaks toward the end of procedure shows the steps when outliers are included. In the right panel estimates of random effects on the coffee trade data during the forward search. A quite smooth behaviour is visible, denoting lack of evidence of suspicious trades

Robust methods frequently have a tendency to detect outliers even when they are absent (false signal). This undesirable characteristic does not affect the forward search. To better understand the false discovery rate of the forward search and compare it to other robust strategies, we will run more simulations in the future with clean and contaminated data.

5 Final remarks

We have introduced the forward search for LMM, where repeated measurements cause correlation. The robust and effective method that has been introduced has two main advantages: when outliers are present, they are correctly identified, and estimations are unaffected by their presence. The forward search does not mark observations as outliers (false signals) when the data are devoid of outliers or when the data do not contain any hidden structure.

In future developments more general models should be taken into account, and there is also the issue of needing a reliable method for choosing the right model.

References

1. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer–Verlag, New York (2000)
2. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc., B, 71, 447–466 (2009)
3. Riani, M., Atkinson, A.C., Cerioli, A.: The forward search: Theory and data analysis. J. Kor. Stat. Soc., 39, 117–134 (2010)
4. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press (2003)

Solicited Session SS15 - *Advanced statistical methods for pattern recognition*

Organizer: Francesca Fortuna
Chair: Stefano Antonio Gattone

1. *Unsupervised classification of NPLs recovery curves* (Carleo A. and Rocci R.)
2. *Living alone in Italian municipalities* (Vellucci P., Benassi F., Naccarato A. and Gallo G.)
3. *Supervised learning from high-dimensional data through dynamic updating of functional classification rules* (Maturo F., Fortuna F. and Di Battista T.)
4. *Assessing the effectiveness of coordination among public authorities in cohesion expenditure* (Coco G., Monturano G. and Resce G.)

Unsupervised classification of NPLs recovery curves

Classificazione non supervisionata di curve di recupero di crediti deteriorati

Alessandra Carleo and Roberto Rocci

Abstract The recovery performance of a portfolio of non-performing loans (NPL) can be measured in terms of recovery rate and liquidation time jointly through a "recovery curve" representative of recovery rates over time. When portfolio heterogeneity is very high, it is more informative to estimate more than just one curve by dividing the portfolio into several homogeneous subsets, i.e. clusters, and calculating a recovery curve for each of them. The aim of this work is to estimate the optimal portfolio partition and the smoothed recovery curves of each cluster by means of non-parametric statistical learning techniques.

Abstract *Il recupero di un portafoglio di crediti deteriorati (NPL) può essere misurato in termini di tasso di recupero e tempo di liquidazione, considerati congiuntamente, attraverso una "curva di recupero" rappresentativa dei tassi di recupero nel tempo. Se l'eterogeneità del portafoglio è molto elevata, è più informativo stimare più di una unica curva, suddividendo il portafoglio in più sottoinsiemi omogenei, i.e. cluster, e calcolando una curva di recupero per ciascuno di essi. L'obiettivo di questo lavoro è stimare la partizione ottimale del portafoglio e le curve di recupero, ipotizzate lisce, ottenute per ciascun sottoinsieme utilizzando tecniche di apprendimento statistico non parametrico.*

Key words: *K*-means, NPL, Recovery curve, Censored data, Smoothing.

Alessandra Carleo
Roma Tre University, Via Sivio D'Amico 77, 00145 – Roma, Italy, e-mail: alessandra.carleo@uniroma3.it

Roberto Rocci
Sapienza University of Rome, P.le Aldo Moro 5, 00185 – Roma, Italy, e-mail: roberto.rocci@uniroma1.it

1 Introduction

Non Performing Loans (NPLs) are exposures in state of insolvency. Their effective recovery depends on the characteristics of the exposure, of the counterparty, on macroeconomic factors, and on internal bank factors. There is a NPL market where banks can get rid of non-performing loans by selling them to specialized operators who deal with recovery. The main method for determining the value of a NPL is that of discounted financial flows, and the valuation methodologies currently used on the market are based primarily on forecast models of the amount of net repayments expected from receivables and related collection times.

The estimation methodology for recovery rate, which is interested in for NPLs, was faced in the more general context of Basel II, where loss given default (LGD) is defined as the proportion of money financial institutions fail to gather during the collection period, and conversely, recovery rate (RR) is defined as the proportion of money financial institutions successfully collect. That means LGD = 1 − RR.

The recovery rate (or LGD) can be estimated using both parametric and non-parametric statistical learning methods. Mainly, the recovery rate is estimated using parametric methods and considering a one-year time horizon. Reference papers for a comprehensive overview of the existing literature are [17, 18, 20], where there are comparative studies on different methods to model the recovery rate. Among parametric models, linear regression is the most common and simplest technique used to predict the mean, whereas to model the overall LGD distribution, or at least certain quantiles of it, linear quantile regression [11, 16] and mixture distributions [2, 4, 14, 24, 25] are proposed as well as other machine learning methods [3, 8, 9, 15]. Also, evidence of multistage models is in [6, 23].

In the case of NPLs, the focus must be both on the recovered amounts and on the duration of the recovery process, the so-called time to liquidate (TTL). Examples of how to consider both size and time of future repayments can be found mainly in papers that use survival modeling techniques applicable to duration processes [1, 5, 13, 26, 27]. In [7] and [22], authors proposed a particular nonparametric method to measure the performance of an NPL portfolio in terms of recovery rate (RR) and time to liquidate (TTL) jointly, by the means of a recovery curve showing how the RR is distributed over time. They also considered a method to estimate such a curve when some data are censored, i.e., when the repayment history for some NPLs is known only until a particular time point.

In this paper, we consider the case where the heterogeneity of the portfolio is very high and a single recovery curve is not representative of the portfolio. Our idea is to estimate an optimal partition of the portfolio in sub-portfolios, i.e. clusters, and to compute a different recovery curve for each one. The plan of the paper is the following. In Sect. 2, we introduce the recovery curve and how it is estimated, also in the case of censored data. In Sect. 3, we propose a method to split the portfolio in homogeneous clusters, each one with its recovery curve, which is smoothed in Sect. 4.

2 Recovery curve of a NPL portfolio

In the case of a single NPL it is clear how to define its RR and TTL. The generalization of such definitions to the case of a NPLs portfolio is not trivial. The two quantities are strictly related and it is crucial to decide when to measure the RR and/or the TTL. As an example, we have to decide if the measurement should be performed when the last NPL has been liquidated or when a significant part of the portfolio has been recovered. In order to avoid such problems, Carleo et al [7] proposed to measure how the RR is distributed during the time. This also helps in deciding at what RR point to measure the TTL and vice-versa. Such a measurement corresponds to build a curve in the following way.

For the k-th ($k = 1, 2, \dots, K$) NPL of the portfolio, the debt exposure at default is indicated with EAD_k, while p_{kt} is the recovery in time interval t ($t = 1, 2, \dots, T$). The quantity

$$r_t = \frac{\sum_{k=1}^{K} p_{kt}}{\sum_{k=1}^{K} EAD_k} \tag{1}$$

is the portfolio recovery rate in the t-th time interval of delay, i.e. the quantity describing the recovery curve of the portfolio when t varies. The computation of this curve is quite simple and intuitive; however it could become impossible when some data are right censored, i.e. the data regarding some NPLs is available only until a time less than T. In order to overcome this problem by using the maximum amount of available information, Carleo et al [7], borrowing some ideas from survival analysis [13], first defined the conditional recovery rate at time t as

$$c_t = \frac{\sum_{k=1}^{K} v_{kt} p_{kt}}{\sum_{k=1}^{K} v_{kt} E_{kt}}, \tag{2}$$

where v_{kt} is 0 if data is not available for the k-th NPL at time t and 1 otherwise, and

$$E_{kt} = EAD_k - \sum_{s=1}^{t-1} p_{ks} \tag{3}$$

is the exposure of the k-th NPL at the beginning of time interval t. Then they suggest computing the recovery rate as

$$r_t = \left(1 - \sum_{s=1}^{t-1} r_s\right) c_t. \tag{4}$$

It can be shown that the two different methods, i.e. formula (1) and (4), coincide when there aren't censored data. It is important to note that the logic of formula (2) of excluding the NPL censored can not be used in formula (1) because could produce incoherent recovery curves.

Table 1 Example of NPL portfolio with censored data.

NPL	EAD_k	p_{k1}	p_{k2}
1	100	10	80
2	100	90	-

Let us consider the example in Table 1. By excluding the censored NPL from the computation of the recovery rates in (1) we would obtain the sequence $r_1 = 0.50$, $r_2 = 0.80$ that is incoherent because it sums to 1.30 > 1.00. With the second method, we obtain $r_1 = 0.50$, $r_2 = 0.44$.

It is interesting to note that the conditional recovery rate can be also defined for a single NPL as $c_{kt} = p_{kt}/E_{kt}$, and the portfolio's recovery rate c_t can be seen as a weighted average of such quantities

$$\begin{aligned} c_t &= \frac{\sum_{k=1}^{K} v_{kt} p_{kt}}{\sum_{h=1}^{K} v_{ht} E_{ht}} = \frac{1}{\sum_{h=1}^{K} v_{ht} E_{ht}} \sum_{k=1}^{K} \frac{p_{kt}}{E_{kt}} v_{kt} E_{kt} \\ &= \frac{1}{\sum_{h=1}^{K} v_{ht} E_{ht}} \sum_{k=1}^{K} c_{kt} v_{kt} E_{kt} \end{aligned} \tag{5}$$

and then, as the solution of the least squares problem

$$c_t = \underset{m_t}{\operatorname{argmin}} \sum_{k=1}^{K} v_{kt} E_{kt} (c_{kt} - m_t)^2. \tag{6}$$

Further information about this method can be found in [7] and [22].

3 Unsupervised classification of recovery curves

In the previous section, we have seen how to estimate the recovery curve of a portfolio as the weighted mean of the recovery curves of the single NPLs. However, in some cases, when the heterogeneity of the portfolio is very high, it would be better to estimate more than only one curve. In other terms, if the recovery curves of the single NPLs are very different, it would be more informative to split the portfolio in several homogeneous sub-portfolios, say clusters, and compute one recovery curve for each one. The optimal partition can be estimated by generalizing the loss in (6) to the case of G clusters as

$$L(\mathbf{U}, \mathbf{M}; \mathbf{V}, \mathbf{E}, \mathbf{C}) = \sum_{g=1}^{G} u_{kg} \sum_{t=1}^{T} \sum_{k=1}^{K} v_{kt} E_{kt} \left(c_{kt} - m_{gt}\right)^2 \tag{7}$$

where

- **U** = [u_{kg}] is the $K \times G$ binary membership matrix defining the partition setting u_{kg} = 1 if NPL k belong to cluster g and 0 otherwise;
- **M** = [m_{gt}] is the $G \times T$ matrix of curve centroids. The g-th row describes the recovery curve of cluster g;
- **V** = [v_{kt}] is the $K \times T$ binary matrix describing the censoring of the NPL trajectories, where v_{kt} = 1 if the datum p_{kt} is available and v_{kt} = 0 otherwise;
- **E** = [E_{kt}] is the $K \times T$ matrix with elements E_{kt}, that is the exposure of NPL k at the beginning of time interval t;
- **C** = [c_{kt}] is the $K \times T$ data matrix containing the conditional recoveries of each NPL in every interval.

The loss in (7) is a sort of weighted K-means [19]. It can be minimized iteratively by using a coordinate descent algorithm.

4 Smoothing the recovery curves

In some cases, we expect to find smooth curves. In order to exploit this a priori information in our methodology, like in functional data analysis [21], we can follow two different approaches. The first is a filtering approach where the data curves are pre-smoothed before the analysis. The second is a regularized approach where the smoothing step is simultaneously performed with the analysis (see for example [10]) adding a quadratic penalty to the loss function in order to penalize non smooth solutions. In our case, loss (7) becomes

$$\begin{aligned} L(\mathbf{U},\mathbf{M};\mathbf{V},\mathbf{E},\mathbf{C}) = &\sum_{g=1}^{G} u_{kg} \sum_{t=1}^{T}\sum_{k=1}^{K} v_{kt}E_{kt}\left(c_{kt}-m_{gt}\right)^2 \\ &+ \lambda \sum_{g=1}^{G}\sum_{s=1}^{T}\sum_{t=1}^{T} m_{gs}d_{st}m_{gt} \end{aligned} \tag{8}$$

where the matrix **D** = [d_{st}] is the penalty matrix used in smoothing spline regression (see [12]).

Further information about the proposed methodology, details about the algorithm and the results of some tests about its effectiveness, performed on real and simulated data, will be given in the complete version of the paper.

References

1. Altman, E.I.: Measuring Corporate Bond Mortality and Performance. J. Financ., 44, 909–922 (1989)
2. Altman, I.E., Kalotay, A.E.: Ultimate Recovery Mixtures. J. Bank. Financ., 40, 116–129 (2014)
3. Bellotti, A., Brigo, D., Gambetti, P., Vrins, F.: Forecasting recovery rates on non-performing loans with machine learning. Int. J. Forecast., 37, 428–444 (2021)
4. Betz, J., Kellner, R., Rösch, D.: Systematic Effects among Loss Given Defaults and their Implications on Downturn Estimation. Eur. J. Oper. Res., 271, 1113–1144 (2018)

5. Betz, J., Kellner, R., Rösch, D.: Time matters: How default resolution times impact final loss rates. J. R. Stat. Soc. Ser. C, 70, 619–644 (2021)
6. Bijak, K., Thomas, L.: Modelling LGD for unsecured retail loans using Bayesian methods. J. Oper. Res. Soc., 66, 342–352 (2015)
7. Carleo, A., Rocci, R., Staffa, M.S.: Measuring the Recovery Performance of a Portfolio of NPLs. Computation, 11(2), 29 (2023)
8. Cheng, D., Cirillo, P.: A reinforced urn process modeling of recovery rates and recovery times. J. Bank. Financ. 96, 1–17, (2018)
9. Gambetti, P., Roccazzella, F., Vrins, F.: Meta-Learning Approaches for Recovery Rate Prediction. Risks, 10, 124 (2022)
10. Gattone, S.A., Rocci R.: Clustering Curves on a Reduced Subspace. Journal of Computational and Graphical Statistics, 21(2), pp. 361-379 (2012)
11. Gostkowski, M., Gajowniczek, K.: Weighted Quantile Regression Forests for Bimodal Distribution Modeling: A Loss Given Default Case. Entropy, 22, 545 (2020)
12. Green, P. J., Silverman, B.W.: Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapmann & Hall, London (1994)
13. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis Failure Time Data. John Wiley & Sons, Hoboken, NJ, USA (2002)
14. Kalotay, A.E., Altman, I.E.: Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses. Rev. Financ., 21, 433–463 (2017)
15. Kaposty, F., Kriebel, J., Löderbusch, M.: Predicting loss given default in leasing: A closer look at models and variable selection. Int. J. Forecast., 36, 248–266 (2020)
16. Krüger, S., Rösch, D.: Downturn LGD Modeling using Quantile Regression. J. Bank. Financ., 79, 42–56 (2017)
17. Loterman, G., Brown, I., Martens, D., Mues, C., Baesens, B.: Benchmarking regression algorithms for loss given default modeling. Int. J. Forecast., 28, 161–170 (2012)
18. Min, A., Scherer, M., Schischke, A., Zagst, R.: Modeling Recovery Rates of Small and Medium-Sized Entities in the US. Mathematics, 8(11), 1856 (2020)
19. Mirkin, B.: Quadratic Error and k-Means. In: Hennig, C., Meila, M., Murtagh, F., Rocci, R. (eds.) Handbook of Cluster Analysis. Chapman and Hall/CRC (2015)
20. Qi, M., Zhao, X.: Comparison of modeling methods for loss given default. J. Bank. Financ., 35, 2842–2855 (2011)
21. Ramsay, J.O. and Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)
22. Rocci, R., Carleo, A., Staffa, M.S.: Estimating Recovery Curve for NPLs. In: Corazza, M., Perna, C., Pizzi, C., Sibillo, M. (eds.) Mathematical and Statistical Methods for Actuarial Sciences and Finance. MAF 2022, pp. 397–403. Springer Cham, Switzerland, (2022)
23. Sun, H.S., Jin, Z.: Estimating credit risk parameters using ensemble learning methods: An empirical study on loss given default. J. Credit Risk, 12, 43–69 (2016)
24. Tomarchio, S.D., Punzo, A.: Modelling the Loss Given Default Distribution via a Family of Zero-and-one Inflated Mixture Models. J. R. Stat. Soc. Ser. A, 182, 1247–1266 (2019)
25. Ye, H., Bellotti, A.: Modelling Recovery Rates for Non-performing Loans. Risks, 7, 19 (2019)
26. Witzany, J., Rychnovsky, M., Charamza, P.: Survival Analysis in LGD Modeling. Eur. Financ. Account. J., 7, 6–27 (2012)
27. Zhang, J., Thomas, L.C.: Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. Int. J. Forecast., 28, 204–215 (2012)

Living alone in Italian municipalities

Vivere da soli nei comuni italiani

Pierluigi Vellucci, Federico Benassi, Alessia Naccarato and Gerardo Gallo

Abstract New family structures have emerged in Italy in recent decades, with a trend towards smaller nuclear families due to demographic and social changes. An aging population, marital instability, declining fertility, and later marriage have contributed to this trend. It is important to understand the changing needs of families, especially the vulnerable, from both an economic and social perspective. Vulnerability is often related to economic factors, but single-person households, are often at risk. The goal of this study is to classify Italian municipalities based on single-member household characteristics, identifying areas of greater or lesser fragility.

Abstract *Negli ultimi decenni sono emerse in Italia nuove strutture familiari, la cui caratteristica principale è la riduzione del numero di componenti e l'aumento del numero di famiglie mono-personali. Fattori come l'invecchiamento della popolazione, l'instabilità matrimoniale, la diminuzione della fertilità e il matrimonio più tardivo hanno contribuito a questa evoluzione. In ottica di policies, bisogna osservare che i nuclei familiari composti da una sola persona, sono maggiormente esposti a situazioni di vulnerabilità. Lo scopo di questo studio è classificare i comuni italiani in base alle caratteristiche dei nuclei familiari monocomponenti, identificando le aree di maggiore o minore fragilità.*

Pierluigi Vellucci
Dept. of Economics, Roma Tre University, Via Silvio D'Amico 77, 00145 Rome, Italy, e-mail: pierluigi.vellucci@uniroma3.it

Federico Benassi
Dept. of Political Sciences, University of Naples Federico II, Via Leopoldo Rodinò 22, 80133 Naples, Italy, e-mail: federico.benassi@unina.it

Gerardo Gallo
Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Rome, Italy, e-mail: gegallo@istat.it

Alessia Naccarato
Dept. of Economics, Roma Tre University, Via Silvio D'Amico 77, 00145 Rome, Italy, e-mail: alessia.naccarato@uniroma3.it

Key words: living alone people, decision algorithm, geographical pattern

1 Introduction

In the last decades new forms of families spread in Italy. The combination of both social and demographic changes resulted in a families' nuclearization process, manifested by the progressive decrease in the average number of members [1]. There are many demographic factors that have contributed most to this process. The aging of the population, which has led many elderly people to live alone, and the marital instability. These drivers are compounded by declining fertility [8], longer length of stay in the family of origin by young adults [3], and higher age of first marriage [5]. From the perspective of both economic and social planning, it is necessary to understand how the needs of families have changed, especially for the most vulnerable ones. The vulnerability of families is generally related to economic dimensions [2] however, people living alone are more likely to experience fragile situations, particularly in cases of elderly person. The aim of this contribution is to classify the Italian municipalities according to the characteristics of single-member households and to study their spatial distribution. This allows us to highlight territories of greater or lesser fragility and identify in which territories it is most important to intervene.

2 Data description

In order to classify Italian municipalities according to the types of single-member households, data from the 2020 Italian Permanent Population and Housing Census (PPHC) on households residing in Italy's 7903 municipalities were used. The characteristics of the single-member households refer to age (AGE), gender (GEN), citizenship (CIT), ownership of one or more dwellings (DWE), ownership of one or more cars (CAR), educational (EDU), whether they receive citizenship income or not (CIN). Moreover, we consider six characteristics of municipalities, geographic macro-area (GA), region (REG), degree of urbanization (DU), number of beds available in elderly housing facilities (BED), local wealth (WEA), and finally a categorical variable that groups municipalities according to their demographic profiles and attractiveness (ATT).

3 Methods

We employ the labelled dataset described in Sect. 2 to identify, by means of a decision algorithm, the amount of municipalities presenting a ”critical” percentage of single-person households. In this first experiment we consider as ”critical” those municipalitieas having percentage grater than the average percentage computed for Italy. Only municipalities so identified were classified according to the prevalent single-member household type (PSMH). The target variable, which our supervised learning engine tries to predict, is the status of being a *critical municipality*. If we denote with i the municipality of our dataset and with y_i the ratio between the number of single-person households in 2020 and the total amount of population in the same year in the municipality i, a municipality i is called *critical* if y_i is larger than the average value of y_i over the entire dataset.

3.1 The decision algorithm

Decision algorithms [6] are commonly used in machine learning and data mining tasks to build predictive models. As a criterion for the splitting we will use the Gini impurity. The Gini impurity measures the probability of incorrectly classifying a randomly chosen element from the set. For a binary classification problem, where there are two classes (0 and 1), the Gini impurity can be defined as:

$$Gini(p) = 1 - \left(p_0^2 + p_1^2\right)$$

where p_0 and p_1 represent the probabilities of each class in the subset of data being considered. The cost function quantifies the quality of a split based on the reduction in impurity achieved. A common cost function is the weighted sum of the impurities of the resulting subsets after the split.

$$Cost = \frac{N_{\text{left}}}{N} \cdot Gini_{\text{left}} + \frac{N_{\text{right}}}{N} \cdot Gini_{\text{right}}$$

where N_{left} and N_{right} are the number of instances in the left and right subsets, N is the total number of instances, and $Gini_{\text{left}}$ and $Gini_{\text{right}}$ are the Gini impurities of the left and right subsets, respectively.

3.2 Spatial analysis

The decision algorithm leads us to the identification of 11 PSMHs on which a first spatial analysis has been done. Using the 11 PSMH typologies, we built a categorial map at municipality level. This map allows us to reconstruct the (local) spatial distribution of each typology and therefore to verify the existence (or absence) of specific geographical patterns. In a second step, for each municipality, we computed and mapped the location quotient [7]. The location quotient (LQ) is used to measure and map relative distributions or relative concentrations of a character in a subarea compared to the area as a whole [4], [9]. LQ varies from 0 to ∞. It is lower that 1 in the spatial units in which a population group is underrepresented or conversely, overrepresented (LQ>1). In our analysis the population group is the typology of single member households in each municipality. The identification of the cut offs in the distribution of LQ is an open issue. In our exercise we defined 3 classes: a first (LQ $\leq$ 1.0) in which there is absence of over representation; a second (1.0 $<$ LQ $\leq$1.5) in which there is a moderate over representation; and a third (LQ $>$ 1.5) characterised by a high over representation. Maps have been done by using QGIS Desktop version 3.20.2.

4 Results and discussion

In Figure 1 we depict the visual representation of the decision tree illustrating the classification of municipalities into 'critical' or 'non critical' categories based on key predictor variables. From Figure 1 we can spot the presence of 11 classes of

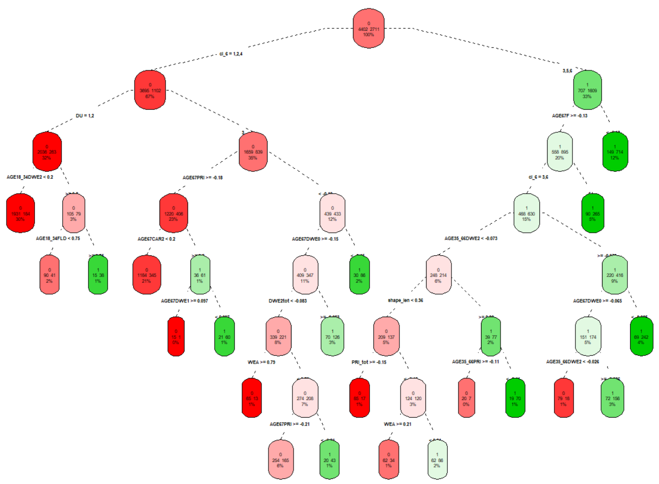


Fig. 1: Decision Tree for the full set of features according to the Gini index

critical municipalities. They can be summarized as follows[1].

Due to brevity constraints, we cannot provide a detailed description of the classes, but we can highlight some key findings. With the exception of one class that includes young individuals with a university education and home ownership, single-person households are predominantly made up of elderly individuals. A smaller number of municipalities are characterized by economically and socially stable sin-

[1] Please refer to the following legend. Degree of urbanization classifies municipalities into Cities (1), Small towns or suburbs (2) and Rural areas (3). The attractiveness variable refers to the systematic depopulation of Italian municipalities as described by [1].

gle elderly individuals, while in the majority of cases they experience situations of vulnerability.

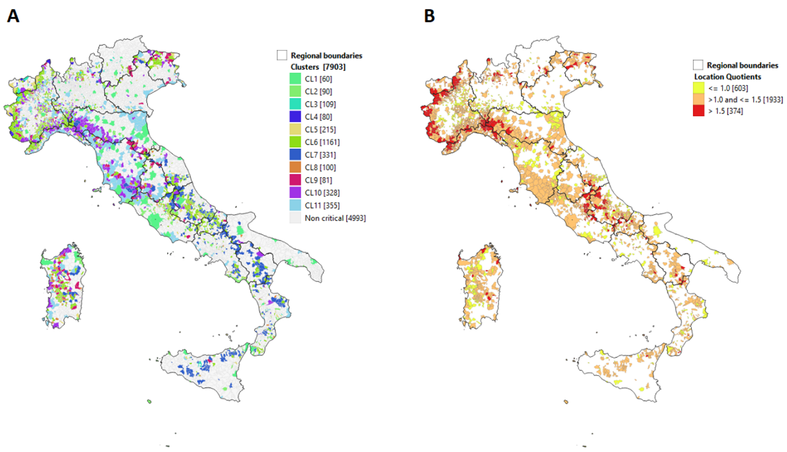


Fig. 2: Categorial map of clusters' typologies (panel A) and LQs map (panel B)

Looking at the two maps of Figure 2 we can infer some interesting (albeit partial) results. The distribution of single-member household typologies does not follow the classic south-north divide that characterised most of the socio-economic processes in Italy. Nevertheless, if we look to the distribution of the single typology some geographical patterns emerge. For a sake of brevity, we recall here just a few: the big cities are divided in two clusters (CL 1 and CL11). The first includes Rome, Bologna, and some other important cities of the Southern Italy like Palermo, Bari Naples, and Reggio di Calabria. The second holds Milan, Genoa, Venice, Turin, and some other metropolitan cities located in the Centre (Florence) and the South (Cagliari, Catania, Messina).These two urban clusters are contrasted by clusters composed of rural and more marginal municipalities like the CL6 and CL7. The first is composed by municipalities that are mainly located on the border of the Italian country especially on the North-West border (Piedmont region) and in some inner areas (mainly located in Abruzzo, Molise, and Sardinia regions). CL 7 is composed by isolated and quite marginal municipalities that are located in majority in the South part of Italy. It is interesting to note a quite different level of "mixing" of each region. It is quite evident that in some regional contexts (i.e. Sardinia and Tuscany) we find different typologies of clusters so that there isn't a typology that is clearly predominant. In other region (i.e. Lazio and Apulia) the typologies of clusters are very few. The LQs map (panel B of Figure 2) reveals some other interesting details. The condition of a severe over representation is related to a small number of municipalities (374 over 2.910). Nevertheless, their geographies is quite peculiar: they are concentrated in specific areas of the country defining some spatial patterns. One is located on the Northwest border, another is located at the intersection of three regions (Emilia-Romagna, Liguria and Piedmont) while the last is again located at the intersection of three regions (Lazio, Marche and Abruzzi).

5 Conclusion

This is a first explorative analysis that shows from one hand the potentialities of using different data source (coming from official statistics and other sources) at very fine geographical scale. And, on the other, the usefulness in the classification of municipalities. These last allowed the identification of specific spatial patterns of certain typologies (or categories). This also results in focusing on particularly critical areas (i.e., municipalities where the level of over representation is comparatively high, LQ >1.5). Results tell us that there is no clear north-south dynamic which is, in a certain manner, a novel aspect in the socio-economic and demographic literature. However, some patterns emerge: 1) a distinction between large cities and the other municipalities; 2) area of internal marginalisation; 3) a different level of heterogeneity inside of each regional context. The study needs further development such as to modify the initial classification criteria, i.e. considering the quantiles of the target variable to identify ”critical” municipalities, and to measure the level of inner heterogeneity of each region using specific indeces.

References

1. AISP – Associazione Italiana per gli Studi di Popolazione: Rapporto sulla popolazione, Le famiglie in Italia. Forme, ostacoli, sfide. Eds. Vignoli, D., Tomassini, C., Il Mulino, 10-34, (2023)
2. Benassi, F., Naccarato, A.: Households in potential economic distress. A geographically weighted regression model for Italy, 2001-2011, Spat. Stat., 21, 362-376 (2017)
3. Billari, F. C., Manfredi, P., Valentini, A.: Macro-demographic effects of the transition to adulthood: Multistate stable population theory and an application to Italy, Math. Popul. Stud., 9(1), 33-63 (2000)
4. Crawley A., Beynon M., Munday M.:Making location quotients more relevant as a policy aid in regional spatial analysis. Urban Stud., 50(9), 1854-1869 (2013)
5. Gabrielli G., Vignoli D.: The Breaking-Down of Marriage in Italy: Trends and Trendsetters, Popul. Rev., 52(1) (2013)
6. Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction. Vol. 2. Springer, New York (2009)
7. Isard W.: Methods of regional analysis: an introduction to regional science. Cambridge: The MIT Press (1960)
8. Kertzer, D. I., White, M. J., Bernardi, L., Gabrielli, G.: Italy’s Path to Very Low Fertility: The Adequacy of Economic and Second Demographic Transition Theories, Eur. J. Popul., 25(1), 89-115 (2009)
9. Wheeler J.O.: Geography. In: Kimberly K.L. (Ed.) Encyclopedia of Social Measurement, Elsevier, 115-123 (2005)

Supervised learning from high-dimensional data through dynamic updating of functional classification rules

Apprendimento supervisionato da dati ad alta dimensionalità attraverso l'aggiornamento dinamico delle regole di classificazione funzionali

Fabrizio Maturo, Francesca Fortuna and Tonio Di Battista

Abstract This article addresses the difficulties involved in analyzing data streams due to their high velocity and volume, which renders traditional statistical techniques impractical. The study proposes a multi-step approach to reduce the dimensionality of data streams using functional data analysis and functional classification trees trained in separate time windows, resulting in an ensemble called the dynamic functional forest. The functional classification trees are trained using the scores obtained from the functional principal components decomposition within each interval, and the majority vote of the ensemble provides the prediction of class labels across the entire time domain. The objective is updating the classification rules systematically as new data becomes available, without the need to store past data.
Abstract *Questo articolo affronta le difficoltà legate all'analisi dei flussi di dati ad elevata velocità e volume, che rende impraticabili le tradizionali tecniche statistiche. Si propone un approccio multifase per ridurre la dimensionalità utilizzando l'approccio funzionale e gli alberi di classificazione funzionali addestrati in finestre temporali separate, creando un insieme detto foresta funzionale dinamica. Gli alberi di classificazione vengono addestrati con i coefficienti ottenuti dalla decomposizione in componenti principali funzionali in ciascun intervallo e il voto di maggioranza dell'insieme fornisce la previsione delle etichette di classe per l'intero dominio. L'obiettivo è quello di aggiornare le regole di classificazione man mano che nuovi dati diventano disponibili, senza memorizzare i dati passati.*

Key words: classification; datastreams; dynamic functional forest; functional principal components; functional data analysis.

Fabrizio Maturo
Universitas Mercatorum, Rome, e-mail: fabrizio.maturo@unimercatorum.it

Francesca Fortuna
Roma Tre University, Rome, e-mail: francesca.fortuna@uniroma3.it

Tonio Di Battista
G. d'Annunzio University, Pescara, e-mail: tonio.dibattista@unich.it

1 Introduction

Technological progress has made it possible to observe vast amounts of data that arrive continuously. These data streams contain a wealth of fast, continuous, mutable, ordered, and unbounded information. Consequently, processing and analysing data streams as soon as they are received is critical. Examples of data streams can be found in various fields, such as mobile data collection platforms, traffic management, and bank transactions. One of the primary challenges in mining data streams lies in the high speed and large volume of incoming information, making it either unnecessary or impractical to store them in their entirety. This leads to problems associated with the curse of dimensionality, rendering traditional data mining and multivariate statistical procedures unsuitable. Additionally, models trained on data streams must be continuously updated and rely heavily on the most recent data within the stream [13]. As a result, this matter has received significant attention in the statistical literature [9, 10]. Typically, data streams are processed by computing suitable data summaries and dividing the time domain into intervals.

Recent studies have proposed analysing high dimensional data using probability density functions via Functional Data Analysis (FDA) [2, 3, 8, 12]. FDA has also gained prominence treating entire streams as single functional objects and has been successfully applied to various empirical concerns, including classification, outlier detection, and forecasting [14, 5, 1, 4, 15, 6]. Several functional classifiers have been proposed for supervised classification, e.g. combining FDA and tree-based methods [7, 9, 11], However, their main limitation is that they quickly become outdated as data streams constantly evolve.

To address this issue, this study offers a methodological tool that combines FDA and dynamic ensembles of classification trees. The proposed approach involves a multi-step procedure for reducing the dimensionality of data streams using functional data representation. In the first step, data streams are divided into non-overlapping time-based windows, and for each, the Functional Principal Component (FPC) decomposition is employed to grasp the fundamental information of the curves within that specific interval. In the second step, functional classification trees (FCTs) [9] are constructed for each time window. These FCTs are trained using the scores obtained from the FPC decomposition, resulting in an ensemble of classifiers named the dynamic functional forest. A majority vote of the ensemble determines each stream's final prediction in terms of class label.

To address the issue of evolving data, a novel algorithm is proposed to update the functional forest by training new FCTs as new information becomes available, removing the need to store old data streams. The remainder of the paper is organised as follows: Sect. 2 presents the proposed algorithm, followed by a simulation study in Sect. 3. Finally, conclusions are provided in Sect. 4.

2 Materials and Methods

Let $\mathbf{S_i} = \{\mathbf{x_{i1}}, \mathbf{x_{i2}}, ..., \mathbf{x_{it}}, ..., \mathbf{x_{iT}}\}$ be the i-th observed data stream, $i = 1, ..., n$, whose generic element x_{it}, $t = 1, ..., T$, is a real-valued scalar observed at time t. S_i can be partitioned into a sequence of J equally spaced time-based windows of length q as follows:

$$W_j = \{x_{q\times(j-1)+1}, .., x_t, ..., x_{q\times j}\}, \quad j = 1, ..., J. \tag{1}$$

For each j-th window, we consider the FPC decomposition to capture the main information of the curve in the specific time period.

In the context of functional supervised classification, the initial step involves having a training set of functions with known labels. Therefore, we have J starting functional training sets of the form $\mathbf{y}, \mathbf{f_j(t)}$, where $\mathbf{f_j(t)}$ represents the set of curves in the j-th time window, and $\mathbf{y}$ is a vector of scalar response values for sample $i = 1, ..., n$, that is assumed to remain constant over time. In this context, we assume the response is categorical and the known labels of the curves are utilized to train a functional classifier, which is a classification rule used to predict the class labels of new functional observations.

The basic idea of this paper is to create a functional classification rule that can be updated when new data are available. In particular, our proposal is to introduce the so-called FCTs proposed by [9, 10] to the context of data streams; therefore, the coefficients of FPC decomposition are used as new features to predict the response at each time interval.

The aim is to define a dynamic functional forest able to be updated as new data are available without the need to store previous data contained in past time windows.

The proposed algorithm is composed by the following steps:

1. Given $S_i = \{x_{i1}, x_{i2}, ..., x_{it}, ..., x_{iT}\}$, create J equally spaced time-based windows of length q;
2. For each W_j window ($j = 1, ..., J$), compute the FPCs;
3. Use the scores of the j-th FPCs to build the j-th FCT for the j-th time window;
4. Delete the original data and store only the j-th FCT building the forest;
5. Predict the curves' labels using the majority vote from the ensemble; the posterior probability for each class is given by $P(\hat{y}_i = G_c \mid W_1, W_2, ..., W_J) = \frac{\sum_{j=1}^{J} I_j(y_i = G_c)}{J}$, where G_c represents the possible labels of Y, i.e. the groups to predict;
6. Evaluate, the performance of the functional forest;
7. After observing q new time observations, repeat steps 2-6 for each new time window W_{j+t}, with $t \in [1, +\infty[$, and update the functional classification rule updating the ensembles.

3 Simulation study

Data streams are simulated considering $n = 200$ observations spanning $T = 600$ time instants (days). Among these observations, $n_1 = 100$ belong to class 1 (red) and $n_2 = 100$ belong to class 2 (black). It is assumed that the class membership remains constant over time. Therefore, as new observations are added, they are utilized to enhance the predictive capability of the classifier by incorporating the new temporal information. The simulation procedure for functional data with two classes follows the same approach as described in scenario 3 of [10].

The first row of Fig. 1 displays how the data evolves over time and is divided into windows of equal width. Once the window width is determined, each new data set is combined with the data from previous intervals. Moving on to the second step of the algorithm, the time-based data for each series are used to compute the FPCs. The charts in the last row of Fig. 1 represent the step three of the algorithm in which separate FCTs are computed for each time window and subsequently pruned using cost-complexity pruning.

Table 1 illustrates the results of the FCTs for each single window and the proposed ensemble. As we can see, the final classification rule, W_{all}, which is based on the majority vote of the ensemble of FTCs, proves to refine the results. Indeed, the proposed algorithm has the power of the ensemble methods in terms of variability reduction and considers the peculiarities of the functions in single windows by capturing the distinctive features of the original classes in detail and specific time intervals.

W_j	Accuracy
W_1	0.83
W_2	0.86
W_3	0.89
W_4	0.95
W_5	0.93
W_{all}	0.96

Table 1 Results of the FCTs for each window and using the proposed weighted ensemble W_{all}.

4 Conclusions

This paper presents a novel algorithm for the classification of data streams using a dynamic classifier that combines FDA and tree-based methods. Our approach extends the concept of FCTs to handle data streams, where a constantly updated functional classifier is required. To assess the accuracy of our functional classifier, we employ the bootstrap method. However, alternative approaches such as cross-

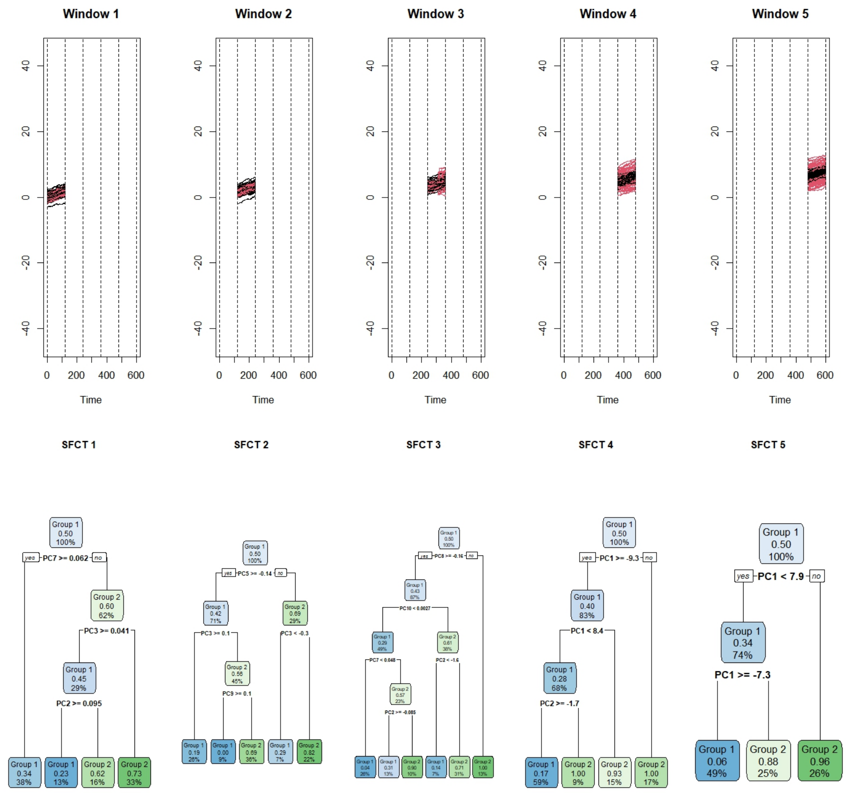


Fig. 1 The visual explanation of the proposed algorithm.

validation or a functional test set, as mentioned in [10], can also be utilized. The primary objective of this study is to propose a classification technique for high-dimensional data that achieves remarkable precision levels while effectively discarding outdated data streams. By retaining only the essential information of the classification rule, our method enables the updating of the functional classification rule with new data, eliminating the need to store large amounts of data that are challenging to accommodate due to their size.

References

1. Aguilera-Morillo, M., Aguilera, A., Escabias, M., Valderrama, MJ.: Penalized spline approaches for functional logit regression. Test. 22(2), 251- 277 (2012)
2. Delicado, P.: Dimensionality reduction when data are density functions. Computat Stat Data Anal. 55, 401–420 (2011)
3. Egozcue, J., Diaz-Barrero, J., Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Math Sin English Series. 22, 1175–1182 (2006)
4. Febrero-Bande, M, de la Fuente, MO.: Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw. 5(4): 1–28 (2012)
5. Ferraty, F, Vieu, P.: Nonparametric Functional Data Analysis. New York, NY: Springer (2006)
6. Maturo, F., Fortuna, F. & Di Battista, T.: Testing Equality of Functions Across Multiple Experimental Conditions for Different Ability Levels in the IRT Context: The Case of the IPRASE TLT 2016 Survey. Soc Indic Res. 146, 19–39 (2019). doi: 10.1007/s11205-018-1893-4
7. Gregorutti, B., Michel, B., Saint-Pierre, P.: Grouped variable importance with random forests and application to multiple functional data analysis. Comput Stat Data Anal. 90, 15–35 (2015)
8. Hron, K., Menafoglio, A., Templ, M., Hruzova, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Comput Stat Data Anal. 94, 330–350 (2016)
9. Maturo, F., Verde, R.: Pooling random forest and functional data analysis for biomedical signals supervised classification: Theory and application to electrocardiogram data. Stat Med. 41(12), 2247–2275 (2022). doi: 10.1002/sim.9353
10. Maturo, F., Verde, R.: Supervised classification of curves via a combined use of functional data analysis and tree-based methods. Comput Stat. (2023). 38, 419–459. doi: 10.1007/s00180-022-01236-1
11. Maturo, F., Verde, R.: Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. Comput Stat. (2022). doi: 10.1007/s00180-022-01259-8
12. Menafoglio, A., Secchi, P., Guadagnini, A.: A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. Math Geosci. 48, 463-485 (2014)
13. Qahtan, A., Wang, S., Zhang, X.: Efficient Estimation of Dynamic Density Functions with Applications in Data Streams (2018) 10.1007/978-3-319-89803-11
14. Ramsay, J, Silverman, B.: Functional Data Analysis. 2nd ed. New York, NY: Springer (2005)
15. Maturo, F., Fortuna, F. and Di Battista, T.: Outliers detection in assessment tests' quality evaluation through the blended use of functional data analysis and item response theory. Ann Oper Res. (2022). doi: 10.1007/s10479-022-05099-z

Assessing the effectiveness of coordination among public authorities in cohesion expenditure

Valutare l'efficacia del coordinamento tra le autorità pubbliche nella spesa per la coesione

Giuseppe Coco, Gianluca Monturano and Giuliano Resce

Abstract Territorial fragilities in reaping the benefits of cohesion policies are partly due to difficulties in allocating and spending resources by central administrations. The efficiency of the allocative mechanism depends on planning and administrative capacities of beneficiary local authorities, project peculiarities, socio-economic dynamics, institutional characteristics, and the need for coordination among authorities. This study focuses on delays in implementing projects for territorial cohesion, using completion dates of each project from OpenCoesione. The results reveal spatial concentration of delays in Italy's most fragile areas, influenced by historical issues affecting Italian growth.
Abstract *Le fragilità territoriali nell'ottenere i benefici delle politiche di coesione sono in parte dovute alle difficoltà nell'allocazione e nella spesa delle risorse da parte delle amministrazioni centrali. L'efficienza del meccanismo allocativo dipende dalle capacità progettuali e amministrative degli enti locali beneficiari, dalle peculiarità dei progetti, dalle dinamiche socio-economiche, dalle caratteristiche istituzionali e dalla necessità di coordinamento tra le diverse autorità coinvolte. Questo studio si concentra sui ritardi nell'attuazione dei progetti per la coesione territoriale, utilizzando le date di completamento di ciascun progetto dal sito OpenCoesion. I risultati rivelano una concentrazione spaziale dei ritardi nelle aree più fragili d'Italia, influenzate da problematiche storiche che incidono sulla crescita italiana.*

Key words: territorial cohesion; cohesion policy; project delay; regional disparities

Giuseppe Coco
Università degli Studi di Bari 'Aldo Moro', Bari, e-mail: giuseppe.coco@uniba.it

Gianluca Monturano
Università degli Studi di Modena-Reggio Emilia, Modena, e-mail: gianluca.monturano@unimore.it

Giuliano Resce
Università degli Studi del Molise, Campobasso, e-mail: giuliano.resce@unimol.it (corresponding author)

1 Introduction and literature review

The guiding principle of the EU's cohesion policy is to reduce territorial inequalities and promote economic, social, and territorial convergence among the regions of the Union [1]. Investments are made in infrastructure, education, research, innovation, environment, and culture to achieve this objective. The cohesion policy utilizes public investments through projects to support economic development and improve the quality of life for European citizens [2]. These projects also contribute to the achievement of the EU's sustainable development goals [3] and foster European integration and economic stability [4]. Place-based policies involve local stakeholders, and the quality of government can influence program implementation [5]. Cohesion policies mitigate territorial disparities in European regions [6]. However, reducing territorial fractures requires long-term commitment and coordinated action at the European, national, and regional levels. Italy is particularly affected by territorial inequalities between the North and South [7]. Southern Italy experiences a development lag compared to the North, which is widening over time [8]. Despite efforts, significant disparities persist throughout the Italian territory, mainly linked to the implementation of cohesion projects [9]. In fact, the eighth report on cohesion policies places the least developed regions in Italy in a development trap for most of the past 15 years [10].

1.1 Delays in cohesion projects

Institutional efforts to improve the planning, management, and implementation of cohesion projects often fail to prevent significant delays compared to scheduled dates. These delays can be caused by various factors, including bureaucracy, technical challenges, lack of financial resources, and insufficient technical and managerial skills, as well as the inability to anticipate and manage associated risks [11]. Numerous studies have highlighted the negative impacts that delays in cohesion projects can have on local and regional economic development. Such delays can compromise the effectiveness of projects and their ability to achieve cohesion objectives, such as job creation, increased competitiveness, and reduced regional inequalities. In less developed regions, delays in cohesion projects can particularly limit their capacity to invest in essential public infrastructure and services [12]. Therefore, it is crucial to understand the causes of delays in cohesion projects and find ways to mitigate them, as emphasized in a report by the European Court of Auditors [13] and various works [12]. Scientific literature underscores the importance of addressing project delays and promoting an evaluation culture to ensure the effectiveness and efficiency of investments. The EU's cohesion policy plays a fundamental role in promoting territorial cohesion and reducing regional disparities, but significant challenges remain in ensuring equitable and sustainable development. In this regard, delays in cohesion projects represent one of the main obstacles to achieving territorial cohesion objectives. In light of the information presented in this paper, we will focus on studying delays in Italian cohesion projects. Our approach aims to identify the extent of delays and their primary causes.

2 Dataset and methodology

For our work, we rely on the availability of data related to cohesion projects, provided by Opencoesione.it, collected in a rich dataset containing all Italian cohesion projects at the municipal granular level. We consider for our estimates only the projects belonging to the 2007-2013 programming cycle, i.e., all projects with a planned end date after January 1, 2007, and at the same time no later than December 31, 2016 (the last date set by the EU for the completion of projects funded by cohesion funds, for the 2007-2013 European programming period). We then analyze, for each cohesion project, the distribution of project delays by calculating, in days, the differences between actual end dates and planned end dates.

$$d_j = e_j - p_j$$

- d_j represents the delay associated with the j-th project;
- e_j represents the end date of the j-th project;
- p_j represents the effective end date for the j-th cohesion project.

For the sake of brevity, we then identify the average delay data d, following the formulation below:

$$\bar{d} = \frac{\sum_{j=1}^{n} d_j}{n}$$

- $\overline{d_j}$ represents the average delay for the associated with the j-th project;
- n represents the number of the j-th project.

We repeat these analyses by dividing the Italian territory into different spatial levels, namely territorial macro-areas (NUTS 1), regions (NUTS 2), provinces (NUTS 3), and municipalities. We conduct the latter analysis to support the territorial estimates made at more aggregated levels and to detect disparities within the same regional or provincial territory. These estimates make it possible to capture any spatial effects that can explain delays. These estimates report the average values of the respective delays for each territorial level. Additionally, in line with the scientific literature that emphasizes the correct planning and evaluation of cohesion projects, we study the type of implementing body in relation to the size of the delays of each project, in order to highlight whether there are temporal differences in cohesion projects between different implementing body. In this regard, we group implementing entities into two specific categories based on their sectors of origin, public sector and private sector. For these two categories, we calculate the average delays:

$$\overline{ds_j} = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i}$$

- $\overline{ds_j}$ represents the average delay for the associated with the j-th project and the i-th category of implementing body;
- d_{ij} represents the delay with the j-th project and the i-th implementing body;
- n_i represents the number of the j-th project and the i-th category of implementing.

We also study the dimension of delays with respect to the amount of cohesion funds associated with them, beyond the territorial effect. In this way, we can verify whether the size of resources assigned to each project may influence their effectiveness in terms of project timing. For computational and result synthesis reasons, we calculate the statistical quintiles of the distribution of funds and replace individual cohesion funds with them, so as to calculate the quintile averages of delays. The division into quintiles allows us to divide the distribution into five groups based on the size of funds. The first quintile represents projects with the lowest funds, while the fifth quintile represents projects with the highest funds. We then obtain five different average statistics on project delays, one for each quintile of the total distribution of cohesion funds. The procedure followed is as follows.

$$q_i = \frac{i/5}{n+1}$$

- q_i represents the value of the i-th quintile number (i = 1,2,3,4,5)

$$\overline{dq_j} = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i}$$

- $\overline{dq_j}$ represents the average delay for the associated with the j-th project and i-th quintile;
- d_{ij} represents the delay associated with the j-th project and for the i-th quintile;
- n_i represents the number of the j-th project and for the i-th quintile.

Finally, we repeat these last two analyses by dividing the Italian territory by NUTS 1 in order to evaluate delays in relation to the amount of funding and the nature of the implementing entities of cohesion projects, observing the territorial differences that characterize them. The statistics and analyses described above are reported in the tables and shown in the figures below.

3 Results and evidence

The results obtained on the delays of cohesion projects (Table and Figure 1) clearly show the strong territorial dualism between North and South, which negatively affects the homogeneous development of our country. In fact, the greatest delays are recorded in the Southern regions of Italy, Campania, Puglia, and Calabria, as well as in the Islands (over 730 days). On average, the highest delays are present in the more inland Sicilian provinces, such as Enna (almost 800 days) and Caltanissetta (737 days). The lowest delays are recorded in the Central-Northern regions such as Emilia-Romagna and Piedmont. The most performing region in terms of delays is Lombardy.

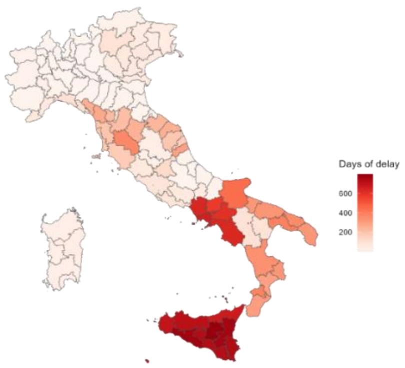
The differences among implementing entities confirm the challenges faced by local and territorial administrations, as projects managed by public entities exhibit an average delay of almost a year, while those initiated by the private sector encounter significantly shorter delays, under ten days. These differences can be attributed to the greater bureaucracy and complex procedures involved in the public sector compared to the private sector. The statistics in Table 2 demonstrate a correlation between funding and delays, indicating that projects with higher funding tend to experience more significant delays compared to those with lower funding. Larger and more complex projects, such as physical and digital infrastructures, generally require more time and resources to complete. Additionally, larger projects may necessitate greater coordination among stakeholders, leading to a higher number of issues and risks to manage. Moreover, projects with higher funding may be subject to increased scrutiny and compliance with regulations, which can impede the completion process. External factors such as market changes or unforeseen environmental and technical issues can further contribute to delays.

Table 1 Territorial statistics on delays in cohesion projects

NUTS	Observed projects	Means	Standard deviation	Median	Min	Max
Abruzzo	20676	7.23	58.67	0	0	1606
Basilicata	7990	95.92	353.54	0	0	3314
Calabria	20835	316.11	625.17	0	0	3318
Campania	25103	634.32	803.60	155	0	2826
Emilia-Romagna	12667	7.45	60.03	0	0	1888
FVG	18142	51.92	115.65	1	0	2326
Lazio	1067	80.10	271.95	0	0	2775
Liguria	10191	94.14	193.62	3	0	3660
Lombardia	404811	0.26	16.34	0	0	2715
Marche	19316	225.25	308.88	158	0	3239
Molise	2856	19.94	104.65	0	0	1453
Piedmont	30325	12.06	84.09	0	0	2754
Apulia	43306	310.70	660.34	0	0	2837
Sardinia	15481	32.54	97.63	0	0	1784
Sicily	21285	730.04	837.79	269	0	3025
Tuscany	12842	238.14	310.92	138	0	3364
Trentino-Alto Adige	3126	68.32	142.53	0	0	1735
Umbria	7753	34.06	153.67	0	0	2154
Aosta Valley	4168	15.39	91.42	0	0	1872
Veneto	4478	48.98	104.66	0	0	1099
Italy	**687175**	**92.62**	**357.12**	**0**	**0**	**3318**

Source: Authors' elaboration on Opencoesione data

Fig. 1 Spatial distribution of delays in cohesion projects



Source: Authors' elaboration on Opencoesione data

Table 2 Statistics on delays in cohesion projects by implementing bodies and quintiles of funding

Implementing body and quintili of funding	Observed projects	Means	Standard deviation	Median	Min	Max
Private sector	517309	8.87	68.0	0	0	2681
Pubblic sector	174314	341.45	638.3	15	0	3660
I°	139005	2.15	29.9	0	0	2245
2°	135898	9.51	66.4	0	0	3364
3°	137401	33.40	167.6	0	0	2973
4°	137434	121.18	423.1	0	0	3083
5°	137435	296.95	603.4	0	0	3660

Source: Authors' elaboration on Opencoesione data

4 Conclusions

The effectiveness of the European Union's cohesion policy is currently compromised by a series of factors that contribute to delays in cohesion projects implementation. Delays in cohesion projects can have negative consequences on the local economy particularly in less developed regions. It is therefore important to identify the causes of delays in cohesion projects and find ways to mitigate them. In this work, we initially focused on the descriptive study of delays in cohesion projects in Italy, through a statistical analysis aimed at identifying the extent of delays and the main causes and factors that determine such timing. In summary, our work aims to contribute to a greater understanding of delays in cohesion projects in Italy, to improve the effectiveness and efficiency of investments and promote territorial cohesion. The next step of the study will be to implement advanced statistical models, such as “machine learning” models and econometric models, to predict delays in future cohesion projects. The main objective of these models is also to offer policy insights to reduce future delays and improve investment effectiveness.

References

1. European Commission: Cohesion policy 2021-2027, Bruxelles (2021)
2. Coco, G. & Lagravinese, R.: Cohesion Policy and Public Investment in the EU, *The Great Reset*, Cerniglia, F., Saraceno, F., and Wattthe, A. (2021)
3. Bachtler, J.: EU cohesion policy in practice: what does it achieve? Rowman & Littlefield (2016)
4. Alesina, A. and Giavazzi, F.: Europe and the Euro, University of Chicago Press (2015)
5. European Commission: Cohesion in Europe towards 20250- Eight Report on Cohesion Policy, Bruxelles (2021b)
6. Rodríguez-Pose, A., and Tselios, V.: The economics of the crisis: a review of post-crisis policies in the EU. Cambridge Journal of Regions, Economy and Society, 10(1), 43-60 (2017)
7. Peragine, V.: The effects of EU regional policy: An analysis of the allocations and spending of the European Regional Development Fund and Cohesion Fund, European Journal of Political Economy, 63, 101874 (2020)
8. Svimez: L’economia e la Società del Mezzogiorno- Rapporto 2022, Il Mulino (2022)
9. Crescenzi, R., Giua, M., and Sonzogno, G. V.: Mind the Covid-19 crisis: An evidence-based implementation of Next Generation EU. Journal of Policy Modeling, 43(2), 278-297 (2021)
10. European Commission: Eighth Report on Economic, Social and Territorial Cohesion, Bruxelles (2022)
11. Cerqua, A. and Pellegrini, G.: The Effects of EU Regional Policy on the Growth of European Regions: Are We Spending Too Much? Documento di valutazione n. 11, Senato della Repubblica (2018)
12. Cheshire, P. C., and Magrini, S.: Urban growth drivers in a Europe of sticky people and implicit boundaries, Journal of Economic Geography, 19(4), 741-763 (2019)
13. European Court of Auditors: Effectiveness of EU support for public investment in member states: Cohesion Fund and European Regional Development Fund. Special report No 18/2019 (2019)

Solicited Session SS16 - *Recent advances in statistical learning and data analysis*

Session of the SIS-CLADAG organized by Domenico Vistocco and Pietro Coretto
Chair: Marcella Niglio
Discussant: Cristina Davino

1. *A Predictive Functional Principal Component Analysis of Well-Being Data* (Marcis L., Pagliarella M.C. and Salvatore R.)
2. *Detecting the partition in the extended hierarchy of a dendrogram: an application on biomedical data* (Policastro V., Palazzo L. and Vistocco D.)
3. *Concordance measure for rankings* (Bissiri P.G. and Nai Ruscone, M.)
4. *Quadratic discriminant scoring for selecting clustering solutions* (Coraggio L. and Coretto P.)

A Predictive Functional Principal Component Analysis of Well-Being Data

Una Analisi Predittiva delle Componenti Principali Funzionali dei Dati sul Benessere

Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Abstract We consider the case of the principal components of a multivariate random vector that obeys a linear mixed model. The random vector itself then lies in a lower dimensional subspace. This situation suggests that this subspace can be modeled by the probabilistic (random-effects) principal components. We employ a linear predictor adjusted by the residual part of the probabilistic principal components, that results not explained by the longitudinal time-varying linear model. A new predictor is given, employing the vector of scores that comes from that principal components, and the resulting scores approximated by spline curves as functionals. The application to the official Italian well-being data shows some of the features of the method.
Abstract *In questo articolo, consideriamo il caso delle componenti principali di un vettore casuale multivariato che segue un modello misto lineare. Il vettore casuale stesso risiede, quindi, in un sottospazio di dimensione inferiore. Questo sottospazio può essere modellato dalle componenti principali probabilistiche (ad effetti casuali). Facciamo uso di un predittore lineare corretto dalla parte residuale dei componenti principali probabilistici, che risulta non spiegato dal modello longitudinale lineare time-varying. Viene dato un nuovo predittore, utilizzando il vettore dei punteggi che proviene da quei componenti principali, e i punteggi risultanti approssimati da curve spline. L'applicazione ai dati ufficiali italiani sul benessere mostra alcune delle caratteristiche del metodo.*

Key words: multivariate random vector, probabilistic principal components, functional principal components, linear mixed model, well-being

Laura Marcis
Department of Economics and Law, University of Cassino and Southern Lazio; e-mail: laura.marcis@unicas.it

Maria Chiara Pagliarella
Department of Economics and Law, University of Cassino and Southern Lazio; e-mail: mc.pagliarella@unicas.it

Renato Salvatore
Department of Economics and Law, Univerisity of Cassino and Southern Lazio; e-mail: rsalvatore@unicas.it

1 Introduction

Principal component analysis (PCA) is recognized as one of the most employed methods to reduce dimensionality, by means of projection of a set of variables in a subspace of them. By summarizing and allowing to visualize data, and, at the same time, minimizing the loss of information in the lower dimensional space, in many cases principal components (PCs) lead to a better assessment of the bundled statistical information, seized by the original variables [2]. In the case of studying data with measurements over time of a multivariate random vector, the goal is to assume that the principal components are curves considered as functionals [4]. These functionals change with a continue variable, that is time. Then, functional principal component analysis is a standard PCA performed by sampling units curves at each time instant. The evaluation of the functionals are related to smoothing time PC points, and the resulting curves considered as the principal sources of variability. This method involves an integral transform, which reflects the functional nature of the data, whose analytic solutions cannot, in general, be determined.

Another important and recurrent way of approaching redundancy and, in general, the recursive informative content of multivariate sample data, is given by considering a common subset of covariates which the population obeys. Two main cases in the literature are deemed representative of the joint dependence on a multivariate vector, the PCs "with covariates" or Partial PCA, and the redundancy analysis. Although these last represent a very useful tool in some cases, the deployment of linear models to explain part of the sample variability has had a remarkable development in the last years. One of these studies brings into play the role of prediction by linear statistical models.

Tipping and Bishop [7] had already introduced the notion of prediction for the PCs. They called "Probabilistic PCA" (PPCA) the model behind the PCA, which parameters are estimated with the Expectation-Maximization algorithm. Instead of a "fixed effects PCs", as the traditional linear regression PCA model, the PPC are random variables. This condition suggests, on the one hand, the Bayesian approach to handle the estimates for the PPC linear model and, on the other hand, to predict PCs under its meaning within random linear models theory [5]. Given normality of the error $\varepsilon \sim N(0,\sigma^2 I)$, for a linear model $\tau = B\lambda + \varepsilon$ - in case of the vector λ random - the likelihood is based on the conditional distribution $\lambda|\tau \sim N[E(\lambda|\tau), var(\lambda|\tau)]$. Moreover, it is known [6] that $E(\lambda|\tau) = \widetilde{\lambda}$ is the "Best Prediction" (BP) estimate, with $var(\widetilde{\lambda} - \lambda) = E_\tau[var(\lambda|\tau)]$. Therefore, given a linear mixed model (LMM, [1]) for τ, with $E(\tau|\lambda) = \lambda$, the model parameters are the realizations of random variables. Thus, given the BP estimates of the PPCs λ, $\widetilde{\lambda} = E(\lambda|\tau)$, the vector $\widetilde{\tau} = B\widetilde{\lambda}$ represents the BP of the p-variate vector.

In the present paper we introduce a model of PPCs combined with a multivariate longitudinal linear predictor, considering then the random vector effectively represented by the subspace of the resulting "shifted" PPCs (SPCs). The new smoothed SPCs can be considered as time-varying functional principal components, because they combine the linear model and the PPC's, and carry out simultaneously the linear predictor and the contribution given by the PPCs not "explained" by the linear predictor itself.

An application to the official Well-being Italian indicators shows some of the features of the method.

2 Theory

In the sequel we report the following symbols, giving the model specification. We consider n as the number of subjects in the longitudinal LMM ($i = 1,...,n$), $N = n \times T$ as the total number of observations considered, with $t = 1,...,T$ the time instants. Then p ($j = 1,...,p$) is the number of the response dependent variables, l the number of the linear model covariates, and s is the dimension of the effective PC subspace. Consider Θ as the $N \times p$ sample matrix of the p-variate $p \times 1$ random vector θ, with N as the total number of the units given by the sample. Moreover, consider that the vector θ obeys the linear model:

$$\theta = \beta' x + u', \tag{1}$$

where x the $l \times 1$ vector of covariates, β the $l \times p$ matrix of the regression effects, u is the vector of the p-variate random effect, with $u \sim N(0, \Sigma_u)$, $\Sigma_u = cov(u)$. Furthermore, we consider at the same time that the multivariate random vector θ obeys the following linear model:

$$\theta = Ab + \varepsilon, \tag{2}$$

in which A is $p \times s$ a loading matrix of eigenvectors, b is the random vector of PPCs, and ε is a vector of isotropic error, with $\theta \sim N(\mu, A\Psi A' + \sigma_\varepsilon^2 I)$, $b \sim N(0, \Psi)$, $\Psi = diag(\psi_1, ..., \psi_s)$, $s < p$, and $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$. When a sample of N observations is given, an $N \times p$ matrix Y of observations from the random vector θ is simply modeled as $Y = \Theta + E$, with the "sampling error" $Np \times Np$ covariance matrix $cov(vec(E)) = (\Sigma_e) \otimes I_N$ ($\otimes$ is the Kronecker product), $e \sim N(0, \Sigma_e)$, $\Sigma_e = var(e)$. Thus, models 1 and 2 are rewritten as $Y = \Theta + E = X\beta + ZU + E$, with the 1 that becomes $\theta = \beta' x + u' + e'$, and $Y = \Theta + E = BA' + \Xi + E = BA' + \Gamma$, with the 2 that is $\theta = Ab + \varepsilon + e = Ab + \gamma$, respectively. The model errors u, ε, and e, are mutually independent. The matrix Z represents the $N \times n$ design matrix of random effects and E is the $N \times p$ matrix of the residual errors of the multivariate LMM, B is the $N \times s$ matrix of the PPCs that lie in the s-dimensional subspace, Ξ is the $N \times p$ matrix of the isotropic errors of the model 2. The models 1 and 2 have the following conditional expectations and variances:

$$E(\theta|y) = \widetilde{\theta}_y = y - E(e|y) = y - cov(e,y)var(y)^{-1}y = y - var(e)Py,$$
$$var(\theta|y) = var(\theta) - cov(\theta,y)var(y)^{-1}cov(y,\theta) = var(e) - var(e)Pvar(e),$$

for the model in 1, where $P = \Sigma_y^{-1}(I - P_X)$, $\Sigma_y = var(y)$, and P_X is the projection matrix. For the model in 2:

$$\begin{aligned}
E(b|\theta) &= E(b) + cov(b,\theta)var(\theta)^{-1}(\theta-\mu)\\
&= cov(b,\mu+Ab+\varepsilon)C^{-1}(\theta-\mu) = \Psi A'C^{-1}(\theta-\mu),\\
var(b|\theta) &= var(b) - cov(b,\theta)var(\theta)^{-1}[cov(b,\theta)]'\\
&= \Psi - \Psi A'C^{-1}A\Psi, \quad C = A\Psi A' + \sigma_\varepsilon^2 I.
\end{aligned}$$

Based on some results on linear projections, i.e., given the random variable y, and the $1\times j, 1\times k$ random vectors x,z, with positive definite covariance matrix of $(y,x,z)'$, then we get for the linear projection $L(y|x,z) = L(y|x) + [z - L(z|x)]\gamma$, where $\gamma = var(z|x)^{-1}[cov(y,z|x)]'$. We have the following:

Proposition 1. *Given the model 2 for the p-dimensional random vector θ, with $b = \overline{F}'(\theta-\varepsilon)$, and under the models in 1 and 2, the multivariate Best Predictor based on (y,b), $E(\theta|y,b)$, is:*

$$E(\theta|y,b) = \widetilde{\theta}_{y,b} = E(\theta|y) + cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \tag{3}$$

with $\widetilde{b} = E(b|\theta)$, $\overline{F}$ the $sN\times pN$ matrix $(\overline{A}'\overline{A})^{-1}\overline{A}$, and $\overline{A}$ is the $pN\times sN$ matrix $A\otimes I_N$. Then, $var(\theta|y,b) = var(\theta|y) - cov(\theta,b|y)var(b|y)^{-1}[cov(\theta,b|y)]'$.

The predictor $\widetilde{\theta}_{y,b}$ in 3 gives the Best Linear Unbiased Predictor $E(\theta|y)$, “embedding” the PPCs through an adjoint component. The last due to knowing that the random vector θ lies in the s-dimensional subspace of the PPCs. In particular, the difference $\widetilde{b} - E(b|y)$ gives the multivariate vector of the PPCs “not explained” by the estimation of the linear model $E(\theta|y)$. The matrix $var(\theta|y,b)$ has rank s, and, consequently, there are $(p-s)$ linear combinations of θ for which their respective variances do not depend on the PPCs.

Proposition 2. *Given the p-dimensional random vector θ,and under the models in (1) and (2), and the Best Predictor $E(\theta|y,b)$ in 3, we get:*

$$\widetilde{b}^* = \overline{F}E(\theta|y) + \overline{F}cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \tag{4}$$

where $\widetilde{b}^* = \overline{F}E(\theta|y,b) = \overline{F}\widetilde{\theta}_{y,b}$ is the s-dimensional vector of the PPCs “shifted” (SPCs) by the linear predictor $E(\theta|y)$. As a particular case, when $\sigma_\varepsilon^2 \longrightarrow 0$, $var(\varepsilon) \longrightarrow 0$, $var(\gamma) \longrightarrow var(e)$, and $\widetilde{b} \longrightarrow \overline{b}$. Therefore, $\overline{F}\widetilde{\theta}_{y,b} = \widetilde{b}^* \longrightarrow \overline{b}$, with $\overline{B} = vec^{-1}(\overline{b}) = ((vecI_s)'\otimes I_N)(I_s\otimes\overline{b})$ the matrix of the PC scores by sample PCs $\overline{b} = A'\theta$. The matrix $\widetilde{B}^* = vec^{-1}(\widetilde{b}^*) = ((vecI_s)'\otimes I_N)(I_s\otimes\widetilde{b}^*)$ can be written as $\underset{1\le i\le n}{col}(\widetilde{b}_i^*(t))$, with $\widetilde{b}_i^*(t) = \underset{1\le t\le T}{col}(\widetilde{b}_{it}^*)$. Thus, the SBCs are the PPCs with the scores predicted by the linear model (1), with subjects as time repeated observations, and $\widetilde{b}_s^*(t)$ expressed in terms of spline basis.

The SPCs $\widetilde{b}^*$ are then the PPCs “adjusted” by the Linear BP $E(\theta|y)$, and $\widetilde{b}^*$ is then the vector obtained by stacking columns of the matrix $\widetilde{B}^*$ of the SPC scores. Given $\sigma_\varepsilon^2 = 0$, the vector θ in the model 2 lies in the s-dimensional subspace of the standard sample PCs $\overline{b}$. In fact, in this case $\widetilde{b} \equiv \overline{b}$, $cov(\theta, b|y) = var(\theta|y)\overline{F}'$, $var(b|y) = \overline{F}var(\theta|y)\overline{F}'$, $E(b|y) = \overline{F}E(\theta|y)$, and then $\overline{F}E(\theta|y,b) = \overline{b}$.

3 Application and Discussion

The data for the application are the Equitable and Sustainable Well-being indicators (in Italian, BES) - annually provided by the Italian Statistical Institute ([3], 2017). In order to highlight the result of the proposed method we use 12 BES indicators relating to the years 2013-2016, collected at NUTS-2 (Nomenclature of Territorial Units for Statistics -2) level. Please, refer to Table 1 for the description and acronyms used for the variables. We use LBE1 as a unique covariate in the LMM model, while the remaining 11 variables are dependent variables. The application uses the REML estimation, a Sas/IML code together with a sequence of Sas-HPMixed procedures. We propose to estimate the model under a uniform correlation structure among the multivariate components of the random effects, and an AR(1) within-subject errors. The uniform correlation between the multivariate components of the random effects is equivalent to compound-symmetry covariance, with a better numerical property in terms of optimization. Further, some studies highlight that using uniform correlation matrices reduces the estimation noise. The slope parameter estimates from the multivariate regression are all significant. The MANOVA multivariate test statistics (Wilks Lambda, Pillai's Trace, Hotelling-Lawley Trace, Roy's Largest Root), based on the characteristic roots are all significant. Figure 1 shows the application of the model described in (3) and (4). The plots represent the spline curves of PPCs and SPCs, for each of the subjects (the 20 Italian administrative Regions). The PPC1 is mainly related with the variables INN1, BS3, AMB9, POL5, S8, while the PPC2 is related with REL4, Q2, L12. The smoothing operated by the linear model part of the SPCs is highlighted especially comparing PPC2 and SPC2, for which the SPC2 shows a less confusing behavior when varying the year. The resulting splines indicate that the prediction stabilizes a lot the scores of the principal components. This is because the SPCs use PPCs, retaining all the main features of the principal components analysis approach, and utilizing at the same time all the capabilities of a linear predictor.

The SPCs in 4 can be viewed from two different perspectives. An "adjusted" linear predictor by the PPCs, by relation (3), as the shifted PCs (SPCs) that modify the probabilistic PCs (PPCs) to accommodate the mixed model regression predicted values. The present work considers then the probabilistic principal components as a "constraint" model, that links together the components of the multivariate random vector in a lower dimensional subspace.

The SPCs are generated considering the multivariate linear predictor 1, by adjusting its standard formulation through the sample parameter vector scores in a convenient subspace. The main advantage of the method proposed is that allows using all the features that a linear model can contribute to improving the principal component data analysis.

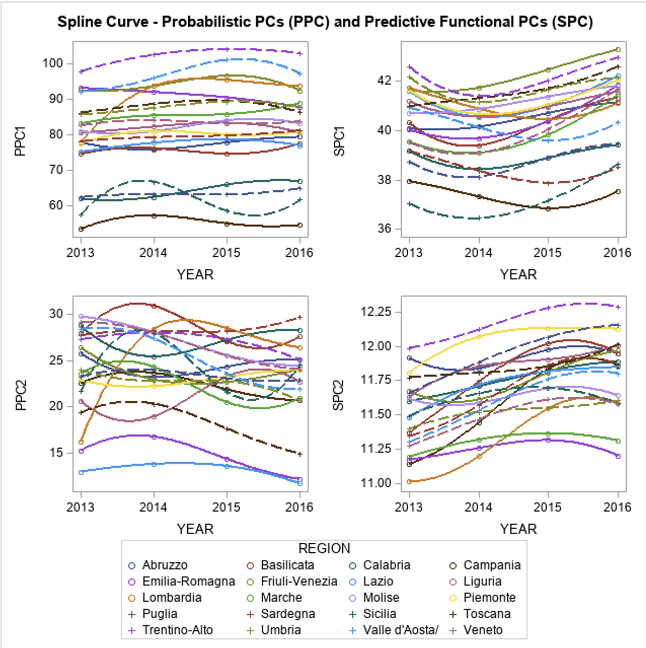


Fig. 1 Joint spline curves of the first two PPCs and SPCs, for the four years (2013-2016) and the twenty Italian administrative Regions. The official Italian Well-Being data are reported and downloadable by the Italian National Statistical Institute web site.

Table 1 Description of the variables used for the application

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgment for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

References

1. Demidenko, E.: Mixed Models: Theory and Applications. Wiley, New York (2004)
2. Hardle, W. K., Simar, L.: Principal Components Analysis. In Hardle, W. K., Simar, L. (eds.) Applied Multivariate Statistical Analysis, 319-358. Springer (2015)
3. ISTAT: BES project www.istat.it/en/well-being-and-sustainability
4. Jolliffe Ian T., Cadima J.: Principal component analysis: a review and recent developments Phil. Trans. R. Soc. A. 374: 20150202.20150202 (2016)
5. Longford N.T.: Random Coefficient Models. In: Lovric M. (eds.) International Encyclopedia of Statistical Science. Springer, Heidelberg (2011)
6. Timm, N. H.: Applied Multivariate Analysis. Springer, New York (2002)
7. Tipping, M.E., Bishop C.M.: Probabilistic principal component analysis. J. R. Stat. Soc., Ser. B (Stat. Methodol.) 61(3), 611–622 (1999)

Detecting the partition in the extended hierarchy of a dendrogram: an application on biomedical data

Rilevamento della partizione nella gerarchia estesa di un dendrogramma: un'applicazione su dati biomedici

Valeria Policastro, Lucio Palazzo and Domenico Vistocco

Abstract This paper shows DESPOTA in action on biomedical data. DESPOTA is an algorithm able to detect the partition through the inspection of the extended hierarchy embedded in a dendrogram. The algorithm exploits permutation tests, and automatically returns a partition of the input data, without requiring to specify the number of clusters to use. The algorithm works with any distance method and agglomeration method used to grow up the dendrogram. Moreover, the provided clusters do not necessarily correspond to a horizontal cut of the tree. We exploited DESPOTA on single-cell transcriptomic data, to the end of clustering cells based on their expression profile. It is one of the challenging fields of applications for the identification of clusters nowadays, given the particular data structures involved.

Abstract *Questo articolo mostra DESPOTA all'opera su dati biomedici. DESPOTA è un algoritmo in grado di rilevare la partizione attraverso l'ispezione della gerarchia estesa incorporata in un dendrogramma. L'algoritmo sfrutta i test di permutazione e restituisce automaticamente una partizione dei dati in ingresso, senza richiedere di specificare il numero di cluster da utilizzare. L'algoritmo funziona con qualsiasi tipo di distanza e metodo di agglomerazione utilizzato per far crescere il dendrogramma. Inoltre, i cluster forniti non corrispondono necessariamente a un taglio orizzontale dell'albero. Abbiamo applicato DESPOTA sui dati trascrittomici a singola cellula, allo scopo di raggruppare le cellule in base al loro profilo di espressione. Si tratta di uno dei campi di applicazione più impegnativi per l'identificazione dei cluster al giorno d'oggi, data la particolare struttura dei dati coinvolti.*

Valeria Policastro
Department of Political Sciences, University of Naples Federico II, e-mail: valeria.policastro@gmail.com

Lucio Palazzo
Department of Political Sciences, University of Naples Federico II, e-mail: lucio.palazzo@unina.it

Domenico Vistocco
Department of Political Sciences, University of Naples Federico II, e-mail: domenico.vistocco@unina.it

Key words: clustering, single-cell, cell clustering, transcriptomic data

1 Introduction

Clustering refers to a family of statistical models and algorithms aiming to partition a set of unlabeled data points into different groups so that subjects with similar traits will be assigned to the same group. In Biology, clustering is used to gain insights into biological processes, i.e. to investigate the natural structure present in the data, and in particular on gene functions, cellular processes, subtypes of cells, and gene regulations. Among the various problems of interest in the biomedical field, single-cell RNA-sequencing technology is growing lately. Despite the development of scRNA–seq, a promising method for studying cell heterogeneity, the analysis of this type of data remains a challenge. Several studies in literature exploit clustering on single-cell transcriptomic data to detect cell sub-populations based on their similarity in the gene expression profiles. Such studies differ in the type of clustering method used. Among others, it is worthwhile to mention the use of *k*–means (Identification of cell types from single-cell data using stable clustering [7]), hierarchical clustering (SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis [4], CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data [5]), and graph-based clustering (Seurat, Comprehensive Integration of Single-Cell Data [6]).

Hierarchical clustering is a valuable solution for clustering data since it provides nested partitions that can be easily visualized through a dendrogram, a tree-like representation. The selection of a unique partition is usually accomplished by cutting the tree through an horizontal line such that the resulting connected branches of the tree define the elements of the partition. Such an approach is based on heuristic criteria, and it is not appropriate for detecting not well-separated uniform clusters. Several criteria have been proposed in the literature to select the ”best” partition by optimizing some cluster validity indexes. Among these, here we resort to DESPOTA, DEndrogram Slicing through a PermutatiOn Test Approach[1]. It inspects the extended hierarchy provided by an hierarchical clustering, including in the search space also partitions not corresponding to horizontal cuts of the dendrogram. Therefore, the final partition could be composed of tree branches corresponding to different heterogeneity levels. Moreover, DESPOTA does not require the user to specify the number of clusters, but it automatically detects the proper partition starting from the data. The algorithm is based on the use of permutation tests at each node of the tree, the tests aiming to verify if the two descending branches belong to only one cluster (null) or to two different clusters (alternative). DESPOTA does not require any distributional assumption and works in a purely data-driven approach. The effectiveness of such an approach is shown starting from Pollen data [3], a low-coverage single-cell transcriptomic dataset of cell populations from heterogeneous tissues.

The paper is organized as follows: Section 2 briefly describes DESPOTA, the underlying idea, the algorithm, and the test statistic, while Section 3 introduces the Pollen data and the main results provided by DESPOTA on such data. Section 4 ends the paper with some concluding remarks.

2 Materials and Methods

DESPOTA retraces down the tree structure of a dendogram following the path inverse to that used for its construction, that is starting from the root node to the leaves. The main procedure is sketched out in Algorithm 1, the essential notation being depicted on an example dendrogram in Figure 1. We denote with $k = 1, \ldots, N-1$ the level at which two generic clusters are merged during the hierarchical agglomerative process, where k = 1 refers to the uppermost link of the dendrogram. The left and right clusters, according to the dendrogram visualization, joined at level k, are denoted with L_k and R_k, respectively, while $h(.)$ refers to the metric used in the hierarchical classification process.

Input: A dataset and its related dendrogram
Output: A partition of the dataset

initialization:
aggregationLevelsToVisit ← $h(L_1 \cup R_1)$
detectedClusters ← [ ]
k ← 1
repeat
  if H_0 *is accepted (the two clusters represent a unique class)* **then**
    ▪ add $L_k \cup R_k$ to detectedClusters
  else
    ▪ add $h(L_k)$ and $h(R_k)$ to aggregationLevelsToVisit
    ▪ sort aggregationLevelsToVisit in descending order according to $h(.)$
  end
  remove the first element from aggregationLevelsToVisit
  k ← k+1
until *aggregationLevelsToVisit is empty*

Algorithm 1: The DESPOTA algorithm.

Starting from the root of the dendrogram, for each level k, a permutation test is carried out to test the Null Hypothesis that the two classes L_k and R_k really belong to the same group C_k. Under H_0, mixing up (permuting) the statistical units of L_k and R_k should not alter the aggregation process resulting in their merging in. The following test statistic:

$$rc_k = \frac{\mid h(L_k) - h(R_k) \mid}{h(L_k \cup R_k) - min[h(L_k), h(R_k)]}$$

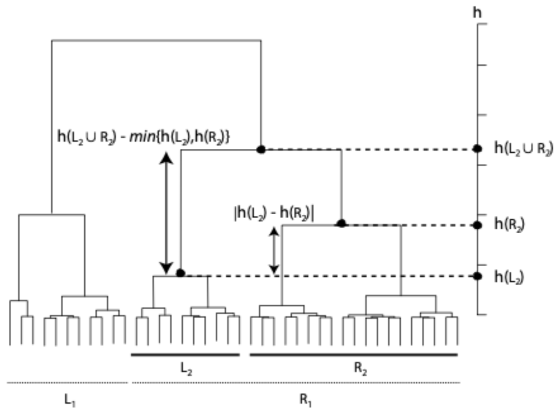


Fig. 1 The main notation used in the DESPOTA algorithm and in formulation of the test statistic.

normalized in $[0,1]$, combines the minimum cost required for joining L_k with R_k (numerator), through the actual cost incurred in the merging process (denominator). In case $rc_k \approx 0$, the cost required to merge the two clusters is very much larger than the minimum cost, and therefore it is reasonable to decide in favour of the presence of two different clusters. At the contrary, when $rc_k \approx 1$, the cost required to merge the two clusters does not differ greatly from the minimum necessary cost, and hence the two groups should be considered as a unique cluster. The Monte Carlo p-value is computed as the proportion of times the observed value rc_k is greater than or equal to the permuted values.

The interested reader is referred to [1] for details.

3 Application

The Pollen dataset [3] contains the gene expression data of 301 cells from 11 subpopulations. The cells are analyzed with single-cell transcriptomic technology. The dataset provides also two known labels for each cell, i.e. the labels at the tissue level, and at the cell line level. The usual pre-processing of this type of data concerns different steps such as normalization, feature selection, and dimension reduction, one of the most used pipelines for this kind of analysis is the one implemented in `Seurat` [6]. We carried out a specific pre-processing of the data, exploiting the `Seurat` package [6]. In particular, we exploited a commonly adopted pipeline for this type of data, consisting of normalization, feature selection, and dimension reduction. We considered the first 25 principal components, computed using the `HGC` package [8], and we used such components to obtain the shared nearest neighbor (SNN) matrix. Subsequently, the resulting data were used for computing the distance matrix, which is the input of the DESPOTA algorithm.

DESPOTA provided the partition in 4 clusters depicted in Figure 2. It is worth highlighting that it is not possible to get such a partition using the traditional horizontal cut approach. Figure 3 compares the detected partition (right side of the plot)

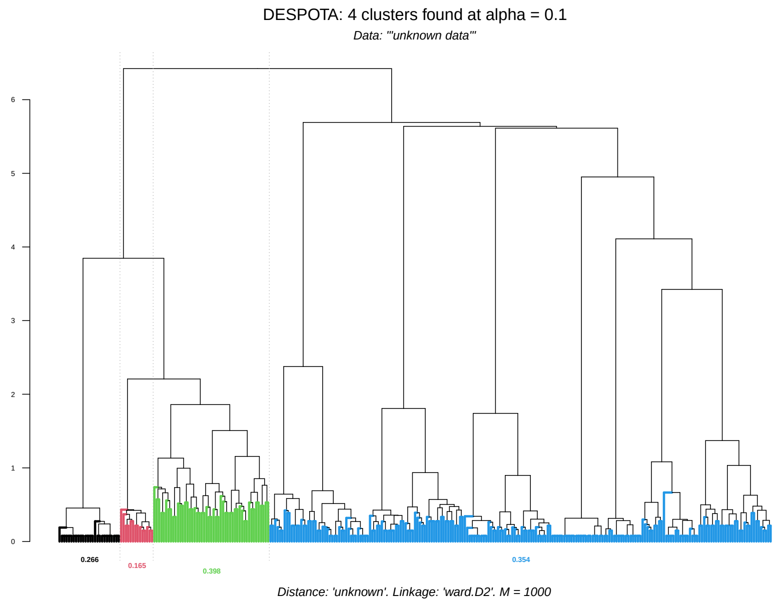


Fig. 2 The dendrogram obtained on the Pollen dataset, and the partition in 4 clusters detected using DESPOTA.

with the known labels at the cell line level (left side of the plot). The Sankey diagram makes it possible to visualize how the different cell line levels flow in the final partition. In particular, cluster 1 gathers all the cells belonging to cell line 2338, cluster 2 collects the cells of cell line 2339, cluster 3 mainly the BJ cell line, and cluster 4 contains all the remaining cell lines. Nonetheless, inspecting the labels at the tissue level of the cells of the fourth class, it appears that GW16, NPC, GW21, and GW21+3 are all neural cells, and K562 and HL60 are both blood cells.

4 Concluding Remarks

The application presented in the previous section showed that DESPOTA is a valuable method for automatically detecting a partition starting from a hierarchical clustering of single-cell transcriptomic data. The algorithm inspects the whole set of partitions hosted in a dendrogram without requiring the number of desired clusters as input and can compose a partition by drawing groups from partitions correspond-

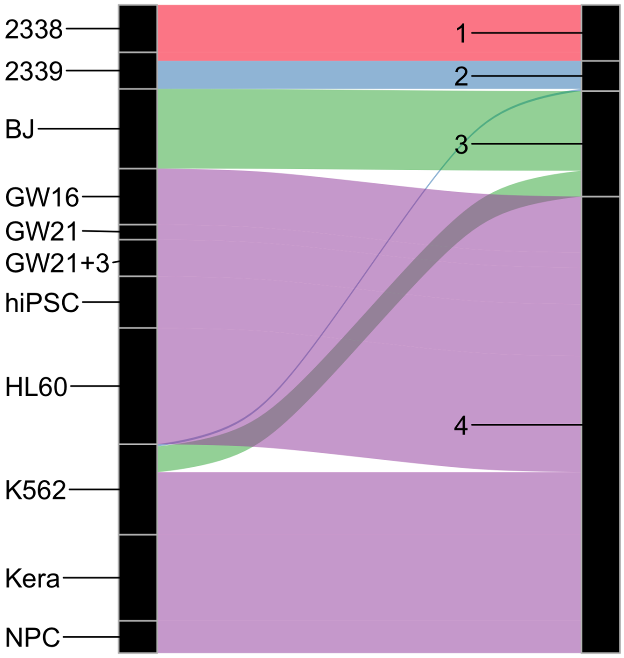


Fig. 3 Comparisons between the clusters obtained with DESPOTA and the cell line labels. The yellow rectangle on the vertical axis includes the neural cell.

ing to different heterogeneity levels. This is the case of the four clusters detected on the Pollen dataset [3]: such a partition does not correspond to a horizontal cut of the dendrogram. Also, the interpretation of the 4 groups is reasonable, as well as the classical quality measures show satisfactory values.

An issue worth investigating concerns the parallel implementation of the algorithm, of particular relevance for the efficient use of cell data.

References

1. Bruzzese, D., Vistocco, D.: DESPOTA: DEndrogram slicing through a pemutation test approach. Journal of classification 32, 285-304 (2015)
2. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification 2, 193-218 (1985)
3. Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang C.: Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature biotechnology 32(10), 1053-1058 (2014)
4. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A., Xu, Y.: SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS computational biology 11(11), e1004575 (2015)
5. Lin, Peijie and Troup, Michael and Ho, Joshua WK: CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome biology. BioMed Central. 18, n1 1-11 (2017)
6. Stuart, Tim and Butler, Andrew and Hoffman, Paul and Hafemeister, Christoph and Papalexi, Efthymia and Mauck III, William M and Hao, Yuhan and Stoeckius, Marlon and Smibert, Peter and Satija, Rahul: Comprehensive integration of single-cell data. Cell. Elsevier. 177, n7 1888-1902 (2019)
7. Peyvandipour, A., Shafi, A., Saberian, N. et al. Identification of cell types from single cell data using stable clustering. Sci Rep 10, 12349 (2020). https://doi.org/10.1038/s41598-020-66848-3
8. Zou, Ziheng and Hua, Kui and Zhang, Xuegong. HGC: fast hierarchical clustering for large-scale single-cell data. Bioinformatics 37, n21 3964-3965 (2021)

Concordance measure for rankings

Misura di concordanza per ranking

Pier Giovanni Bissiri and Marta Nai Ruscone

Abstract We define a new concordance measure for ranking data taking into account the discrete nature of the variables. This concordance measure is used to evaluate the dissimilarity between subjects expressing their preferences by rankings. The proposed distance builds upon the Spearman's correlation coefficient as a measure of concordance, based on copula, of the rank denoting the level of the importance assigned by subjects under classification to k objects.
Abstract *Definiamo una nuova misura di concordanza per dati rank tenendo conto della loro natura discreta. Questa misura di concordanza viene utilizzata per valutare la dissimilarità tra i soggetti che esprimono le loro preferenze su alcuni oggetti. La distanza proposta si basa sul coefficiente di correlazione di Spearman come misura della concordanza, basata sulla copula, del rango che denota il livello dell'importanza assegnata dai soggetti nel classificare k oggetti.*

Key words: Concordance Measure, ranking, copula

1 Introduction

Ranking data occur when a number of subjects are asked to rank a list of objects (items) according to their personal preference order. A distance measure is useful in measuring the discrepancy between two rankings.

The issue when dealing with ordinal data lies in computing an appropriate distance matrix. Several distance measures have been proposed for ranking data [1].

Pier Giovanni Bissiri
University of Bologna, Department of Statistical Sciences 'Paolo Fortunati', Via Belle Arti 41, Bologna, e-mail: piergiovanni.bissiri@unibo.it

Marta Nai Ruscone
University of Genova, Department of Economics, Via Vivaldi, 5, Genova, e-mail: marta.nairuscone@unige.it

The most important are referred to Kendall's s, Spearman's q and Cayley distances [2, 3, 4, 5].

In this work we propose a generalization of this kind of distances using concordance measure. In this way we have a more flexible instrument to model different types of data dependence structures and to consider the discrete nature of the data.

2 Spearman's correlation coefficient as a measure of concordance

Concordance between X and Y is the tendency of one random variable to take large values when the other random variable takes large values, too, whereas discordance is the tendency of one random variable to take large values when the other random variable takes small values. Let us make the concept precise. Given two $\mathbb{R}^2$-valued random vectors (X_1,Y_1) and (X_2,Y_2) with identical marginal distributions, (X_2,Y_2) is said to be *more concordant* than (X_1,Y_1), denoted as $(X_1,Y_1) \prec_c (X_2,Y_2)$, if

$$P(X_1 \leq s, Y_1 \leq t) \leq P(X_2 \leq s, Y_2 \leq t)$$

holds for every pair $(s,t) \in \mathbb{R}^2$.

Let F_{XY} be the joint distribution function of X and Y (namely $F_{XY}(x,y) = P(X \leq x, Y \leq y)$, $(x,y) \in \mathbb{R}^2$), and let F_X and F_Y be the marginal distribution functions of X and Y, respectively (namely $F_X(x) = P(X \leq x)$, $x \in \mathbb{R}$, and $F_Y(y) = P(Y \leq x)$, $y \in \mathbb{R}$).

There is comonotonicity if and only if the random pair(X,Y) is distributed as $(F_X^{-1}(U), F_Y^{-1}(U))$, being U uniformly distributed on $[0,1]$. In the particular case of $F_X = F_Y$, there is comonotonicity if and only if $X = Y$, almost surely.

For more details about comonotonicity see [6] and citations provided therein.

A well known measure of concordance is the Spearman's rho correlation coefficient. Given a random pair (X,Y), its Spearman's correlation coefficient ρ is equal to

$$\begin{aligned} \rho = 3\int_{R^2} \{ & P(X \leq x, Y \leq y) + P(X < x, Y \leq y) + \\ & P(X \leq x, Y < y) + P(X < x, Y < y)\} \, \mathrm{dF_X(x)} \, \mathrm{dF_Y(y)} - 3 \end{aligned} \tag{1}$$

The Spearman's correlation coefficient takes values into the interval $[-1,1]$, but it is known that the bounds -1 and 1 are not reached in the discrete case even in case of comonotonicity and countermonotonicity, whereas can be reached in the continuous case (see [7], [8]). Moreover, in the continuous case, it is possible to consider the concept of copula that is useful for evaluating concordance.

Given a random pair (X,Y) where X and Y are continuous, the copula C is the joint distribution function of the random pair $(F_X(X), F_Y(Y))$, whose marginal distributions are known to be uniform on the unit interval $[0,1]$, namely:

$$C(u,v) = P(F_X(X) \leq u, F_Y(Y) \leq v), \qquad u,v \in [0,1].$$

As proved by [9], C is the unique function from $[0,1]^2$ into itself such that:

$$C(u,v) = F_{XY}(F_X^{-1}(u), F_Y^{-1}(v)), \quad u,v \in [0,1] \tag{2}$$

and also the unique function such that:

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)), \quad x,y \in \mathbb{R}. \tag{3}$$

3 Concordance measure in the discrete case

Let us consider a particular scenario, where X and Y are two real valued random variables such that

$$P((X,Y) \in \mathbb{N} \times \mathbb{N}) = 1,$$

where $\mathbb{N} = \{1,2,3,\ldots\}$.

The two random variables X and Y are discrete and therefore an unique copula does not exists for (X,Y), but there exists uniquely a semicopula, that is a function C from
$\{F_X(i) : i \in \mathbb{N}_{k_1}\} \times \{F_Y(j) : j \in \mathbb{N}_{k_2}\}$ into $[0,1]$
such that

$$F_{X,Y}(i,j) = C(F_X(i), F_Y(j)) \ \text{ for } i \in \mathbb{N}_{k_1},\ j \in \mathbb{N}_{k_2}$$
$$C(u,v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)) \text{ for every} u,v \in [0,1].$$

Dealing with continuous random variables (i.e. random variables with a continuous distribution function) would allow us to define the copula uniquely. Therefore, we consider the following pair of continuous random variables, which is a continuous extension of (X,Y):

$$X^* = X + U - 1 \qquad Y^* = Y + V - 1 \tag{4}$$

where U and V are two continuous random variables valued in the unit interval $(0,1)$ with strictly increasing distribution functions F_U and F_V and such that (X,Y) and (U,V) are independent. A natural choice for the marginal law of U and V is the uniform distribution on $(0,1)$, but no result in this work will depend on the marginal distributions of U and V.

As a concordance measure for (X,Y) we shall consider the population Spearman's correlation coefficient between X^* and Y^*. An expression for such index is given by [10].

Theorem 1. *Let (X,Y) be a random pair valued into $\mathbb{N} \times \mathbb{N}$, and let s be the Spearman's correlation coefficient between U and V.*

Then the Spearman's correlation coefficient between $X^* = X + U - 1$ and $Y^* = X + V - 1$ is equal to:

$$\begin{aligned}\rho_s^* = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \Big\{ &(s+3)(F_{XY}(i,j) + F_{XY}(i-1,j-1)) \\ &+ (3-s)(F_{XY}(i,j-1) + F_{XY}(i-1,j)) \Big\} p_X(i)\, p_Y(j) - 3,\end{aligned} \tag{5}$$

Moreover, if ρ is the Sperman's correlation coefficient between X and Y then:

$$\rho_s^* = \rho + s \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} p_{XY}(i,j)\, p_X(i)\, p_Y(j). \tag{6}$$

The coefficient ρ_s^* satisfies the following properties:

1. $\rho_s^* = 1$ if and only if $s = 1$ and Y is a strictly increasing function of X;
2. $\rho_s^* = 0$ if $s = 0$ and X and Y are independent;
3. $\rho_s^* = -1$ if and only if $s = -1$ and Y is a strictly decreasing function of X.

4 Concordance Measure for rankings

Let us assume that two subjects are asked to rank a list of k items according to their personal preferences. Let x_l be the rank that subject one assigns to the l–th item and let y_l the rank assigned by subject two to the same item, $l = 1, \ldots, k$. We assume that ties and missing values are not allowed. So, two rankings $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $\boldsymbol{y} = (y_1, \ldots, y_k)$ are considered, being $\boldsymbol{x}$ and $\boldsymbol{y}$ permutations of $(1, \ldots, k)$. Our aim is to evaluate how the two rankings are different, but giving a higher weight to high ranked items and lower weight to low ranked items. Let such weights be $w_1, \ldots, w_k$ where $w_l > 0$ for $l = 1, \ldots, k$. Assume that such weights sum up to one (if not, replace each weight w_l with $w_l / \sum_{h=1}^k w_h$). We shall construct two random variables X and Y such that evaluate their concordance will be equivalent to evaluate how the two rankings differ according to the weights $\boldsymbol{w} = (w_1, \ldots, w_k)$.

Consider $k = 3$ items, and let the rankings of the two subjects be $2, 1, 3$ and $1, 2, 3$, respectively. Then the joint probability mass function of the random pair (X, Y) is given by Table 4.

			y	
	X \ **Y**	**1**	**2**	**3**
	1	0	$(w_1+w_2)/2$	0
x	**2**	$(w_1+w_2)/2$	0	0
	3	0	0	w_3

Table 1 Example: joint probability mass function of (X, Y).

We are interested to evaluate how X and Y are different from each other. For this reason, we pick $s = 1$ because our aim is to evaluate how X and Y are different from each other.

Therefore we define a dissimilarity measure between the two rankings given by:

$$d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \sqrt{\frac{1 - \rho_1^*}{2}}, \tag{7}$$

where ρ_1^* is the concordance coefficient between X and Y defined in the previous section with $s = 1$.

Setting $s = 1$, it is guaranteed that:

- $d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$;
- $d(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = d(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{w})$ for every two rankings $\boldsymbol{x}, \boldsymbol{y}$.

5 Discussion

In this work we generalize the approach by Denuit and Lambert [11] who consider U and V being independent. The case $U = V$ is particularly useful to obtain the coefficient ρ_1^* which is equal to one when there is a one-to-one increasing relationship between X and Y. This allows us to define a novel dissimilarity measure. In the future works we will extent the dissimilarity measure in order to consider missing and ties as well.

References

1. Alvo, M., Yu, P.L.H.: Statistical methods for ranking data. New York, Springer (1992)
2. Mallows, C.L.: Non-null ranking models. Biometrika, 44(1/2), 114–130 (1957)
3. Critchlow, D.E., Fligner, M.A.,Verducci, J.S.: Probability models on rankings. Journal of mathematical psychology, 35(3), 294-318 (1991)
4. Diaconis, P.: Group representations in probability and statistics. Lecture notes-monograph series, i-192 (1988)
5. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. 15, 72-101 (1904)
6. Puccetti, G. and Scarsini, M.: Multivariate comonotonicity. Journal of Multivariate Analysis, 101(1), 291-304 (2010)
7. Genest, C. and Neslehova, J.: A primer on copulas for count data. ASTIN Bulletin: The Journal of the IAA, 37(2), 475-515 (2007)
8. Neslehova, J.: On rank correlation measures for non-continuous random variables. Journal of Multivariate Analysis, 98(3), 544-567 (2007)
9. Sklar, M.: Fonctions de repartition an dimensions et leurs marges. Publ. inst. statist. univ. Paris, 8, 229–231 (1959)
10. Mesfioui, M. and Tajar, A.: On the properties of some nonparametric concordance measures in the discrete case. Nonparametric Statistics, 17(5), 541-554 (2005)
11. Denuit, M. and Lambert, P.: Constraints on concordance measures in bivariate discrete data. Journal of Multivariate Analysis, 93(1), 40-57 (2005)

Quadratic discriminant scoring for selecting clustering solutions

Funzione discriminante quadratica per la selezione di soluzioni di clustering

Luca Coraggio and Pietro Coretto

Abstract Selecting an optimal clustering solutions is a difficult problem. There exist many data-driven validation strategies in the literature to perform this task. In this paper, we focus on a recent proposal, based on quadratic discriminant scores and bootstrap resampling, namely the BQH and BQS from Coraggio and Coretto [4]. These strategies proved to be extremely successful with elliptic-symmetric clusters and, in general, when clusters can be separated by quadratic boundaries. In this work, we review the BQH and BQS strategies, and try to shed more light on their functioning, by comparing them with alternative likelihood-based validation indexes, and with different resampling schemes.
Abstract *La selezione di soluzioni di clustering ottimali è un problema complesso. In letteratura, esistono molte strategie di validazione per svolgere questo compito. In questo lavoro, ci concentriamo su una proposta recente, basata su funzioni di discriminante quadratica e tecniche di ricampionamento bootstrap, ovvero i metodi BQH e BQS proposti in Coraggio and Coretto [4]. Queste strategie si sono dimostrate estremamente efficaci con cluster ellittico-simmetrici e, più in generale, quando i cluster possono essere separati da confini quadratici. In questo lavoro, riprendiamo le strategie BQH e BQS, e indaghiamo il loro funzionamento in uno studio comparativo, utilizzando nuovi indici di validazione, basati sulla funzione di verosimiglianza, assieme a schemi di ricampionamento alternativi.*

Key words: Cluster validation, Mixture models, Model-based clustering, Resampling methods

---

Luca Coraggio
Department of Economics and Statistics, University of Naples Federico II, e-mail: luca.coraggio@unina.it

Pietro Coretto
Department of Economics and Statistics, University of Salerno e-mail: pcoretto@unisa.it

1 Introduction

Typically, in cluster analysis, the researcher produces many solutions, running several clustering algorithms with various settings. The problem is that, while a single final solution may be required, it may well be the case that multiple of them provide a good description of the data, according to different clusters' concepts [17]. Recently, in [4], we proposed a novel validation index aimed at selecting clustering solutions in cases where clusters can be expected to have elliptic-symmetric shapes, or to be separable by quadratic boundaries. The proposed strategy proved to be effective in conjunction with a bootstrap resampling strategy that hedges against overoptimism in the selection process, and helps avoid choosing overly-complex solutions. In this paper, we review this methodology and investigate the different components that make it successful. We do this by comparing it with an alternative likelihood-based validation index, and combining it with different resampling schemes. The outline of the paper is as follows. Section 2 reviews the BQH and BQS indexes, and introduces the alternative strategies; Section 3 presents the empirical analysis; Section 4 concludes.

2 Quadratic Discriminant Scoring

Let $\mathbb{X}_n$ be an observed sample of size n, with feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$. Let $\mathscr{G}^{(m)} = \left\{G_k^{(m)},\ k = 1,\ldots,K_m\right\}$ be a clustering solution, allocating the n objects into K_m groups; $\mathscr{G}^{(m)}$ is obtained from a clustering method $m \in \mathscr{M}$.We assume that $\mathscr{G}^{(m)}$ can be meaningfully described by a collection of K_m triplets $\boldsymbol{\theta}^{(m)} = \left\{\boldsymbol{\theta}_k^{(m)},\ k = 1,\ldots,K_m\right\}$, where each $\boldsymbol{\theta}_k^{(m)}$ collects unique elements of the following objects: *(i)* π_k: the expected fraction of points belonging to the *k-th* group; *(ii)* $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the *k-th* cluster's center; *(iii)* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is a positive definite scatter matrix that either coincides with or is proportional to the *k-th* cluster's covariance matrix. $\boldsymbol{\theta}^{(m)}$, together with K_m, provides a description of the *m-th* clustering configuration $m \in \mathscr{M}$.[1]

For a point $\boldsymbol{x}$ and a triplet $\boldsymbol{\theta}_k$, we define the quadratic scoring of point $\boldsymbol{x}$ for the *k-th* cluster as

$$\mathrm{qs}(x,\boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \tag{1}$$

The quadratic score is inspired to Quadratic Discriminant Analysis, where (1) represents the optimal classification boundaries under the Gaussian assumption (e.g., see [10]). It can be seen as a measure of how well point $\boldsymbol{x}$ is accommodated into

[1] Depending on m, the triplets $\boldsymbol{\theta}_k^{(m)}$ may be taken to coincide with elements defined by the approach itself (e.g., in model-based clustering, each triplet coincides with the parameters of one mixture component; e.g., see [13]), or can be estimated with sample quantities, computing within-cluster empirical counterparts (using estimated point-to-cluster assignments).

cluster k (higher values correspond to a better fit). Using (1) we define the quadratic partition, $\mathscr{Q}$, as

$$\mathscr{Q} = \{Q_k,\ k \in \{1,\ldots,K\}\}, \quad Q_k = \left\{ \boldsymbol{x} \in \mathbb{R}^p : k = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) \right\}. \tag{2}$$

Finally, for a given $\boldsymbol{\theta}$ we define the hard and smooth quadratic scoring criteria respectively as

$$H(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} s_H(\boldsymbol{x}, \boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} \mathbb{I}\left\{ \boldsymbol{x} \in Q_k(\boldsymbol{\theta}^{(m)}) \right\} \mathrm{qs}\left(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)} \right); \tag{3}$$

$$T(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} s_T(\boldsymbol{x}, \boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} \tau_k\left(\boldsymbol{x}_i, \boldsymbol{\theta}^{(m)} \right) \mathrm{qs}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(m)} \right), \tag{4}$$

where τ_k defines a smooth weight of point-to-cluster membership as measured by the quadratic score:

$$\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\mathrm{qs}(\boldsymbol{x}, \theta_k)}{\sum_{k=1}^{K} \mathrm{qs}(\boldsymbol{x}, \theta_k)}.$$

2.1 Likelihood-based validation

The quadratic score (1) is strongly connected to likelihood theory, and it is easy to show that it is proportional to the Gaussian density function [4]. Thus, a natural alternative to the scoring criteria (3) and (4) appears to be the likelihood function of a Gaussian mixture model. A similar proposal was also made by Smyth [16], used in combination with cross-validation. Thus, the data log-likelihood can be used to score a clustering solution (solutions achieving higher likelihood are preferred), and is defined as:

$$l(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \log \left(\sum_{k=1}^{K^{(m)}} \pi_k^{(m)} \phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)}) \right), \tag{5}$$

where $\phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})$ is the density function of a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. Note that, differently from (3) and (4), the usage of (5) is only justified for model-based clustering, where a probabilistic model is assumed, and it is needed to motivate the construction of the likelihood function.

Algorithm 1 Bootstrap quadratic scoring

input: observed sample $\mathbb{X}_n$ (with ecdf $\mathbb{F}_n$), $\alpha \in (0,1)$; clustering method $m \in \mathscr{M}$; integers $B > 0$
output: bootstrap quadratic scoring for method m: $\widetilde{L}_n^{(m)}$.
(to ease notation, dependence on m is dropped and reintroduced in step 3)
for $b \in \{1,\ldots,B\}$ do
(step 1.1) $\mathbb{X}_n^{(b)} \leftarrow$ non-parametric bootstrap resample from $\mathbb{X}_n$ (sample of size n from $\mathbb{F}_n$)
(step 1.2) $\hat{\boldsymbol{\theta}}_n^{(b)} \leftarrow$ triplets of parameters from clustering solution m fitted on $\mathbb{X}_n^{(b)}$
(step 1.3) $S_n^{(b)} \leftarrow$ score solution on points in $\mathbb{X}_n$ not in $\mathbb{X}_n^{(b)}$

$$S_n^{(b)} = H(\hat{\boldsymbol{\theta}}_n^{(b)};\mathbb{X}_n) \quad \text{or} \quad S_n^{(b)} = T(\hat{\boldsymbol{\theta}}_n^{(b)};\mathbb{X}_n) \quad \text{or} \quad S_n^{(b)} = l(\hat{\boldsymbol{\theta}}_n^{(b)};\mathbb{X}_n)$$

end for
(step 2) $\widetilde{W}_n \leftarrow \frac{1}{B}\sum_{b=1}^{B} S_n^{(b)}$ $\qquad R_n^{(b)} \leftarrow \sqrt{n}\left(S_n^{(b)} - \widetilde{W}_n\right)$
(step 3) Compute $(\alpha/2)$-level and $(1-\alpha/2)$-level empirical quantiles:

$$\widetilde{L}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \leq t\right\} \geq \frac{\alpha}{2} \right\}; \qquad \widetilde{U}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \leq t\right\} \geq 1-\frac{\alpha}{2} \right\}$$

2.2 Resampling schemes

Choosing the solution that maximizes (3) or (4) may give poor results: since the sample data $\mathbb{X}_n$ is used both for estimation and scoring, overly-complex solutions may be selected due to overoptimism in the evaluation process. It is well known that increasingly complex solutions can fit the data better, and many selection methods takes this into account in several ways. For example, information criteria typically include a penalization term for the model's complexity (e.g., BIC [15] and ICL [3]). To cope with this issue, in Coraggio and Coretto [4], we propose to estimate clustering solutions on non-parametric bootstrap resamples [7] from $\mathbb{X}_n$, and to evaluate (3) and (4) on $\mathbb{X}_n$. The procedure is reviewed in Algorithm 1, and here we extend it to include a likelihood-based scoring, using (5) as the scoring function.

3 Empirical analysis

The experimental analysis is a scaled-down version of that in Coraggio and Coretto [4], using the Iris, Olive and Banknote data sets [1, 9, 8]. Since we are going to compare the quadratic scores with likelihood-based ones, $\mathscr{M}$ includes only model-based clustering methods: *(i)* 140 model-based clustering methods, using Gaussian mixture models with parsimonious representation of the covariance matrices [2], implemented with the Mclust software [14] (setting $K = 1,\ldots,10$, and 14 covariance models); *(ii)* 180 model-based clustering methods, using Gaussian mixture models with eigen-ratio contraints (ERC; [12]), implemented with Otrimle software

Table 1: Selected clustering solution by selection criteria (left-most column). Each sub-table shows results from a real data set: the first column shows the selected solution, and the second column reports its ARI, computed against true classes.

	(b) Iris		(c) Olive		(d) Banknote	
Criterion	Selected m	ARI	Selected m	ARI[1]	Selected m	ARI
QH	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
QS	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
LK	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
CVQH	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=6, VVV	0.79	Mclust, K=4, VVE	0.68
CVQS	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=6, VVV	0.79	Otrimle, K=3, γ=10	**0.86**
CVLK	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=9, VEE	0.65	Mclust, K=4, VVE	0.68
BQH	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Otrimle, K=3, γ=10	**0.86**
BQS	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Otrimle, K=3, γ=10	**0.86**
BLK	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Mclust, K=6, EEE	0.47

(Olive) 1. refers to the finer 9-regions classification.

[5, 6] (setting $K \in \{1, \ldots, 10\}$, ERC $\gamma \in \{1, 5, 10, 10^2, 10^3, 10^4\}$, and 3 initialization methods). The criteria compared to select optimal solutions are respectively based on equations (3), (4), and (5), and are divided as follows. QH, QS, and LK: clustering solutions are estimated and scored using the full data, $\mathbb{X}_n$; CVQH, CVQS, CVLK: clustering solutions are estimated on a "train set" and scored on a non-overlapping "test set", using a 10-fold cross-validation scheme, as in [16]. BQH, BQS, BLK: clustering solution are estimated and scored according to Algorithm 1, selecting the method m maximizing $\widetilde{L}_n^{(m)}$. For each criterion, the selected solutions are evaluated against the true class labels, reporting the achieved Adjusted Rand Index (ARI, [11]).

Results are presented in Table 1. The comparison gives a better understanding on the mechanism that lies behind the effectiveness of the BQH and BQS criteria. First, notice that all criteria where solutions are estimated and scored on the full data (QH, QS, LK) always select poor, overly-complex solutions: this is because of the problem discussed in Subsection 2.2. On the contrary, both resampling mechanisms choose lower complexity solutions, acting similar to complexity penalization. However, the cross-validation scheme produces poor (Olive) or inconsistent (Banknote) results, with respect to the bootstrap scheme. This is likely due to the fact that 10-fold cross-validation splits leave too few shared information between train and test sets, and is less adequate for clustering where, differently from prediction settings, the goal is to select a clustering method that has a good in-sample performance. This is also true for the likelihood-based criteria, CVLK and BLK, which perform worse with respect to the quadratic score criteria. Indeed, it can be shown that compared to the likelihood function, the quadratic scores add extra penalization for overlapping clusters. Overall, both the quadratic scores, (3) and (4), and the resampling scheme in Algorithm 1 seem equally important to consistently achieve good results.

4 Conclusion

In this paper, we reviewed the BQH and BQS procedures from [4]. We ran an empirical analysis comparing it with a likelihood-based validation index, and testing different resampling schemes. Our experiments show that neither the bootstrap resampling scheme in Algorithm 1 nor the quadratic discriminant score in (1), alone, are sufficient to obtain good results. Rather, both of them contribute to the procedure in different ways. The comparison with likelihood-based alternatives highlights that the extra penalty that the quadratic scores impose on overlapping clusters is needed to select better solutions in cases where clusters are not well separated. Similarly, the comparison with the 10-fold cross-validation shows that the bootstrap resampling scheme is better suited for cluster analysis, where the final goal is to describe the data at hand, rather than predicting cluster membership for unseen points.

References

1. Anderson, E.: The species problem in Iris. Ann. Missouri Bot. Gard. 23 (3) 471–483 (1936)
2. Banfield, J. D., Raftery, A. E.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics, 49 (3) (1993)
3. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 22 (7), 719–725 (2000)
4. Coraggio, L., Coretto, P.: Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. J. Multivariate Anal. , 196, 105181, (2023)
5. Coretto, P., Hennig, C.: Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. J. Mach. Learn. Res. 18 (142), 1–39 (2017)
6. Coretto, P., Hennig, C.: OTRIMLE: Robust model-based clustering. R package version 2.0 (2021)
7. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7 (1) (1979)
8. B. Flury and H. Riedwyl: Multivariate statistics: a practical approach., 1st ed. Chapman and Hall/CRC, London—New York (1988)
9. Forina, M., Armanino, C., Lanteri, S., Tiscornia, E.: Classification of olive oils from their fatty acid composition. Food Res. Data Anal. 189–214 (1983)
10. Hastie, T., Tibshirani, R. J., Friedman, J.: The Elements of Statistical Learning. 2nd ed., Springer New York (2009)
11. Hubert, L., Arabie, P.: Comparing partitions., J. Classification 2 (1), 193–218 (1985)
12. Ingrassia, S.: A likelihood-based constrained algorithm for multivariate normal mixture models. Stat Methods Appt 13 (2) (2004)
13. McLachlan, G., Peel, D.: Finite mixture models. In: Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York (2000)
14. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R J. 8 (1), 205–233 (2016)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. 6 (2) 461–464 (1978)
16. Smyth, P.: Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. 10 (1) 63–72 (2000)
17. von Luxburg, U., Williamson, R. C., Guyon I.: Clustering: Science or Art?. In Proceedings of ICML Workshop on Unsupervised and Transfer Learning, Bellevue, Washington, USA, 27, pp. 65–79 (2012)

Solicited Session SS17 - *Statistical methods for education and educational services*

Organizer and Chair: Enrico Ripamonti

1. *Association between INVALSI scores and students' mobility in Italy: a preliminary analysis* (Bacci S., Bertaccini B., Lombardi G. and Tocchioni V.)
2. *Modelling Responses and Response Times: an application to Mathematics INVALSI data* (Bungaro L., Desiderati R. and Mignani S)
3. *Latent potential outcomes: An analysis of the effects of programs aimed at improving students' non-cognitive skills* (Pennoni F., Bartolucci F. and Vittadini G.)
4. *Cognitive Skills and Non Cognitive Skills to Analyze School and Students Performances* (Vittadini G.)

Association between INVALSI scores and students' mobility in Italy: a preliminary analysis

Associazione tra risultati ai test INVALSI e mobilità studentesca in Italia: un'analisi preliminare

Silvia Bacci, Bruno Bertaccini, Gabriele Lombardi and Valentina Tocchioni

Abstract This work explores the relationship between INVALSI scores and students' mobility, referred to the decision of enrolling in a university distant more than two hours drive from their hometowns. Through data drawn from the Italian Student Register for the cohort of graduates in Italian High Schools during the school year 2018/2019 – and enrolled in the Tertiary Education system in the academic year 2019/2020 – we estimate a Logit model for the decision of being a *mover*. The INVALSI scores in Italian language and Mathematics emerge as positively correlated with the decision of moving.

Abstract *Il presente studio esplora la relazione fra punteggi INVALSI e mobilità studentesca, con riferimento alla decisione di iscriversi in un'università che disti più di due ore di automobile dalla città di residenza. Attraverso i dati dell'Anagrafe Nazionale Studenti (ANS) per la coorte di diplomati nell'anno scolastico 2018/2019 – ed iscritti nel sistema universitario nell'anno accademico 2019/2020 – stimiamo un modello Logit per la decisione di essere un*o studente fuori sede*. I risultati dei test INVALSI in Italiano e Matematica sono positivamente correlati alla decisione di spostarsi.*

Key words: Students' mobility, students' performance, INVALSI score, Italy

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

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

Gabriele Lombardi
University of Florence, Department of Statistics, Informatics, Applications "G. Parenti"; e-mail: gabriele.lombardi@unifi.it

Valentina Tocchioni
University of Florence, Department of Statistics, Informatics, Applications "G. Parenti"; e-mail: valentina.tocchioni@unifi.it

1 Introduction

Reasons behind interregional and intraregional mobility for higher education purposes regard both personal motivations and the socio-economic characteristics of the host region. Students in higher education may be driven to leave their place of residence to seek better career and life opportunities if there is a lack of prospects in their home environment. They may aspire to higher wages, to better life conditions and quality of life, or to improve their human capital by searching for an educational offer not available nearby and/or for the universities considered most prestigious [8,7]. From this perspective, moving for academic reasons is not always a choice: indeed, a moving student experiments with several settling costs both from economic, social and psychological points of view who can prevent him/her from enrolling in a faraway university [4]. Anyway, sometimes moving for higher education reasons is mandatory, if young people want to pursue academic studies and no university is located nearby. But even in this case, they may opt for the nearest university to their residence, or for a more distant one, because for several reasons that may range from personal ambitions and aspirations to better road connections or train services.

Given these premises, who is most likely to pursue mobility? The literature identifies both individual resources and family background as features related to students' mobility at different levels of education. Broadly speaking, individual skills are universally accepted in influencing students' mobility to higher education, towards both bachelor's [9] and master's programs [3]. Among individual skills that may be associated with students' mobility, the final high-school grade has already been taken into account in empirical applications on the Italian case and are in line with the expectation of greater mobility of students with high final grades [9,4,5]. The high-school final mark may be considered an overall result of performance for high-school graduates, a sort of synthesis of several achievements and marks in different subjects. In Italy, individual skills of high-school students may also be measured through the administration of the tests of the National Institute of Evaluation of the Education and Training System (INVALSI), consisting of standardized tests on a national scale that students carry out at different stages of their career to identify their level of competence in subjects like literacy, numeracy, and English reading and listening competencies.

As far as we know, the capability of the INVALSI test scores in predicting the decision to become a mover has not been tested yet. Thus, this work aims at investigating if individual performances – in terms of high-school final grade and INVALSI scores obtained during the last high-school year - have an impact on student's propensity to move for enrolling in a university. We define a student as a mover if s/he enrolls in a University that is at least two-hour drive from home. Here we concentrate on first-level mobility, which refers to the transition from the high-school to the bachelor's (or single cycle) degree [6]. To investigate whether and how individual attitudes and students' mobility are associated, we address the following research question: Does the likelihood of being a mover differ according to high-school final grade and Invalsi scores? Conscious of the fact that the association between individual attitudes and students' mobility might change when accounting for other individual characteristics and features of program degrees, we control for a series of other individual characteristics, family background, and bachelor's features.

2 Data and Methods

Data - belonging to the Database MOBYSU.IT [Mobilità degli Studi Universitari in Italia, Data Source ANS-MUR/CINECA[1] - consists of 211,095 freshmen enrolled in the Italian Higher Education System in the academic year 2019/2020 and who have graduated in a high school in the school year 2018/2019[2].

Table 1 Descriptive Statistics: Means for continuous variables and proportions for categorical variables.

Variables	***Full Sample***	***Stayers***	***Movers***
Gender			
Females	0.5444	0.5423	0.5598
Males	0.4556	0.4577	0.4402
Type of High School			
Scientific Lyceum	0.3714	0.3619	0.4406
Classic Lyceum	0.1108	0.1012	0.1801
Other Lyceum	0.2394	0.2437	0.2076
Technical	0.2120	0.2232	0.1308
Other	0.0664	0.0700	0.0409
Residence Macro-Region			
North-West	0.2509	0.2755	0.0715
North-East	0.1773	0.1869	0.1071
Centre	0.2079	0.2148	0.1579
South	0.2618	0.2337	0.4665
Islands	0.1021	0.0891	0.1970
Socio-Economic Background			
ESCS	0.3189	0.2987	0.4659
High School Final Grade			
1st Quartile (60-70)	0.2487	0.2602	0.1649
2nd Quartile (71-80)	0.2985	0.3043	0.2566
3rd Quartile (81-90)	0.2065	0.2035	0.2279
4th Quartile (91-100 cum laude)	0.2463	0.2320	0.3506
Invalsi Scores			

[1] Data - drawn from the Italian "Anagrafe Nazionale della Formazione Superiore"- has been processed according to the research project "From high school to the job market: analysis of the university careers and the university North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero Università e Ricerca", and INVALSI

[2] We excluded students enrolled to a course related with "Health and Welfare", due to the national contest needed to enrol medicine-related courses in Italy. Moreover, some variables were imputed through a Conditional Tree Regression with recursive partitioning: i.e. Residence Macro-Region, ESCS, High School Final Grade and INVALSI scores.

Italian language	215.00	214.2	220.5
Mathematics	215.47	215.01	218.77
Fields of Study	*N* = 9	*N* = 9	*N* = 9
No. of Observations	*N* = 211,095	*N* = 185,593	*N* = 25,502

Our main variable of interest is a dichotomous indicator assuming value 1 if a student enrols in a university being at least two-hour drive from home (*movers*), 0 otherwise (*stayers*). For each freshman, we can observe several characteristics reported in Table 1. Gender is well-balanced within our data, with a small prevalence of females (54%), who prevail also within movers (56%). The relative majority of students comes from a scientific lyceum (37%), the share increasing up to 44% among movers. The socio-economic background can be controlled through the Economic, Social, and Cultural Status Index (ESCS), whose mean is substantially higher among movers rather than stayers (0.4659 vs 0.2987). Even if more than 62% of students come from the central-northern areas of the country, they represent only 33% of movers. Moreover, 58% of those students who took a high school final grade greater than 80 are movers. Regarding INVALSI scores, we analyse results in Italian language and Mathematics: in both cases mean values for movers are higher than the overall means. Finally, we can distinguish among nine fields of studies as in the ISCED classification: Agriculture, forestry, fisheries and veterinary; Arts and humanities; Business, administration and law; Education; Engineering, manufacturing and construction; Information and Communication Technologies (ICTs); Natural sciences, mathematics and statistics; Services; Social sciences, journalism and information.

The likelihood of being a mover is estimated through a Logit regression, where INVALSI scores are standardized. Predicted probabilities will be plotted for INVALSI scores, distinguished by class of high school final grade and residence macro-region. Other characteristics are fixed at means for continuous variables and modes for categorical variables.

3 Results and discussion

Table 2 reports model coefficients for the logit regression estimating the likelihood of being a mover. Both the high school final grade and the INVALSI scores are positively associated with the decision of being movers. Then, females have more chance of moving than men, while students from classic lyceums are the only ones at being more likely of moving than their scientific lyceums' peers. Students from the South and Islands are those with higher chances of moving far away their hometowns, while decreasing as much of a student comes from a northern area of the country. Also, individual ESCS is positively correlated with the decision of moving. As it can be noticed by looking at Figure 1 (for INVALSI score in the Italian language) and in Figure 2 (for INVALSI score in Mathematics), the difference in predicted probabilities across quartiles of high school final grades is negligible, even if they slightly increase as the quartiles increase. On the contrary, the predicted probabilities of being a mover increase as the INVALSI score increases (irrespective of the subject). Coherently with the academic literature, the macro-region of residence plays a key role in the decision of moving, with students from the South and Islands exhibiting far higher chances of becoming movers for Higher Education studies. Our results suggest that high INVALSI scores are strongly associated with the decision of moving: further development will explore the propensity of being a mover according to students' individual performance by university reputation. This concept will be operationalized using the composite indicator of university reputation at the first-level created by Bacci and Bertaccini [1].

Table 2 Logit Model Estimates.

Dep.Var: 0=Stayers;1=Movers	***Coeff.***	***Std.err.***	***p-value***
Intercept	-3.72	0.04	0.00
Gender (ref: Females)			
Males	-0.10	0.02	0.00
Type of High School (ref: Scientific Lyceum)			
Classic Lyceum	0.26	0.02	0.00
Other Lyceum	-0.07	0.02	0.00
Technical	-0.32	0.02	0.00
Other	-0.28	0.04	0.00
Residence Macro-Region (ref: North-West)			
North-East	0.81	0.03	0.00
Centre	1.14	0.03	0.00
South	2.26	0.03	0.00
Islands	2.42	0.03	0.00
Socio-Economic Background			
ESCS	0.25	0.01	0.00
High School Final Grade (ref: 1st Quartile)			
2nd Quartile (71-80)	0.13	0.02	0.00
3rd Quartile (81-90)	0.24	0.02	0.00
4th Quartile (91-100 cum laude)	0.28	0.02	0.00
Invalsi Scores			
Italian language	0.13	0.01	0.00
Mathematics	0.23	0.01	0.00
No. of Observations: 211,095			

Fig. 1 Predicted Probabilities for Italian Language INVALSI Score, by High School Final Grade and Residence Macro-Region.

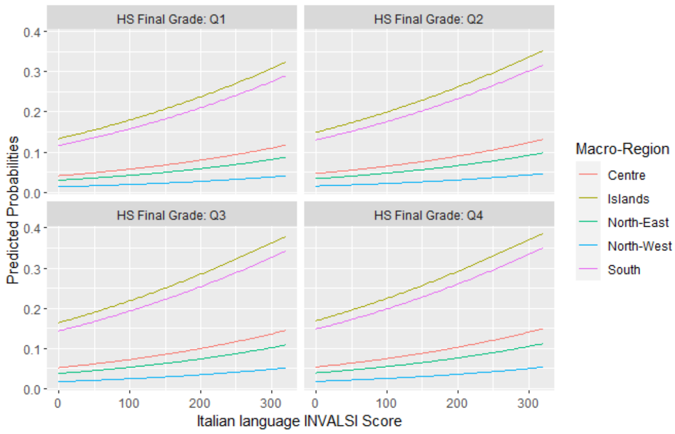
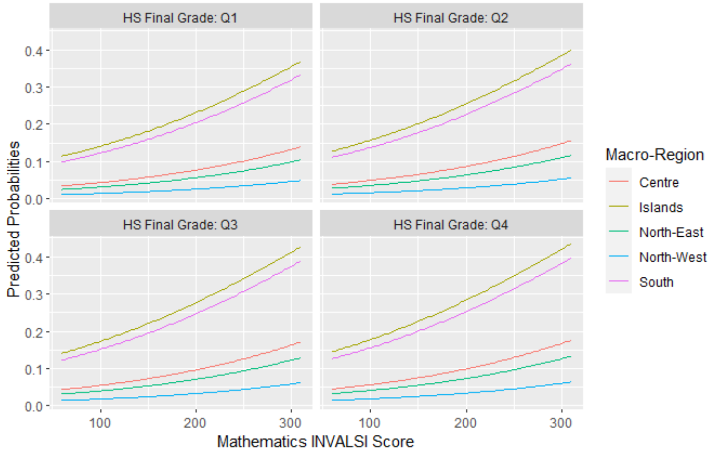


Fig. 2 Predicted Probabilities for Mathematics INVALSI Score, by High School Final Grade and Residence Macro-Region.



References

1. Bacci, S., Bertaccini, B.: Assessment of the university reputation through the analysis of the student mobility. Social Indicators Research, 156, 363-388 (2021)
2. Columbu, S., Porcu, M., Primerano, I., Sulis, I., Vitale, M.P.: Analysing the determinants of Italian university student mobility pathways. Genus, 77, 1-20 (2021)
3. D'Agostino, A., Ghellini, G., Lombardi, G.: Movers and stayers in STEM enrollment in Italy: Who performs better? Genus, 77 (2021)
4. D'Agostino, A., Ghellini, G., Lombardi, G.: Gender Effect at the Beginning of Higher Education Careers in STEM Studies: Does Female Recover Better Than Male? In: Khine, M.S. (eds.) Methodology for Multilevel Modeling in Educational Research. Springer, Singapore (2022)
5. Enea, M., Attanasio, M. La mobilità degli studenti universitari nell'ultimo decennio in Italia. In: G. De Santis, E. Pirani, & M. Porcu (eds.), Rapporto sulla popolazione. L'istruzione in Italia, pp. 43–58. Il Mulino, Bologna (2019)
6. Lombardi, G., Ghellini, G.: The effect of grading policies on Italian universities' attractiveness: A conditional multinomial logit approach. Electronic Journal of Applied Statistical Analysis, 12(4), 801–825 (2019)
7. Ragozini G., Scolorato C., Santelli F.: Le determinanti della mobilità degli studenti universitari campani. In: Buono P., Gallo M., Ragozini G., Reverchon E., Rostirolla P. (eds). Il sistema universitario campano tra miti e realtà. Aspetti metodologici, analisi e risultati. Franco Angeli, Milano (2016)
8. Tosi, F., Impicciatore, R., Rettaroli, R.: Individual skills and student mobility in Italy: A regional perspective. Regional Studies, 53(8), 1099-1111 (2019)

Modelling Responses and Response Times: An application to the Mathematics INVALSI data

Modellare le risposte e i tempi di risposta: un'applicazione ai dati Invalsi di Matematica

Luca Bungaro, Roberto Desiderati and Stefania Mignani

Abstract In this paper, the responses and response times of the INVALSI Mathematics test for Grade 13 students are jointly analyzed. This approach allows to improve the accuracy of the ability estimation taking into account the speed in responding. The results obtained consolidate the hypothesis that students with higher ability are more engaged to respond, even during a test that does not directly affect their school average, while students with lower ability tend to be less interested and more hasty.
Abstract *In questo lavoro vengono analizzate congiuntamente le risposte e i tempi di risposta del test di Matematica INVALSI per gli studenti di grado 13. L'approccio congiunto permette di migliorare la precisione della stima dell'abilità tenendo in considerazione anche la velocità nel rispondere. I risultati confermano l'ipotesi che lo studente preparato vuole impegnarsi e risponde prendendo il tempo necessario mentre lo studente meno preparato tende ad essere meno interessato e più frettoloso.*

Key words: Item Response Theory, Computer Based Testing, Response time, INVALSI data

1 Introduction

Large-scale survey assessments have been utilized for many years to track students' knowledge and abilities. In recent years, the implementation of computer based testing (CBT) has been receiving a growing interest because of its operational advantages. CBT allows to automatically collect data not only on the students' response

Luca Bungaro
Department of Statistical Sciences, e-mail: luca.bungaro2@unibo.it

Roberto Desiderati
Department of Statistical Sciences, e-mail: Roberto.desiderati@studio.unibo.it

Stefania Mignani
Department of Statistical Sciences, e-mail: Stefania.mignani@unibo.it

accuracy (RA) based on item responses, but also on their response times (RT). Using the RTs, the assessment results can be further improved in terms of precision, fairness, and minimizing costs. RTs can also reveal new information about test characteristics, test takers' response alone behavior, and test takers' ability that would not be identified when using response information [3].

The fundamental idea is that a higher level of ability is associated to a higher probability of providing a correct response. Similarly, this concept can be applied to response time modelling where higher speed of working is linked to a lower expected response time. This paper is a novelty in the application of response times to improve the estimation of scores from computer-based tests. In particular, the results of a large-scale standardized test administered in Italy by the National Institute for the Evaluation of the Education and Training System (INVALSI) have been analyzed.

2 Focus of the study

INVALSI every year administers standardized tests via CBT to students attending grades 8, 10, and 13 to evaluate competence in Mathematics, Italian and English language. In this study, we use the mathematics data for grade 10 administered at the end of 2017-2018 scholastic year.

The number of involved students is very large and tests must be administered in multiple sessions and locations so several test forms are assembled to overcome security concerns. The tests of mathematics are carried out in 12 different parallel forms each of which is characterized by 35 items coming from a calibrated items bank. The same item may be present in several tests and the tests are assembled considering some constrains (length, difficulty, domain of the item,..).

The computer-based tests are time limited to 90 minutes.

2.1 The models

The data on Response Accuracy, i.e. correct/incorrect response, and Response Time are collected for each item and modelled following a Bayesian joint model with a hierarchical structure ([4], [2], [1]).

At the first level, separate models for responses and response times are defined. At the second level, a distributional structure is defined for the model parameters and hyperprior distributions are specified for the parameters.

At level 1, the two-parameter normal ogive (2PNO) model was used to define the mathematical relationship between the probability of correct response and the person and item parameters as follows

$$P(Y_{ik} = 1 \mid \theta_i,\, a_k,\, b_k) \;=\; \phi(a_k\theta_i \,-\, b_k)\,, \qquad (1)$$

where Y_{ik} is the binary response variable taking value 1 when the response is correct and 0 otherwise, with $i = 1,...,N$ test-takers and $k = 1,...,K$ items, b_k is the difficulty parameter of item k, a_k is the discrimination parameters, θ_i denotes the ability of test-taker i, and $\phi(.)$ is the normal cumulative distribution function.

Then, a log-normal distribution is used to model the RTs and the log RTs are stored in a $\mathbf{N} \times \mathbf{K}$ matrix RT. In this way, the generic element $lnRT_{ik}$ is assumed to be normally distributed as follows

$$lnRT_{ik} \; = \; \lambda_K \; - \; \psi_K \zeta_i \; + \; \varepsilon_{ik}, \; \varepsilon_{ik} \sim N(0, \sigma^2_{\varepsilon_{ik}}) \; , \tag{2}$$

where λ_K is the time-intensity parameter of item k, representing the population-average time (on a logarithmic scale) needed to complete an item, ζ_i is the speed parameter of test-taker i, representing the constant working speed of that test-taker, as the systematic differences in RTs given λ_K, ψ_K is the time-discrimination parameter of item k, representing the sensitivity of the item for different speed levels of the test takers. Lastly, ε_{ik} is an additional error term that can model variations in RTs that cannot be explained only by the structural mean term, such as when test-takers operate with different speed values, take small pauses during the test, or change their time management.

At level 2, a distributional structure is defined for the level 1 person and item parameters. For the ability and speed, a bivariate normal distribution is defined with an inverse-Wishart distribution for the covariance matrix. Model parameters are estimated through the Gibbs sampling algorithm. To identify the model, some restrictions are imposed: the product of the time discrimination is fixed to one and the mean of ability and speed are fixed to zero.

The model can be extended with a multivariate multilevel regression structure which allows the inclusion of covariates to increase estimation precision of person parameters.

Let $\mathbf{X}_\theta$ denote the predictors for the ability parameter and $\mathbf{X}_\zeta$ for the speed parameter. The mean component for the person parameters can be expressed as

$$\begin{aligned} \mu_\theta \; &= \; \mathbf{X}_\theta \beta_\theta \\ \mu_\zeta \; &= \; \mathbf{X}_\zeta \beta_\zeta \end{aligned} \tag{3}$$

Non informative normal priors are defined for the regression parameters with a mean of zero and a large variance.

3 The results

In this paper, we analyze the results of a sample of students who responded to the same INVALSI test form (one of the 12 parallel test forms). The data set included, for each item, the binary response and the response time; for each student, some covariates such as gender, type of school, socio-economic background, are also collected.

First, we may want to investigate the correlation between the time intensity and the item difficulty parameters. Second, the analysis could focus on the structural relationships between explanatory information at the individual level and the test takers' ability and speed.

3.1 Some descriptive statistics

The data set has been cleaned eliminating all the students who registered a total test time exceeding 5400 seconds or answered at least one item in less than 10 seconds. The definitive sample size was 1813 units.

The Fig. 1 shows that most students employed on average from one to one and half minute to give the answers.

Fig. 2 illustrates the distribution of overall time to conclude the test that is left-skewed. To estimate the model parameters a logarithmic transformation has been applied.

3.2 The estimated parameters

Table 1 shows a summary of the item parameter estimations.

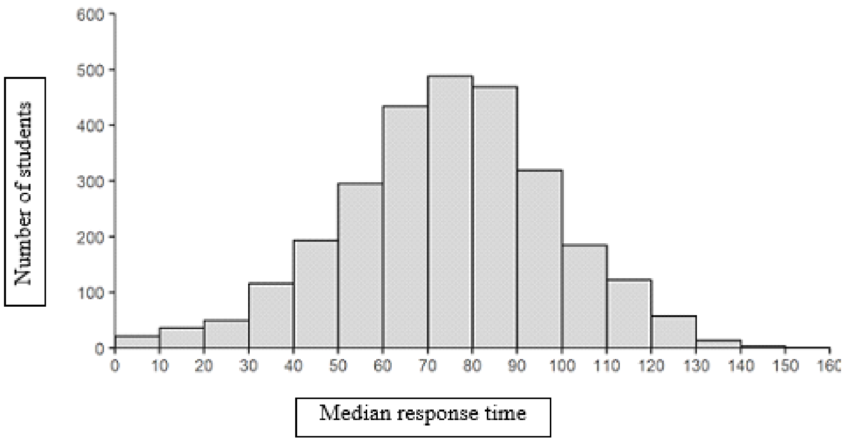


Fig. 1 Distribution of median response time

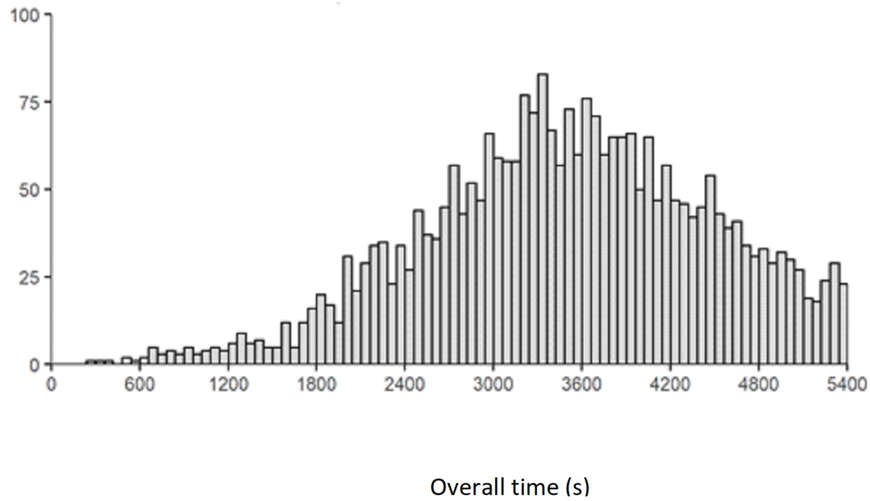


Fig. 2 Distribution of overall time to conclude the test

Table 1 The summary of item estimated parameters

Parameters	Mean	S.D.	Min	Max
Discrimination a	1.037	0.062	0.500	1.504
Difficulty b	-0.241	0.085	-1.668	0.805
Time discrimination ψ	1.088	0.059	0.251	1.964
Time intensity λ	4.392	0.077	3.301	5.336

The mean of the difficulty parameters indicates that the overall test is quite easy, as very difficulty items are not present in the test. The item discrimination parameters show an average slight impact. The estimated time discrimination parameters varied over [.25, 1.96], indicating that the items discriminate substantially between test takers of different speed.

Concerning the person parameters, the ability follows a normal distribution on the range [-1.744; 1.600], while the speed distribution curve is slightly skewed on the range [-0.338; 0.699].

The model goodness of fit based on residual analysis is quite satisfactory: only the 10.81% of students show aberrant response time patterns, while no items show aberrant patterns.

The correlation for person is -0.373 and the correlation matrix for item parameters is given in Table 2.

There is a positive relationship between the item difficulty and the item intensity: the most difficult (easy) items are also the ones that discriminate better (worse) and require more (less) time to perform. The negative correlation between time-discrimination and time-intensity, on the other hand, indicates that on average the

Table 2 Correlations of item parameters

Parameters				
	a	b	ψ	λ
a	1			
b	0.324	1		
ψ	-0.282	0.365	1	
λ	0.153	0.165	0.125	1

items that require more (less) time are the ones that discriminate worse (better), but with a very low and not significant magnitude. n particular, it goes to consolidate that hypothesis for which those who are prepared want to engage and show their skills, even during a test that does not directly affect their school average, while those who are less prepared tend to be less interested and more hasty. The estimated model with covariates (gender and type of school) highlighted the usual gender gap in favor of boys for the ability but not for speed, while the type of school showed the better performance of student attending a lyceum, as usual, but no effect on the speediness.

4 Concluding remarks

The use of a joint model for RA e RT makes available more accurate estimates of student achievement. The results of this study need to be confirmed through additional research relaxing the assumption that students respond to all items considering constant speed and also including the sequence in which a student has responded to the items to highlight aberrant response patterns.

References

1. Fox, J.-P., Klotzke, K., Simsek, A. S., LNIRT: An R Package for Joint Modeling of Response Accuracy and Times. ARxIV (2021)
2. Klein Entink, R. H., Fox, J.-P., van der Linden, W. J.: A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. Psychometrika, 74(1), 21-48 (2009)
3. van der Linden, Wim J.: A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics 31(2), 181-204 (2006)
4. van der Linden, W. J.: A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. Psychometrika, 72(3), 287-308 (2007)

Latent potential outcomes: An analysis of the effects of programs aimed at improving students' non-cognitive skills

Variabili latenti potenziali: Un'analisi degli effetti dei programmi per migliorare le abilità non cognitive degli studenti

Fulvia Pennoni, Francesco Bartolucci and Giorgio Vittadini

Abstract We illustrate a causal latent transition model to evaluate the effects of educational programs administered to pupils in the 6th and 7th grades during their middle school period. The programs are conducted in an Italian region and focus on improving non-cognitive abilities. The interest is in evaluating the effects on the skills acquired in the 8th grade in Italian and Mathematics. The model can be cast in the hidden Markov literature and is formulated as an extension of Rubin's causal model based on potential versions of discrete time-varying latent variables.

Abstract *Un modello di transizione causale viene utilizzato per la valutazione degli effetti di programmi sviluppati in una regione italiana per studenti che frequentano il primo ed il secondo anno della scuola secondaria di primo grado. Si intende valutare l'effetto di programmi volti a sviluppare le abilità non cognitive sugli apprendimenti riscontrati in italiano e matematica al termine del ciclo della scuola secondaria. Il modello rientra nella classe dei modelli di Markov a variabili latenti ed estende il modello causale di Rubin utilizzando variabili latenti potenziali discrete e dinamiche nel tempo.*

Key words: causal inference, expectation-maximization algorithm, hidden Markov models, human capital, observational studies

---

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy, e-mail: fulvia.pennoni@unimib.it

Francesco Bartolucci
Department of Economics, University of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Giorgio Vittadini
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy, e-mail: giorgio.vittadini@unimib.it

1 Applicative context

Extensive literature demonstrates that there are strong links between Non-Cognitive Skills (NCSs) and Cognitive Skills (CSs) [4]. NCSs are defined as personality resources linked to motivation in learning, relational capabilities, emotional stability, and autonomy in pursuing personal objectives. We investigated if NCSs programs also causally determine an improvement of the CSs in pupils evaluated from the 5th grade through the 8th grade during the 2015-2018 school years. We considered a sample of primary and middle class students of the Autonomous Province of Trento (PAT) with data derived from the integration of five different sources. We consider 25 schools with 1,561 pupils (out of 77 with a total population of 5,502 students) that freely accepted participating in the PAT survey in 2015, among which 12 schools with 845 students voluntarily adopted the educational programs proposed to improve the NCSs. As response variables we considered the scores students achieved in the INVALSI (Italian National Institute for the Evaluation of the Educational System) tests in the 5th and 8th grades for Italian and Mathematics. A large set of measurements on individual covariates regarding family background is considered in the analysis. Interest lies in identifying and estimating the average causal effects on the treated (ATETs) of these programs on the student's CSs.

The remainder of this paper is organized as follows. In Section 2 we present the model formulation, and in Section 3 we show some of the results illustrated in [2]. Section 4 provides some concluding remarks.

2 Causal latent transition model

In the applicative context illustrated in Section 2 the interest lies in estimating the ATETs using observational data, when a response variable before ($t = 0$) and after ($t = 1$) the treatment of interest is considered. A causal latent transition (CLT) model is proposed in [2] for estimating the effect of an intermediate treatment considering multivariate responses under the framework of hidden Markov (HM) models [1]. The CLT model relies on potential versions of the latent variables and it extends the traditional approach of potential outcomes; see [6], among others.

We consider vectors of continuous response variables ($\mathbf{Y}_{i0}$, $\mathbf{Y}_{i1}$) for a sample of n students available at the two occasions, further to baseline covariates which are time-constant and collected in the vectors $\mathbf{X}_i$. For every student i we assume that the individual-specific response variables depend on a vector $\mathbf{H}_i = (H_{i0}, H_{i1})'$ of two latent variables following a first-order Markov chain with state space having a discrete distribution with support $\{1,\dots,k\}$. States represent subpopulations of students sharing common latent characteristics. Moreover, we assume conditional independence between the response variables given the latent process at the two consecutive time occasions. Underlying every H_{it} a set of potential latent variables (potential latent outcomes, PLOs) is introduced and denoted as $H_{it}^{(z)}$, $i = 1,\dots,n$,

$t=0,1$, $z=0,1$. Each $H_{it}^{(z)}$ corresponds to the latent state of student i at occasion t if h/she had taken treatment ($z=1$) or not ($z=0$). The ATET is specific of the two potential latent states at each time occasion and it is defined as

$$\begin{aligned}\mathrm{ATET}_{1\bar{h}h}(\mathbf{x}) = &\log\frac{p(H_{i1}^{(1)}=h|H_{i0}^{(0)}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)}{p(H_{i1}^{(1)}=\bar{h}|H_{i0}^{(0)}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)}\\ &-\log\frac{p(H_{i1}^{(0)}=h|H_{i0}^{(0)}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)}{p(H_{i1}^{(0)}=\bar{h}|H_{i0}^{(0)}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)},\end{aligned}$$

where $(\bar{h},h)$ is the pair of consecutive latent states. The model relies on: (*i*) stable unit treatment value assumption, meaning that there is no interference between individuals; (*ii*) exogeneity of the covariates in $\mathbf{X}_i$ that are not influenced by the treatment; (*iii*) no effect of treatment at $t=0$; (*iv*) common support, since every individual has a positive probability of receiving any type of treatment; and (*v*) common trend.

Under the above assumptions, the ATET is identified on the basis of the latent variables H_{it}:

$$\begin{aligned}\mathrm{ATET}_{1\bar{h}h}(\mathbf{x}) = &\log\frac{p(H_{i1}=h|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)}{p(H_{i1}=\bar{h}|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=1)}\\ &-\log\frac{p(H_{i1}=h|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=0)}{p(H_{i1}=\bar{h}|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=0)}.\end{aligned}$$

We estimate the ATET by a reduced form HM model with covariates as:

$$\log\frac{p(H_{i0}=h|\mathbf{X}_i=\mathbf{x},Z_i=z_i)}{p(H_{i0}=1|\mathbf{X}_i=\mathbf{x},Z_i=z_i)} = \beta_{0h}^{(0)}+z_i\bar{\beta}_{0h}+\mathbf{x}'\beta_{1h},\ \ h=2,\dots,k, \tag{1}$$

$$\log\frac{p(H_{i1}=h|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=z_i)}{p(H_{i1}=\bar{h}|H_{i0}=\bar{h},\mathbf{X}_i=\mathbf{x},Z_i=z_i)} = \gamma_{0\bar{h}h}^{(0)}+z_i\bar{\gamma}_{0\bar{h}h}+\mathbf{x}'\gamma_{1\bar{h}h},\quad \bar{h},h=1,\dots,k,\ h\neq\bar{h}, \tag{2}$$

where $\mathrm{ATET}_{1\bar{h}h}(\mathbf{x})$ directly corresponds to $\bar{\beta}_{0h}$ at the initial period and $\bar{\gamma}_{0\bar{h}h}$ in the second time occasion for the transition between states, indeed constant with respect to the covariates.

Estimation is carried out on the basis of the maximum likelihood approach through the EM algorithm [3], where the manifest distribution of the observed responses is computed through suitable recursions [1]. Standard errors for the parameter estimates may be obtained with different methods: by computing an approximation of the information matrix or through bootstrap procedures. Model selection is performed with the Bayesian Information Criterion (BIC).

3 Results

The CLT model is estimated accounting for missing values in the covariates through dummies as missing indicators. The BIC index obtained for an increasing number of latent states ranging from 1 to 4 leads to selecting a model with two latent states. According to the estimated conditional means reported in Table 1 two latent subpopulations of students are identified clustered in low (1st subpopulation) and high cognitive levels (2nd subpopulation) both in Italian and Mathematics. At the 6th grade, the average probability of belonging to the 2nd subpopulation is 0.352. We notice that the national average of the test scores measured on the Rasch scale for each grade is fixed at 200. Therefore, a student in the 2nd subpopulation having an average score of around 235 for Italian and 246 for Mathematics is above the national average, and s/he shows a gain of around 40 points on both subjects with respect to a student in the 1st subpopulation. There is a moderate estimated conditional correlation ($\hat{\rho} = 0.381$) between the two scores.

Estimates of the logit regression parameters on the initial probabilities (at the 5th grade) according to Equation (1) are reported in Table 2 for some baseline covariates among those considered in the model. Standard errors are obtained through a nonparametric bootstrap based on 1,000 bootstrap samples. Gender significantly affects the initial distribution of the latent variables: the odds ratio for females versus males to belong to the 2nd subpopulation is 1.587 and discomfort at school negatively affects cognitive abilities: the odds ratio is 0.249.

Estimates of the logit regression parameters on the transition probabilities according to Equation (2) are shown in Table 3 where effect 1 is referred to the probability of moving from the 1st to the 2nd subpopulation and vice-versa for effect 2. A negative and significant ATET is estimated for effect 2. Thus students in the 2nd subpopulation who had taken the program to increase NCSs show a reduced probability of decreasing in CSs from the 6th to the 8th grade. Family background is important and what has been acquired during primary school (Italian and Mathematics scores at the 5th grade) helps to increase the CSs.

Figure 1 shows the estimated average transition matrices of the two state CLT model for treated (a) and non-treated (b) students. We observe that the probability to transit from the 2nd to the 1st subpopulation is higher for students who have not taken the educational programs aimed at improving the NCSs. Standard errors for the estimated probabilities obtained with the nonparametric bootstrap show that all these values are statistically significant.

Table 1: Estimated conditional averages of the two state CLT model

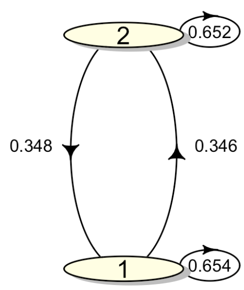
	Latent state (h)	
Scores	1	2
Italian	192.312	235.133
Mathematics	196.071	246.636

Table 2: Estimates of the logit regression parameters of the initial probabilities of the two state CLT model, significant **at 1%

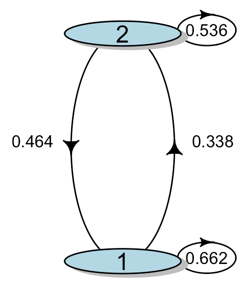
Covariate	Effect	s.e.
Intercept	-3.275**	0.455
Female	0.462**	0.199
Discomfort at school	-1.389**	0.163
Italian nationality of the father	0.904**	0.308
Missing indicator for father's nationality	-0.301	1.101
⋮	⋮	⋮

Table 3: Estimates of the logit regression parameters of the transition probabilities of the two state CLT model, significant †at 10%, **at 1%

Covariates	Effect 1	s.e.	Effect 2	s.e.
Intercept	-43.393**	0.507	45.543**	0.802
Treatment effect (ATET)	0.847	0.814	-3.583**	1.683
Employment status of the father	0.394†	0.236	-0.026	0.426
Italian score at the 5th grade	0.078**	0.011	-0.076**	0.022
Math score at the 5th grade	0.114**	0.010	-0.141**	0.020
⋮	⋮	⋮	⋮	⋮



(a) Treated



(b) Non-treated

Fig. 1: Estimated averaged transition probabilities between latent state 1 and 2 of the CLT model for treated (a) and non-treated (b) students: nodes depict states and arrows transitions

4 Conclusions

In the context of observational studies, evaluating the causal effect of a treatment or policy on individual characteristics, such as students' abilities, is often of interest. A novel approach is proposed in [2] to account for confounding factors with pre- and post-treatment multivariate outcomes. Under a set of assumptions, the causal latent transition model clusters individuals into a finite number of homogenous subpopulations and causal effects are identified and estimated for each subpopulation. The proposal is also compared with some Difference-in-Difference (DiD) methods [5], which are well developed in economic literature. Sensitivity analyses were carried out considering a set of different model specifications and simulations implemented to verify the proposal confirm its validity.

Concerning the evaluation of non-cognitive skills, many studies confirm the importance of personality traits such as openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability to increase cognitive skills. With the causal latent transition model, we infer the positive effects of the programs aimed at developing non-cognitive abilities of pupils in Italian and Mathematics for the subgroup of students belonging to the subpopulation with higher cognitive abilities. Employing a DiD approach, these effects are not detected, and only a positive effect of the treatment on Mathematics is significant under a certain DiD model specification.

The causal latent transition model may be used to estimate the average treatment effects on the treated in many other contexts of interest since it may be simply generalized to the case of more time occasions and more than one type of treatment.

Acknowledgements We greatly acknowledge the Autonomous province of Trento (PAT) and the Italian National Institute for the Evaluation of the Educational System (INVALSI) for providing access to surveys and administrative data.

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Chapman & Hall/CRC Press (2013)
2. Bartolucci, F., Pennoni, F., Vittadini, G.: A causal latent transition model with multivariate outcomes and unobserved heterogeneity: Application to human capital development. J. Educ. Behav. Stat. (2023) doi: 10.3102/10769986221150033
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B 39, 1–38 (1977)
4. Heckman, J.J., Kautz, T.: Hard evidence on soft skills. Labour Econ. 19, 451–464 (2012)
5. Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J Econ Lit 47, 5–86 (2009)
6. Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. J. Am. Stat. Assoc. 100, 322–331 (2005)

Cognitive Skills and Non Cognitive Skills to Analyze School and Students Performances

Skill cognitivi e non cognitivi per valutare scuole e studenti

Giorgio Vittadini

Abstract Defining the Human Capital in terms not only of Cognitive Skills (CSs) but also of Non-Cognitive Skills (NCSs) allows us to study the connection between the students' characteristics, ability to know. Furthermore, it can be verified whether appropriate educational projects can increase NCs and how efficient schools are in transforming NCs into CSs.
Abstract *Definire il Capitale Umano in termini non solo di Cognitive Skills (CS) ma anche di Non Cognitive Skills (NCS) permette di studiare il nesso tra caratteristiche degli studenti e capacità di conoscenza e Inoltre si può verificare se opportuni progetti educativi possono incrementare le NCS e quanto le scuole siano efficienti nel trasformare NCs in CS.*

Key words: Non Cognitive Skills, Cognitive Skills, Causal inference, Efficiency analysis

1 Human Capital: Cognitive Skills and Non Cognitive Skills

A classical approach to the analysis of schooling outcome is the Human Capital (HC) analysis. HC consists of the investment in years of education and vocational training, where the returns on investment are higher expected individual earnings connected to cognitive skills (CS) [1], Recently, HC has been also connected to NCS, a multifaced definition of human character traits. The economist and Nobel Prize James Heckman was the first one to recognize the importance of NCS and quantitatively measure their effects on a wide variety of outcomes, from a child's success in their educational

Giorgio Vittadini
University of Milano-Bicocca, Department of Statistics and Quantitatve Methods; email: giorgio.vittadini@unimib.it

journey and social and economic dimensions to a person's physical and mental health as an adult [2, 3, 6].

Subsection Heading Starting from these previous studies we propose three research questions:
a. Are NCS related to academics CS outcomes?
b. Is it possible to implement educational programs aimed at improving NCS?
c. Are NCS useful to propose a new version of the efficiency school analysis taking into account students' knowledge measured in terms of their CS?

2 Classification of NCS

The database on which Heckman's studies are based (the NLSY79, National Longitudinal Survey of Youth) are extremely rich and of high quality, respect to CS. Regarding NCS indicators, he uses only two indicators (the Rosemberg Self-Esteem Scale and the Rotter Locus of Control Scale and the Temperament Scale and the Behavior Problems Index for 2010) because these measurements were the only ones in the NLSY79 sample.
Instead, we wish to use a more complete set of indicators able to measure the multifaced definition of NCS [4, 9].
- *BIG5* model composed by five personality dimensions*: openness to experience, the tendency to be open to reality; conscientiousness, the propensity to be organized, and hardworking* (Inner stability synthesis of openness and conscientiousness); *extraversion*, the positive orientation toward the outer world of people and things; *agreeableness*, the disposition to act in a cooperative manner; *emotional stability*, the absence of rapid mood changes.
- *Psychological capital*, a positive psychological state able to provide competitive advantage; it includes *self-efficacy* in achieving goals as well as optimism about succeeding [4].
- *Locus of control* being dependent on the judgment of others
- *Learning orientation*, the propensity to increase personal ability.
- *Performance orientation*, the desire to achieve specific goals.
- *Self-regulation of motivation to study the reasons* that induce people to engage in tasks.
Social capital
ESCS socio-economic family index.

3 Statistical Methodology

a) **NCS estimation.** From a statistical point of view, NCS may be interpreted as latent variables underlying observed items obtained from student replies to the NCS survey by means of a confirmatory factor analysis [4].

b) The **influence of NCS on CS** can be investigated by means a generalized least squares (GLS) model [8].

c) The difference-in-differences approach [5] tests whether **educational programs had a causal effect improving NCS** We have the following equations concerning times $t=0,1$

$$NCS_{it} = k+mF+pX_i+u_{it} \qquad i=1,...,n$$

where NCS_{it}–NCS scores ($t=0,1$)**;** F Educational program: $F_{=}0$ for classes with no educational programs $F_{=}1$ for classes with programs**;** X_i invariant covariates; u_{it} random error assumed to have a zero mean and a normal distribution ($t = 0,1$)

d) In classical efficiency analysis, the problem is to maximise monetary school revenues given monetary costs Effective- ness is the capacity of the school to increase student knowledge measured in terms of CS. Instead in our framework, we propose a "Non-Cognitive Skills Efficiency" approach that measures the different school ability of transforming Non-Cognitive Skills into Cognitive Skills, by means of a Stochastic Frontier Approach (SFA) [8]. We have

$$y_{wj} = \alpha+\lambda k_j+\sum \beta\, x_{hwjj} - u_{wj} + v_{wj}$$

with $w = (1, ..., k_j)$ classes $j = (1, ..., n)$ schools; y_{wj} average CS values for the w-th class of the j-th school; α intercept; k_j school effect; x_{wj} average w-th class of the j-th school value of NCS or control variables; u_{wj} time-invariant stochastic inefficiency of the w-th class of the j-th school with U_σ variance expressed in terms of natural logarithm; v_{wj} stochastic disturbance with V_σ variance of stochastic disturbance expressed in terms of natural logarithm.

$$\gamma = \frac{\exp(U_\sigma)}{\exp(U_\sigma + V_\sigma)}$$

average relative weight of the time-invariant stochastic inefficiency over the total variance. It shows the importance of time- invariant stochastic inefficiency with respect to stochastic disturbance. We also assume that.

1. The random vectors $u_{wj}+v_{wj}$ are independent in probability;
2. For every w, j, u_{wj} has a half normal distribution with zero expected value and variance s^2 left-truncated at zero;
3. For every w, j, v_{wj} has a normal distribution with zero expected value and variance σ^2.

4 The dataset joining the INVALSI dataset with an NCS survey

The sample analysed consists of 1522 8th grade students attending schools in the Provincia Autonoma di Trento (PAT), in Italy. The sample was collected as follows: during the 2017-2018 school year, IPRASE, the regional authority for research on education and schooling, launched a project to evaluate student NCS. 25 middle schools with a total of 108 classes (out of 77 PAT middle schools and a total of 5502 students) joined the research project on a voluntary basis. 12 schools (with a total of 845 students) carried out curricular educational programmes aimed at improving student NCS. The integrated dataset consisted of five datasets, which were matched by IPRASE and INVALSI.

5 Results

This section presents the results obtained from the analysis of the relationship between eighth graders' CS (first survey, May 2018) and their NCS. The dependent variables, measuring the CS, were the grades of the 2018 INVALSI tests for Italian and mathematics.

First of all, the R2 of the two models (0.55, 0.56) indicate that the explicative variables satisfactorily explain the results of the INVALSI 2018 tests. Given that:

a) The first evidence concerns the link between NCS and CS. Inner stability and emotional stability are positively linked with the INVALSI 2018 tests while Locus of control is negatively linked.

b) Regarding social capital, the variables related to the increase in NCS, such as reading books, doing homework are positively related to the results of the 2018 INVALSI tests. Helping at home, not directly related to studying, has a negative relationship with the 2018 INVALSI tests.

c) The 2018 INVALSI test scores in the eighth grade in both Italian and mathematics are positively correlated with the corresponding tests obtained by students in the fifth grade in 2015. This highlighted how primary education is crucial to the acquisition of CS [6].

d)Variables regarding Kindergarten, full-time, and the prospect of embarking on a challenging educational path are positively related to the 2018 INVALSI results, describing a decisive and important investment in the school path. [6].

e) With regard to socioeconomic aspects, the ESCS index is significant demonstrating that initial inequalities are not overcome by the education system

The educational interventions improving NCS have positive causal effects on optimism and emotional stability [9].

The variance of time-invariant stochastic inefficiency parameters is significant, and the γ relative weights have good values for both CS Italian and Math 2018. This suggests that there are considerable differences among schools in terms of inefficiency measured in terms of capacity of transforming NCS in CS. Above all, these differences depend on school diverse ability of utilizing Inner Stability. Emotional Stability, Optimism to improve students' knowledge, see Table 1 for thr efficiency analysis and Table 2 for the SFA analysis.

Table 1 Efficiency

Treatment	*Treatment*	*R-squared*
Optimism	0.097***	0.364
Emotional stability	0.209***	0.181
Performance orientation	1.790***	0.253
Obs	1,521	1,521

*** *p<0.0001*

Table 2 NC CS analysis

Variables	*Italian 2018*	*Mathematics 2018*
INVALSI Italian 2015	0.340***	0.114***
INVALSI math 2015	0.157***	0.457***
Inner stability	12.10***	10.49***
Emotional stability	3.452***	2.109***
Locus of control	-4.860***	-3.872***
Playing with friends		-2.112***
Helping at home	-3.380***	-2.528***
Reading a book	2.945***	
Doing homework	2.038***	2.654***
ESCS	1.746***	2.050***
High school	4.155***	4.606***
Full time	3.466***	5.147***
Obs	1,521	1,521
Number of schools	25	25
R^2	0.5498	0.5652
Wald c^2	1806.48	1922.29
	Prob > chi2 = 0.0000	Prob > chi2 = 0.0000

*** *p<0.0001*

6 Conclusions

Our results:
a) confirm the analytical approach introduced by Heckman, which indicated the links between NCS and CS.
b) indicate that it is possible to improve some NCS through educational programs in schools.
c) show that is useful to include both NCS and CS in an efficiency analysis
Our analysis can be improved if more appropriate measures of NCS variables will be available. In Italy, there is already an ongoing debate on how to measure students NCS. Recently, it has been decided to introduce a portfolio for students at the last year of high school, which collects their NCS characteristics; however, the definition of NCS is still more a qualitative list and not yet a quantitative evaluation.

References

1. Becker, G.S.: Investment in human capital: A theoretical analysis. In: Journal of Political Economy 70: 9–49 (1962)
2. Cunha, F., Heckman, J.J.: The Technology of Skill Formation. In: American Economic Review 97: 31-47 (2007)
3. Cunha, F., Heckman, J.J., Schennach, S.M.: Estimating the Technology of Cognitive and Noncognitive Skill Formation. In: Econometrica 78: 883–931 (2010)
4. Fabbris, L., Fornea, M.: Psychological capital and locus of control as determinants of graduate employability beyond human and social capital. In: Statistica Applicata- Ital J Appl Stat 2019;1: 29–50 (2019)
5. Greene, W.: Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. In: J Econom 2005; 126(2): 269-303 (2005)
6. Heckman, J.J., Humphries, J.E., Kautz, T.: The Myth of Achievement Tests: The GED and the Role of Character in American Life. Chicago: University of Chicago Press (2014)
7. Lechner, M.: The Estimation of Causal Effects by Difference-in-Difference Methods. In: Foundations and Trends in Econometrics 4: 165–224 (2010)
8. Vittadini, G., Folloni, G., Sturaro, C.: The Development of Cognitive and Noncognitive Skills in Students in the Autonomous Province of Trento. In: Economies, 10(7), 1-17 (2022)
9. Vittadini, G., Folloni, G., Sturaro, C.: Non-Cognitive Skills and Cognitive Skills to measure school efficiency. In: Socio-Economic Planning Sciences, 81 (2022)

Solicited Session SS18 - *Statistical methods for the assessment of transport services and sustainable emissions*

Organizer and Chair: Antonello D'Ambra

1. *Sustainability assessment of urban transport by an LCA comparison on different technologies vehicles* (Della Ragione L. and Meccariello G.)
2. *Passenger comfort prediction via time-series classification* (Vanacore A., Pellegrino M.S. and Ciardiello A.)
3. *A statistical model to analyse driving behavior: a case study* (Rodia G., Sarnacchiaro P. and Acciarino V.)
4. *Aggregating judgments in non negotiable group decisions in transport system* (Amenta P. and Lucadamo A.)

Sustainability assessment of urban transport by an LCA comparison on different technologies vehicles

Valutazione della sostenibilità del trasporto urbano mediante un confronto LCA su veicoli di diverse tecnologie

Livia Della Ragione and Giovanni Meccariello

Abstract This paper intends to contribute to the environmental assessment in the context of sustainability vehicular transport in urban areas. A possible evaluation is based on a comparative analysis through the LCA Life Cycle Analysis methodology of the entire vehicle life cycle. The comparison may concern couples made up of an electric vehicle and an internal combustion vehicle of the same segment and category. The study will be designed for different vehicle segments to evaluate their efficiency and overall environmental sustainability also in relation to current social and political scenarios.

Abstract *In questo lavoro si intende contribuire alla valutazione ambientale nel contesto della sostenibilità del trasporto urbano delle autovetture. Una possibile valutazione si fonda su un'analisi comparativa attraverso la metodologia LCA Life Cycle Analysis dell'intero ciclo di vita. La comparazione può riguardare coppie composte da un veicolo elettrico ed un veicolo a combustione interna dello stesso segmento e categoria. Lo studio verrà proposto per differenti segmenti di veicoli al fine di valutarne l'efficienza e la sostenibilità ambientale complessiva anche in relazione agli scenari sociali e politici attuali.*

Key words: LCA, Battery Electric Vehicle, Internal Combustion Engine, CO_2 emissions.

Livia Della Ragione
Institute of Sciences and Technologies for Sustainable Energy and Mobility (STEMS) - CNR, e-mail: livia.dellaragione@stems.cnr.it

Giovanni Meccariello
Institute of Sciences and Technologies for Sustainable Energy and Mobility (STEMS) - CNR e-mail: giovanni.meccariello@stems.cnr.it

1 Introduction

The LCA process is a methodology for evaluating the ”environmental footprint” of the product, ”from the cradle to the grave”. It is carried out by quantifying the impacts deriving both from the use of resources (energy, raw materials, water) and from emissions into the environment (into the air, water and soil), in relation to human health, the quality of ecosystem and resource depletion. All the processes that contribute to obtaining the product are analyzed and evaluated, as if they were examined with a magnifying glass, from the production and transport of raw materials to the actual manufacturing process, from distribution on the market and its use up to for disposal or sending to recovery plants. At an international level, the LCA methodology is regulated by the ISO 14040 series standards.

The LCA activity is, therefore, of fundamental importance for the estimation and analysis of the economic and social impacts through the comparative analysis of technological solutions in scenarios of ”accelerated technological evolution” and/or ”sustainable mobility”. The Life Cycle Assessment procedure of a vehicle, according to the integral approach ”from the cradle to the grave”, provides for the identification and collection of multiple information regarding the various components of the product which translate into numerous critical issues within the various phases of the life cycle of the vehicle. In the assessment of the impacts in the production phase it is necessary to consider components that are very different from each other in terms of material and production process, while in the analysis of the use phase it is important to follow the scientific and regulatory progress on the cycles of reference guide for vehicle type approval. The impact must be calculated according to the expected life expressed in units of measurement appropriate to the typical use of the product, i.e. in km for vehicles. It is, therefore, necessary the availability of numerous data and the prescription of hypotheses which are sometimes not general. Primary energy consumption can be re-calculated differently based on the boundary conditions of the calculation itself. Therefore, the overall impact can be divided into [1]:

- Well To Tank (WTT - ”from the well to the tank”), due to the supply of primary energy to the vehicle (e.g. extraction of fossil fuel, production of biofuels, production of electricity, distribution), before this is used by the vehicle itself;
- Tank To Wheels (TTW - ”from the tank to the wheels”), calculated according to the performance of the vehicle after it has been refuelled.

2 Comparative LCA approach

The first step in an LCA is to define the objective and purpose of the LCA. From the point of view of Global Warming Potential (GWP) all the impacts assessed in the context of an LCA are expressed in kg of CO_2 equivalent [2, 3].

Given the comparative character of LCA, the functional unit plays a critical role and must clearly define the functions (performance characteristics) of the system under consideration. Finally, it constitutes the reference to which all input and output study data will be reported, and therefore must be clearly defined and measurable. It should be emphasized that comparisons between systems must be made because of the same function and quantified through the same functional unit. Fig. 1 describes the comparative LCA approach in vehicle production stages, fuel and electricity production, vehicle use and maintenance, as well as vehicle end-of-life, highlighting the comparative focus of the LCA analysis performed in this work.

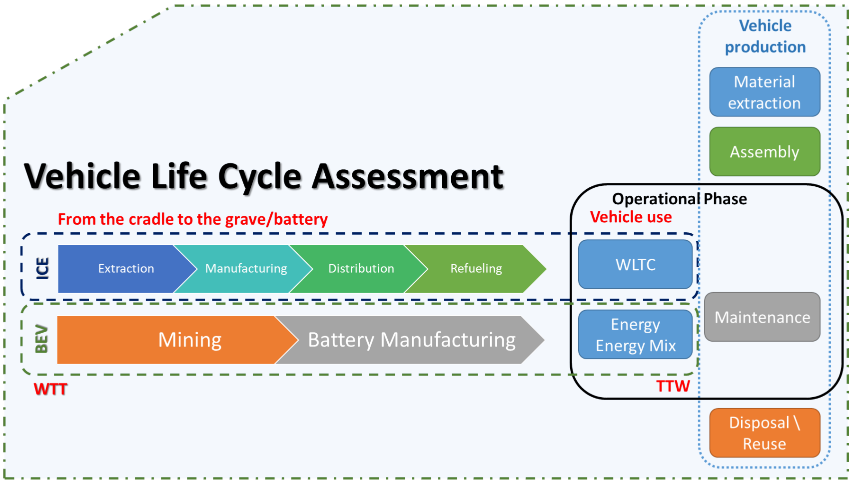


Fig. 1 Schematic representation of the different phases of vehicle LCA

Overall, an estimation model referred to the benchmarking should include the calculation of the equivalent CO_2 emissions of several factors such as the average energy consumption in kWh per km traveled for electric cars, the production and disposal of the vehicle and batteries in the case of electric vehicles, the raw materials needed to produce fuel, energy, cars and batteries.

Finally, in relation to the disposal/recycling phase, some data were considered, even if aggregated data above all on what concerns the batteries. There are few data in the literature to allow a reliable comparison, for the reuse of batteries as accumulators in fixed installations.

To evaluate the production of CO_2 equivalent produced by an electric vehicle, it is important to consider the analysis of the Italian and/or European energy system in terms of energy produced from renewable sources and their evolution with respect to the reference analysis period or scenario expressed in years life of the vehicle or distance travelled in km of the vehicles

3 Result

The comparative LCA activity was carried out through the assessment of scenarios relating to the different factors consolidated in the approach. These scenarios foresee a range of different values for the factors that contribute to the terms related to the use phase, Combustion Mobility/Electric Mobility [4, 5, 6], to the terms related to the extraction of the fuel up to its introduction into the tank WTT, and to the production/replacement of the battery. Thus, a comparative analysis of a scenario with a high economic and investment probability in the current political situation was obtained.

Three comparisons were made considering different segments of vehicle such as small city cars, mid-size and SUV. Elements considered in carrying out the comparison between the couple of vehicles, i.e. the data characterizing the possible reference scenarios, are various factors which constitute the basic information for determining the following aggregates of CO_2 equivalent in g/km for the total LCA of a specific configuration:

- Car Body: for vehicle production.
- Battery: for the production of the first installation battery.
- Combustion Mobility: for emissions based on the WLTP cycle.
- Electric Mobility: for evaluating the energy mix.
- Battery Replacement: for replacing the battery at the end of its life cycle and until the end of the vehicle's life.
- Well to Tank Fuel: for evaluating the extraction component of the fuel up to its introduction into the tank.

The analyzes are carried out considering a duration of 15 years and an average annual mileage of 15,000 km, this means that the observation window will correspond to 225,000 km intended as the total life on which to calculate the CO_2 equivalent value.

The LCA analysis, in terms of CO_2 equivalent production in relation to the kWh used, is made to the Euro/WLTP scenario. For this case we consider a forecast trend of the energy mix to 2030 decreasing by approximately 2.6%, The current electricity production efficiency, used as starting point of the comparison is 320 kg CO_2/kWh.

The acronym BEV (Battery Electric Vehicle technology) stands for vehicle powered by a battery electric motor, without using fossil fuels. For the comparison scenario considered for BEV vehicle must be included in addition to the value of the battery, in terms of "battery capacity" [KWh], initially also considered its replacement during the life period. In this scenario, the expected average life is approximately 160,000 km. Unlike hybrid cars, electric cars have an almost absent polluting impact. In Mild Hybrid engines (MHEV) the electric part supports the combustion engine, improving consumption and performance. In the fourth and fifth rows of the following table 1, the values presented in terms of CO_2 equivalent represent the aggregate both for the vehicle and for the battery production . In the last two rows, values are referred to mobility use, in the purely electric mobility part for the BEV vehicle while for the MHEV vehicle it refers to the values on the WLTP driving cycle which is the reference standard for internal combustion vehicles.

Table 1 Vehicles parameters of the scenarios considered.

Vehicles	FIAT 500e BEV	FIAT 500 MHEV 2020	TESLA MODEL 3 50kwh	AUDI A5 MHEV 2019	KIA E-NIRO 67kwh 2018	KIA SPORTAGE MHEV 2019
Vehicle Segmentation	A	A	D	D	SUV	SUV
Vehicle Electrification	BEV	MHEV	BEV	MHEV	BEV	MHEV
CO_2 eq for the production and disposal of the vehicle	4000	4000	5500	5500	6000	6000
CO_2 emission to produce the battery [kg/kWh]	200	-	200	-	200	-
electric efficiency driving cycle [kWh/100 km]	13.1	-	11.6	-	14.1	-
CO_2 emitted for 100km WLTC cycle [gr/km]	-	124	-	120.7	-	138

In the following table 2, the first row summarizes the LCA overall values calculated referred to the total life observation window (225,000 km) on which the different CO_2 equivalent aggregates are splitted into the remaining lines. Assumed that the BEV vehicle battery must be replaced after 160000 km of useful life, the relative term is present twice in the table. On the second raw, the term Car body is the same for each pair of vehicles in the same sizing of the component and powertrain category. The differences in the phase of mobility of use and in the phase of production and transport of the fuel (fossil or electric) are obviously decisive. On the other hand, the last aggregate (WtTf) must also take into account the presence of existing infrastructure for electric mobility and also the greater or lesser diffusion on the territory of electric charging stations for the BEV vehicle. Overall, for each

Table 2 LCA vehicle comparison.

Vehicles	FIAT 500e BEV	FIAT 500 MHEV 2020	TESLA MODEL 3 50kwh	AUDI A5 MHEV 2019	KIA E-NIRO 67kwh 2018	KIA SPORTAGE MHEV 2019
LCA (gr/km)	125.7	174.3	142.8	171.5	181.8	196.9
Car body	17.8	17.8	24.4	24.4	26.7	26.7
Battery	37.3	0.1	44.4	0.0	59.6	0.4
Combustion mobility	-	124.0	-	120.7	-	138.0
Electric mobility	33.3	-	29.5	-	35.8	-
Battery replacement	37.3	-	44.4	-	59.6	-
Well to Tank fuel	-	32.4	-	26.4	-	31.8

pair of vehicle category, it is possible to evaluate the percentage variations, in terms of efficiency of a BEV vehicle compared to an ICE, equal to 28%, 17% and 8%.

4 Conclusion

This article presents comparative analyzes between vehicles of different categories to present results in terms of LCA. These analyzes could support political and economic decisions scenario, in terms of infrastructure investments and policies to support the diffusion powertrain electric. Moreover, above all it is fundamentally to take into account the cost in terms of CO_2 equivalent, i.e. the environmental sustainability.

BEV and especially MHEV Hybrid electric vehicle have known a quickly grow in the last 10 years. Between conventional vehicles which are criticized for their CO_2 emission and electric vehicles which have a big issue about autonomy, hybrid electric ones seem to be a good trade of. In any case, the evaluations are still questionable, since both the costs for the production and disposal of raw materials to produce the batteries, and the diffusion of electric charging columns by number of circulating BEV vehicles are still today on discussion. Furthermore, research in the field of internal combustion engines is still ongoing and new technologies may be competitive, at least for some vehicle segments, with hybrid engines. In any case, from the point of view of environmental sustainability and noise reduction, hybrid vehicles perform better in terms of CO_2 equivalent, but you must always consider the weight of the vehicle and its size.

References

1. Attributional and consequential LCA: Methodology overview, assessment and recommendations focusing on the JEC Well-to-Tank and Well-to-Wheel reports; https://www.eucar.be/lca-in-wtt-and-wtw-review-and-recommendations/
2. European Commission - Joint Research Centre - Institute for Environment and Sustainability: International Reference Life Cycle Data System (ILCD) Handbook - General guide for Life Cycle Assessment - Detailed guidance. First edition March 2010. EUR 24708 EN. Luxembourg. Publications Office of the European Union (2010)
3. ENEA - Analisi trimestrale del sistema energetico italiano, III trimestre 2021; 4/2020. https://www.pubblicazioni.enea.it/
4. Samsu Koroma M., Costa D., Philippot M., Cardellini G., Sazzad Hosen Md, Coosemans T., Messagie M.: Life cycle assessment of battery electric vehicles: Implications of future electricity mix and different battery end-of-life management, Science of The Total Environment, Volume 831, 154859, ISSN 0048-9697, (2022) https://doi.org/10.1016/j.scitotenv.2022.154859
5. Hall D., Lutsey N.: Effects of battery manufacturing on electric vehicle life-cycle greenhouse gas emissions. ICCT The International Council on Clean Trasportation (2018)
6. Kawamoto R., Mochizuki H., Moriguchi Y., Nakano T., Motohashi M., Sakai Y., Inaba A.: Estimation of CO_2 Emissions of Internal Combustion Engine Vehicle and Battery Electric Vehicle Using LCA. Sustainability, 11, 2690 (2019) doi:10.3390/su11092690

Passenger comfort prediction via time-series classification

Previsione del comfort del passeggero attraverso la classificazione di serie temporali

Amalia Vanacore, Maria Sole Pellegrino and Armando Ciardiello

Abstract The seat comfort experience during a flight is one of the most crucial factor impacting the passenger's intention of flying with the same airline in future occasions, therefore an effective strategy to increase passenger satisfaction cannot be separated from improving the comfort of the seat. Due to its cost-effectiveness, the analysis of the pressure distribution at the seat-occupant interface is one of the most widely used methods in the diagnostic evaluation of seat (dis-)comfort. Dealing with the research issue of predicting subjective (dis-)comfort via seat pressure measurements, this study aims to identify the seat pressure index that best predicts seat (dis-)comfort by investigating five different pressure indexes via eleven time series classification algorithms.
Abstract *Il comfort del sedile è uno dei fattori che condiziona maggiormente la scelta del passeggero di volare nuovamente con la stessa compagnia aerea, pertanto un'efficace strategia per aumentare la soddisfazione dei passeggeri non può prescindere dal migliorare il comfort della seduta. L'analisi della distribuzione della pressione all'interfaccia sedile-occupante è uno dei metodi più utilizzati nella valutazione diagnostica del (dis-)comfort del sedile. Questo studio è volto ad identificare l'indice di pressione che meglio predice la sensazione di (dis-)comfort, analizzando cinque indici di pressione attraverso undici algoritmi di classificazione di serie temporali.*

Key words: Seat (dis-)comfort, Pressure indexes, Time Series Classification algorithms

---

Amalia Vanacore
University of Naples Federico II; p.le Tecchio 80, 80125 Naples, Italy e-mail: `amalia.vanacore@unina.it`

Maria Sole Pellegrino
University of Naples Federico II; p.le Tecchio 80, 80125 Naples, Italy

Armando Ciardiello
University of Naples Federico II; p.le Tecchio 80, 80125 Naples, Italy

1 Introduction

Increasing competition among airlines has prompted them to make greater efforts to improve the passenger comfort experience: improving the feeling of comfort associated with a flight produces an increase of the proportion of passengers who wish to use the same airline on future occasions. Passengers spent most of the time sitting, thus seat comfort has a great impact on the overall passenger experience and, as a consequence, seat comfort assessment can provide useful information to improve passenger satisfaction. Two types of approaches can be discerned to evaluate seat comfort: subjective methods (i.e. directly asking passenger how comfortable her/his seat is) and objective methods (i.e. based on pressure distribution or posture analysis). Seat comfort assessment based on subjective evaluations is costly and time-consuming, whereas objective methods represent a more cost-effective alternative. Among the objective methods, pressure distribution at seat-occupant interface is undoubtedly the most widely used in seat industry, because it is less invasive and provides easily interpretable information.

The relationship between seat pressure measures and (dis-)comfort perception is worth investigating since finding an effective strategy to predict (dis-)comfort from seat-occupant interface pressure measurements would enable two benefits: to overcome the criticisms of the conventional methods based on surveys and jury evaluations [2]; and to take advantages from intelligent technologies (i.e. embedded sensors) able to modify in real time seat characteristics so as to improve passenger's comfort [3].

The problem addressed in this study is identifying the seat pressure index that best predicts seat (dis-)comfort. For the problem at hand, five different pressure indexes have been investigated via eleven Time Series Classification (TSC) algorithms. The paper is articulated as follows: a brief description of the adopted TSC algorithms and seat pressure indexes are provided in Sections 2 and 3, respectively; Section 4 is devoted to data analysis and conclusions are drawn in Section 5.

2 Time Series Classification algorithms

Over recent years, several Time Series Classification (TSC) algorithms have been proposed [4, 5, 6, 7]; they are commonly grouped as follows: *interval based algorithms*, *distance based algorithms*, *shapelet based algorithms*, *dictionary based algorithms*, *hybrid algorithms*, *kernel based algorithms*, and *deep learning based algorithms*. In our study, a total of 11 algorithms has been selected to represent each category: Time Series Forest (TSF) [8], Random Interval Spectral Ensemble (RISE) [10] and Canonical Interval Forest (CIF) [9] classifiers as representative of interval based TSC algorithms; Dynamic Time Warping (DTW) classifier [11] as representative of distance based TSC algorithms; Shapelet Transform Classifier (STC) [12] for the shapelet based TSC algorithms; Bag Of Symbolic Fourier approximation Symbols (BOSS) [13] and Multivariate Unsupervised Symbols and dErivatives (MUSE) classifiers [17] as representative of the group of dictionary based TSC algorithms; Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE) classifier [14] as representative of the hybrid TSC algorithms; Random Convolutional Kernel Transform (ROCKET) classifier [15] as representative of the group of kernel based TSC algorithms; finally, Residual Neural Network (ResNet) and Long Short-Term Memory (LSTM) classifiers [6, 7] as representative of deep learning based TSC algorithms.

3 Seat pressure indexes

The pressure distribution at the seat-occupant interface can be described by means of the following synthetic pressure indexes: mean pressure (i.e. the average of the pressure values over the contact area), pressure peak (i.e. the maximum pressure value over the contact area), seat pressure distribution (SPD) that is a measure of the homogeneity of pressure distribution [18], and comfort loss [19] with respect to an ideal pressure distribution which can be represented as an even load distribution over the seat (CL1) or an uneven load distribution (CL2) to take into account that human body areas show different sensitivity for seat pressure.

4 Data analysis

The data were collected during 115 laboratory test sessions for the comparative (dis-)comfort assessment of aircraft seats. The participants were selected according to pre-defined anthropometric criteria (i.e. age, gender, height, BMI) so as to represent a wide range of the passenger population. On the other hand, the seats under comparison differed for the shape and dimension of seat pan, backrest, armrest, headrest, and for the presence of reclining so as to represent a wide range of seat conditions producing different seat pressure distributions and different (dis-)comfort feelings. In each test session, passenger-seat interface pressures were collected for 40 minutes using a Tekscan (South Boston, MA, USA) pressure mat with 1024 resistive sensors; at the end of the session, the participant was asked to provide her/his subjective discomfort evaluation (viz. no discomfort, light discomfort, moderate discomfort, strong discomfort). For each test session, 120 temporally equidistant pressure matrices were analysed via the pressure indexes under comparison (viz. mean pressure, peak pressure, SPD, CL1, CL2) so as to obtain five data sets, each consisting of 115 rows (i.e. test sessions) and 121 columns (i.e. time series of length 120 for the specific pressure index and the associated discomfort label).

The time series data were classified into the aforementioned discomfort classes via the TSC algorithms described in Section 2 to investigate which pressure index yields the greatest effectiveness in classifying seat discomfort. The experiment was implemented in Python using *sktime* [16]. For each algorithm-data set combination, a repeated $k-$fold cross validation with $r = 10$ repetitions and $k = 3$ folds was performed, so 30 stratified resamples were taken maintaining the skewed original class distribution, that is 12, 45, 35 and 23 sessions belonging to class no discomfort, light discomfort, moderate discomfort, and strong discomfort, respectively.

The set configurations used for each TSC algorithm are the same adopted for the comparison proposed in [4], which, at the best of our knowledge, is the most comprehensive and complete review on TSC algorithms.

Table 1 Mean value of performance measures based on $R = 30$ stratified resamples

		Mean	**Peak**	**SPD**	**CL1**	**CL2**
Precision	Min	0.49	0.57	0.57	0.48	0.50
	Max	0.63	0.62	0.65	0.66	0.62
$\boldsymbol{F_1}$**−score**	Min	0.56	0.60	0.61	0.55	0.56
	Max	0.64	0.62	0.65	0.65	0.63
Balanced Accuracy	Min	0.61	0.61	0.63	0.60	0.61
	Max	0.68	0.67	0.69	0.67	0.67
Balanced $\boldsymbol{AC_2}$	Min	0.57	0.57	0.60	0.56	0.57
	Max	0.65	0.64	0.66	0.64	0.64

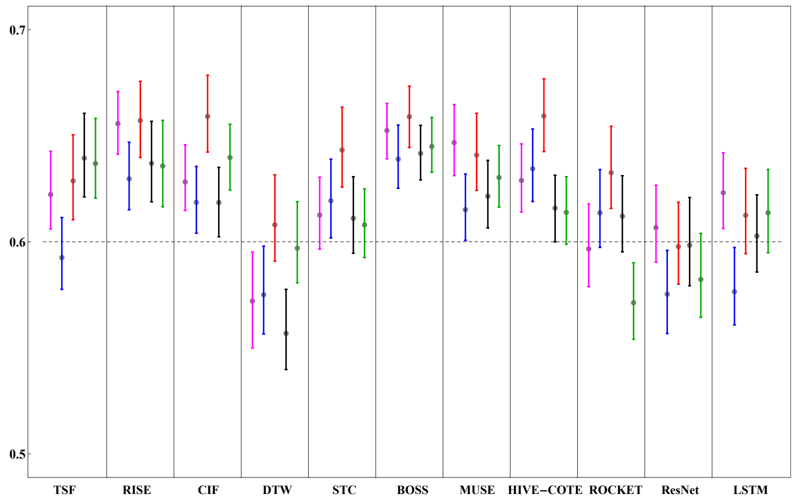


Fig. 1 Mean value and percentile bootstrap CI of the predictive performance of all TSC algorithms estimated via Balanced AC_2 for Mean Pressure (in magenta), Peak Pressure (in blue), SPD (in red), CL1 (in black), and CL2 (in green)

The predictive performance has been estimated via the linearly weighted version of Precision, F_1-score, Balanced Accuracy [20] and Balanced AC_2 [1]. The mean predictive performance over the 30 stratified resamples has been assessed for each TSC algorithm and the minimum and maximum mean value for each performance measure are reported in Table 1. Moreover, since Balanced AC_2 is able to handle data imbalances while accounting for predicted classifications matching the actual class by chance alone, the percentile bootstrap confidence intervals (CIs) of the Balanced AC_2 have been built to further investigate the predictive performance comparison, and reported in Figure 1 for each combination of TSC algorithm and pressure index.

The results in Figure 1 reveal that the predictive performance is comparable across TSC algorithms with the exception of DTW, ROCKET, ResNet and LSTM, for which it is not possible to reject the hypothesis of moderate performance being the lower bounds of percentile CIs lower than 0.6 for most of pressure indexes. Among all, DTW shows a predictive performance significantly lower than the other algorithms for almost all indexes. This result is due to the poor performance of DTW with noisy series, since the noise may overpower subtle shape differences that are useful for class discrimination. The classifications provided by BOSS, RISE, CIF and MUSE can be assumed in substantial agreement with actual classifications being their predictive performance significantly higher than 0.6 for all the pressure indexes. These results are not surprising since BOSS uses Symbolic Fourier Approximation for transforming the signal into words, being thus suitable for time series classification in the presence of noise, like pressure data where small movements detected by pressure sensors embedded in the seats and unrelated to discomfort, as nodding of the head, introduce noise within the data set [13]. RISE and CIF, instead, being interval based algorithms, can remove some of the noise caused by the length of series since they extract diverse and informative descriptive features at different time intervals. Looking at pressure indexes, study results suggest that SPD generally best predicts passenger's discomfort, whereas pressure peak provides the lowest predictive performance.

5 Conclusion

The comparison across the predictive performance of 11 TSC algorithms and 5 pressure indexes reveals that the best performing strategy for seat discomfort prediction is given by the adoption of BOSS, RISE, CIF or HIVE-COTE algorithm with SPD pressure index. The study results can be further exploited to provide new insights in the relationship between seat pressure and discomfort perception by identifying typical pressure patterns characterizing specific discomfort classes. These findings will be of interest to the airline and train industries keen to improve the passenger comfort level under space and weight constraints for seat design.

References

1. Vanacore A., Pellegrino M. S. and Ciardiello, A.: Evaluating classifier predictive performance in multi-class problems with balanced and imbalanced data sets. Qual. Reliab. Eng. Int. 39(2): 651-669 (2023)
2. Kolich M., Seal N. and Taboun S.: Automobile seat comfort prediction: statistical model vs. artificial neural network. Appl. Ergon. 35(3): 275-284 (2004)
3. Cieslak M., Kanarachos S., Blundell M., Diels C., Burnett M. and Baxendale A.: Accurate ride comfort estimation combining accelerometer measurements, anthropometric data and neural networks. Neural Comput. Appl. 32: 8747-8762 (2020)
4. Ruiz A.P., Flynn M., Large J., Middlehurst M. and Bagnall A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Discovery. 35(2): 401-449 (2021)
5. Faouzi J.: Time series classification: A review of algorithms and implementations. Machine Learning (Emerging Trends and Applications) (2022)
6. Ismail Fawaz H., Forestier G., Weber J., Idoumghar L. and Muller P.A.: Deep learning for time series classification: a review. Data Min. Knowl. Discovery. 33(4): 917-963 (2019)
7. Ebrahim S.A., Poshtan J., Jamali S.M. and Ebrahim N.A.: Quantitative and qualitative analysis of time-series classification using deep learning. IEEE Access. 8: 90202-90215 (2020)
8. Deng H., Runger G., Tuv E. and Vladimir M.: A time series forest for classification and feature extraction. Inf. Sci. 239: 142-153 (2013)
9. Middlehurst M., Large J. and Bagnall A.: The canonical interval forest (CIF) classifier for time series classification. 2020 IEEE international conference on big data (big data). 188–195 (2020)
10. Flynn M., Large J. and Bagnall T.: The contract random interval spectral ensemble (c-RISE): the effect of contracting a classifier on accuracy. Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, 2019, Proceedings 14. 381-392 (2019)
11. Senin P.: Dynamic time warping algorithm review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA. 855(40): 1-23 (2008)
12. Lines J., Davis L.M., Hills J. and Bagnall A.: A shapelet transform for time series classification. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 289-297 (2012)
13. Schäfer P.: The BOSS is concerned with time series classification in the presence of noise. Data Min. Knowl. Discovery. 29: 1505-1530 (2015)
14. Lines J., Taylor S. and Bagnall A.: Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles. ACM Trans. Knowl. Discovery Data. 12(5) (2018)
15. Dempster A., Petitjean F. and Webb G.I.: ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min. Knowl. Discovery. 34(5): 1454-1495 (2020)
16. Löning M., Bagnall A., Ganesh S., Kazakov V., Lines J. and Király F.J.: sktime: A unified interface for machine learning with time series. arXiv preprint arXiv:1909.07872 (2019)
17. Schäfer P. and Leser U.: Multivariate time series classification with WEASEL+ MUSE. arXiv preprint arXiv:1711.11343 (2017)
18. Ahmadian M., Seigler T.M., Clapper D. and Sprouse A.: Alternative test methods for long term dynamic effects of vehicle seats. SAE Trans. 684-692 (2002)
19. Vanacore A., Lanzotti A., Percuoco C., Capasso A. and Vitolo B.: Design and analysis of comparative experiments to assess the (dis-) comfort of aircraft seating. Appl. Ergon.76: 155-163 (2019)
20. Sokolova M. and Lapalme G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45(4):427-437 (2009)

A statistical model to analyse driving behavior: a case study

Un modello statistico per analizzare il comportamento alla guida: Un caso studio

Guido Rodia, Pasquale Sarnacchiaro and Vincenzo Acciarino

Abstract Nowadays one of the highest causes of road accidents is the use of mobile phones while driving. The goal of this study is to investigate all together knowledge, attitudes, and behaviour toward using a mobile phone behind the wheel. The data acquired from 774 questionnaires collected in Naples (Italy) revealed that 69 % have used their mobile phone while driving at least once in their lifetime. The results indicate that cell phone usage while driving is common in the study population. The statistical analysis shows how knowledge is not correlated to the behavior held. On the contrary, attitudes are strongly correlated to knowledge and behavior, meaning that good attitudes bring forth positive behavior.

Abstract *Attualmente una delle maggiori cause di incidenti stradali è l'uso del cellulare durante la guida. L'obiettivo di questo studio è quello di indagare sulle conoscenze, le attitudini e i comportamenti di coloro che guidano l'auto e decidono di utilizzare il telefono cellulare durante la guida. A tal fine sono stati raccolti 774 questionari somministrati a guidatori residenti nella città di Napoli (Italia). La ricerca ha evidenziato che il 69% ha utilizzato il cellulare durante la guida almeno una volta nella vita. I risultati ottenuti attraverso l'utilizzo di modelli statistici hanno mostrato come la conoscenza non sia correlata al comportamento tenuto. Al contrario, gli atteggiamenti sono risultati fortemente correlati alla conoscenza e al*

Pasquale Sarnacchiaro

Università of Naples Federico II – Department Economia, Management e Istituzioni, e-mail: sarnacch@unina.it

Guido Rodia

Consorzio Unico Campania, e-mail: rodia@unicocampania.it

Vincenzo Acciarino

e-mail: ditarant@unina.it

comportamento, nel senso che buoni atteggiamenti producono comportamenti positivi.

Key words: Mortality rate, KAP Survey Model, Research method, Survey-based study, Multiple Linear Regression Analysis (MLRA), Dependent and independent variables

1 Introduction

Traffic accidents are a leading cause of death globally, particularly among individuals aged 5-29 years. In Italy, road fatalities decreased by 1.6% in 2018 compared to the previous year, with 3,325 recorded deaths. The country's mortality rate stands at 5.5 deaths per 100,000 population. During the first half of 2019, there were 82,048 road accidents resulting in injuries and 1,505 victims, with a mortality index of 1.8 [2]. Notably, road accidents have a significant economic impact, costing Italy 17 billion euros in social costs in 2018, equivalent to 1% of the country's GDP (ISTAT Press Release, 2019).
Risk Factors: Various factors contribute to the increased risk of road traffic accidents and resulting injuries or deaths worldwide. These include driving under the influence of alcohol or drugs, failure to use safety devices (such as helmets, seat belts, and child restraints), and distractions like mobile phone usage (WHO Global Status Report on Road Safety, 2018). In Italy, distraction, especially from mobile phone use, was the primary cause (16.3%) of road crashes in 2018, surpassing speeding (10.2%), alcohol-related DUI (3.9%), and drug-related DUI (3.2%) [1].
Objective: This study examines Italian drivers' behaviors related to mobile phone use while driving, as well as the frequency and level of their mobile phone involvement. The aim is to analyze their knowledge, attitudes, and behavior concerning mobile phone use while driving in a large metropolitan area of Italy. Understanding these factors can help identify determinants and develop strategies to raise public awareness and promote appropriate driving behavior [3].

2 Theoretical Framework

Behavior is influenced by knowledge, attitudes, and their interplay, as shown in literature research. The KAP Survey Model (Knowledge, Attitudes, and Practices, 2011) uncovers an individual's fundamental attributes in terms of knowledge, attitude, and behavior, while also exploring their ideas on the subject. By employing this model, the aim is to assess a phenomenon using questionnaires, collecting quantitative data, and statistically analyzing the gathered information. The advantage of a KAP survey lies in its ability to gather substantial data in a single survey, which can later be subjected to statistical analysis [6]. Given the rise in mobile phone use while driving and the resulting road fatalities, it is crucial to investigate knowledge, attitudes, and their influence on behavior. This research can facilitate the development of health education and community-based interventions to promote knowledge and foster positive attitudes.

3 Research Method

This study employed a cross-sectional survey conducted between June 2019 and January 2020 in the metropolitan city of Naples, Italy. The survey targeted adults residing in the area who possessed a driver's license and a smartphone. The questionnaire collected information on participants' demographics, such as age, gender, education level, profession, years of driving experience, and smoking habits. It also included three sections focusing on knowledge, attitudes, and behaviors related to mobile phone use while driving, comprising a total of 40 questions.
Data analysis utilized IBM SPSS (vers. 22) statistical software. Descriptive statistics summarized the participants' characteristics, and Multiple Linear Regression Analysis (MLRA) was conducted to predict the dependent variables. MLRA allows for the modeling of linear relationships between independent variables and the outcomes of interest. Three MLRA models were developed to analyze knowledge, attitudes, and actual behavior regarding mobile phone use while driving. Independent variables included sex, age, education level, years of driving license, type of vehicle driven, and smoking status. Knowledge was added as an independent variable in the attitudes model, and both knowledge and attitudes were included in the behavior model [5].
The study population demonstrated good knowledge and positive attitudes towards mobile phone use while driving. Participants generally agreed that using a mobile phone while driving was unacceptable, although their behaviors contradicted this belief and were knowingly inappropriate according to Italian laws. The results indicated an inverse association between education level, driving experience, and the examined behaviors, suggesting the need for targeted educational programs, community-based interventions, and legal regulations. Additionally, addressing individuals' behavior directly is crucial for improving driving habits [4].

References

1. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. 50 (2), 179–211 (1991)
2. Al-Jasser, F.S., Mohamed, A.G., Choudry, A., Youssef, R.M.: Mobile phone use while driving and the risk of collision: a study among preparatory year students at King Saud University, Riyadh, Saudi Arabia. J. Fam. Commun. Med. 25 (2), 102–107 (2018)
3. Alraeesi, F.H., Farzin, F.J., Abdouli, K.A., Sherif, F.Y., Almarzooqi, K.A., AlAbdool, N.H.: Smoking behavior, knowledge, attitude, and practice among patients attending primary healthcare clinics in Dubai, United Arab Emirates. J. Family Med. Prim. Care 9 (1), 315–320 (2020)
4. Arvin, R., Khademi, M., Razi-Ardakani, H.: Study on mobile phone use while driving in a sample of Iranian drivers. Int. J. Inj. Contr. Saf. Promot. 24, 256–262 (2017)
5. Baig M., Gazzaz J.Z., Atta H., Alyaseen M.A., Albagshe A.J., Alattallah H.G.: Prevalence and attitude of university students towards mobile phone use while driving in Jeddah, Saudi Arabia. Int. J. Inj. Contr. Saf. Promot. 25 (4), 372–377 (2018)
6. Bianchi, A., Phillips, J.G.: Psychological predictors of problem mobile phone use. Cyberpsychol. Behav. 8 (1), 39–51 (2005)

Aggregating judgments in non negotiable group decisions in transport systems

Aggregazione dei giudizi nelle decisioni di gruppo non negoziabili nei sistemi di trasporto

Pietro Amenta and Antonio Lucadamo

Abstract The work aims to assess user preferences on four different bus lines connecting Benevento to Rome. For this purpose, pairwise comparison methods were used. In fact, passengers were asked to compare possible alternatives of choice, based on three service-related criteria. In order to understand which bus line was preferred by the users, an aggregation method was used to identify the overall priority vector, thanks to the weights assigned to each decision maker, based on congruence with the judgments of all other users.
Abstract *Il lavoro ha l'obiettivo di valutare le preferenze degli utenti su quattro diverse linee di trasporto che collegano Benevento a Roma. A tal fine sono stati utilizzati i metodi di confronto a coppie. Ai passeggeri è stato infatti chiesto di confrontare le possibili alternative di scelta, sulla base di tre criteri legati al servizio. Per poter comprendere quale linea di autobus fosse preferita dagli utenti, è stato utilizzato un metodo di aggregazione che consente di individuare il vettore di priorità complessivo grazie ai pesi che vengono assegnati ad ogni decisore, sulla base della congruenza con i giudizi di tutti gli altri utenti.*

Key words: Decision analysis, Analytic Hierarchy Process, Optimal Weights, Transport system

1 Introduction

The choice of means of transport is one of the problems often studied in the context of both passenger travel and the transport of goods. In our work the focus is on

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

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

the evaluation of passenger preferences with regard to public transport services. In particular, we intend to understand what citizen preferences are with respect to four bus lines connecting the city of Benevento to the city of Rome.

A questionnaire was administered to detect passenger choice preferences, asking them to compare alternatives on the basis of three possible criteria. Taking into account the type of problem, the most appropriate method to approach the study is the analytic hierarchy process (AHP). It is a procedure used to analyse a multicriterion problem and to rank the alternatives by means of pairwise comparisons [7]. The eigenvector method and the singular value decomposition are the most commonly proposed methods in the AHP literature for deriving priority vectors [5]. When decision problems involve a group of decision makers is necessary to synthesize the individual judgments to obtain the group preferences.

Many authors have studied the AHP for group decisions and proposed different methods for aggregating the priorities expressed by group members [1, 3, 4, 6].

The most common methods applied in the literature for aggregating individual judgments in the AHP are the weighted arithmetic mean method and the weighted geometric mean method [4].

Regardless of the aggregation method, a critical problem in aggregating individual judgments is to derive the weights of the decision makers (DMs) when negotiation is not possible.

The simplest case is to assume that they have equal importance, but this produces two issues: the first concerns the mathematical disadvantage of the arithmetic mean method, which does not satisfy the property of robustness; the second is the pragmatic inconvenience that most of the real cases require applying different weights because each DM may have experience and competence in only a part of the problem. Who can assign these weights? They can be assigned by a manager of the decision making, even if the group members may not agree with this manager, or by means of a participatory method. The weights could be determined by the group members themselves, but this necessitates that the DMs interact with each other. Moreover, Ramanathan and Ganesh [6] propose that the group determine the weights by using their subjective opinions in comparing themselves to each other. This implies that the DMs are not anonymous, but rather get to know each other. In many real situations, as in our problem, there is a panel of anonymous DMs, so the above methods appear useless.

For this reason, in this paper we consider a new formal method, proposed by Amenta et al. [2], to compute a suitable set of coefficients for aggregating individual judgments in a common group preference matrix.

2 The procedure

Let consider the following basic notation:

- $\mathbf{A}_k = ({}^k a_{ij})$ is the $n \times n$ positive reciprocal matrix, with ${}^k a_{ij} = 1/{}^k a_{ji}$ for $i, j \in \{1, \ldots, n\}$. ${}^k a_{ij}$ represents the judgment expressed by the kth decision maker when comparing the ith with the jth alternative.
- The entries of the matrix are expressed by using Saaty's fundamental scale $\{1/9, 1/8, \ldots, 1/2, 1, 2, \ldots, 8, 9\}$.
- $\mathbf{X}_k = \left({}^k x_{ij}\right)$; ${}^k x_{ij} = \log({}^k a_{ij})$; ${}^k x_{ji} = \log({}^k a_{ji}) = \log(1/{}^k a_{ij}) = -\log({}^k a_{ij}) = -{}^k x_{ij}$. This converts the discrete values of Saaty's scale to continuous values, with 0 as the value for a judgement of indifference.
- ${}_{ca}A = ({}_{ca}a_{ij})$: Arithmetic Synthesis Judgment Matrix;
- ${}_{cg}A = ({}_{cg}a_{ij})$: Geometric Synthesis Judgment Matrix;
- ${}_{ca}a_{ij} = \sum_{k=1}^{K} \alpha_k ({}^k a_{ij})$
- ${}_{cg}a_{ij} = \prod_{k=1}^{K} {}^k a_{ij}^{\alpha_k}$
- ${}^k a_{ij}$: judgment expressed by the kth expert;
- α_k: positive weight of the kth expert in forming the group opinion;
- ${}_{cg}\tilde{\mathbf{A}} = ({}_{cg}\tilde{a}_{ij})$
- ${}_{cg}\tilde{a}_{ij} = \log({}_{cg}a_{ij}) = \log(\prod_{k=1}^{K} a_{ij}^{\alpha_k})$.

Amenta et al. [2] proposed to use the Frobenius norm to define the weights to assign to decision makers, considering the following criterion:

$$\begin{cases} \max\limits_{\alpha_k} \|{}_{cg}\tilde{\mathbf{A}}\|_F^2 = \max\limits_{\alpha_k} \|{}_{ca}\mathbf{X}\|_F^2 \\ \text{subject to} \sum_{k=1}^{K} \alpha_k^2 = 1 \end{cases} \tag{1}$$

The coefficient α_k reflects the positive or negative role played by the kth expert, in terms of globally conflicting judgments, in forming the group opinion.

$$\|{}_{ca}\mathbf{X}\|_F^2 = tr({}_{ca}\mathbf{X}({}_{ca}\mathbf{X})^T) = \sum_{k,k'=1}^{K} \alpha_k \alpha_{k'} tr(\mathbf{X}_k \mathbf{X}_{k'}^T).$$

Let $\mathbf{C} = (c_{kk'})$ be a $K \times K$ Gram matrix, defined by $c_{kk'} = tr(\mathbf{X}_k \mathbf{X}_{k'}^T)$ with $|c_{kk'}| \leq \|\mathbf{X}_k\|_F \|\mathbf{X}_{k'}\|_F$

This square matrix contains the Frobenius norms and the numerators of the Tucker's congruence coefficients associated with the DMs, respectively.

The diagonal elements of the matrix represent the measures of the departure of each DM from the indifference condition in his/her evaluations, and the off-diagonal elements represent the measures of compatibility between all possible pairs of DMs. The choice of the coefficients α_k is based on all these measures simultaneously. The criterion can be rewritten according to the maximization of the quadratic form $\mathbf{a}^T\mathbf{C}\mathbf{a}$

$$\begin{cases} \max\limits_{\alpha_k} \sum_{k,k'=1}^{K} \alpha_k \alpha_{k'} c_{kk'} \\ \text{subject to} \sum_{k=1}^{K} \alpha_k^2 = 1 \end{cases} \text{or} \quad \begin{cases} \max\limits_{\mathbf{a}} \mathbf{a}^T\mathbf{C}\mathbf{a} \\ \|\mathbf{a}\|_2^2 = 1 \end{cases} \tag{2}$$

with $\mathbf{a} = (\alpha_1, \ldots, \alpha_K)^T$. The solutions of problem (2) are obtained using Lagrange multipliers. The Lagrangian function is defined as

$$\mathfrak{L}(\mathbf{a}) = \mathbf{a}^T\mathbf{C}\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{a} - 1)$$

where λ is the Lagrange multiplier associated to the constraint $\sum_{k=1}^{K} \alpha_k^2 = 1$.

Differentiating $\mathfrak{L}(\mathbf{a})$ with respect to $\mathbf{a}$ and setting the result equal to zero, we obtain $\mathbf{Ca} = \lambda\mathbf{a}$, showing that $\mathbf{a}$ is an eigenvector of $\mathbf{C}$ and so $\mathbf{a}^T\mathbf{Ca} = \lambda$. The maximum of $\mathfrak{L}(\mathbf{a})$ is then reached when $\mathbf{a}$ is the eigenvector of $\mathbf{C}$ associated to the greatest eigenvalue λ. The choice of coefficients α_k is based on the eigen analysis of the structure of the compatibility matrix $\mathbf{C}$ and so on the closeness measures of all possible pairs of DMs. These coefficients are then applied in the determination of the synthesis matrix.

3 Analysis and final remarks

As stated before, the objective of the work is to assess the preferences of passengers who are asked to choose among four bus routes connecting the city of Benevento to Rome. To this end, a survey was conducted, to compare the four choice alternatives, according to three characteristics related to the different companies that provide transportation. Specifically, people were asked to make a judgment, considering pairwise comparison matrices and using Saaty's scale, on the following aspects: price of the ticket, services offered on board and comfort of travel. The users were then asked to evaluate, using the same procedure, the importance of the three criteria so that weights could be assigned to them. Using the new procedure we obtain the weights assigned to each DM and that are then applied to calculate the priority vectors for each criterion.

The results for the three criteria are in table 1.

Table 1 Priority vectors (and ranking) according to the three criteria

Criteria / Companies	Company A	Company B	Company C	Company D
Price	0.0055 (3)	0.9344 (1)	0.0592 (2)	0.0008 (4)
Services	0.0032 (4)	0.7749 (1)	0.2029 (2)	0.0190 (3)
Comfort	0.0788 (3)	0.7540 (1)	0.1495 (2)	0.0177 (4)

It is evident that Company B is the most preferred, followed by Company C. Less clear is the situation regarding the third and fourth positions in the rankings. Company A, for two of the criteria considered has higher scores than D, but in order to assess the overall preference, it is necessary to consider the global priority vector, which takes into account the weights, assigned by the decision makers to the individual criteria.

Actually, we can see in table 2, that the preference ranking for all users is B-C-A-D. Looking at the values, it is clear that the final vector predominantly reflects the one related to the price of the service. It happens because, in the weighting obtained

Table 2 Global priority vector

Priority Vector			
Company A	Company B	Company C	Company D
0.0003	0.9590	0.0406	0.0001

with the pairwise comparison matrices that the decision makers filled in, with reference to individual criteria, it is the one with the greatest impact and thus led to the determination of the final priority vector. The results obtained in this analysis might seem trivial, but this happens because there is a fairly sharp distinction between the companies, and, most importantly, the weight assigned to the price affected the final decision the most. In many other situations, the results associated with individual criteria, may lead to an overall solution that does not necessarily reflect the majority of partial results. However, the aspect to be emphasized, is related to the method used to obtain the weights of the individual decision makers. The suggested procedure has interesting properties, because it provides a solution by optimizing a mathematical criterion; highlights the discordant behaviors of the DMs and introduces the concepts of majority and minority in computing the final priority vector. In addition, the method, can be further enriched by introducing additional information about decision makers. In fact, their characteristics (e.g., belonging to a certain type of users) can be considered as external information in the model [8]. In this way we could calculate an appropriate set of weights to aggregate the judgments into a common group preference matrix. When we have not prior information, we can obtain groups using a previous cluster analysis. Furthermore we could consider also external information about alternatives.

References

1. Aczel, J., Saaty, T.: Procedures for synthesizing ratio judgements. Journal of Mathematical Psychology, 7, 93–102 (1983)
2. Amenta, P., Lucadamo, A., Marcarelli, G.: On the choice of weights for aggregating judgments in non-negotiable AHP group decision making. Eur. J. Oper. Res. 288(1), 294-301 (2021)
3. Dyer, J., Forman, E.: Group decision support with the analytic hierarchy process. Decision Support Systems, 8, 99–124 (1992)
4. Forman, E., Peniwati, K.: Aggregating individual judgements and priorities with the analytic hierarchy process.European Journal of Operational Research, 108, 165–169 (1998)
5. Gass, S., Rapcsak, T.: Singular value decomposition in ahp. European Journal of Operational Research, 154, 73–584 (2004)
6. Ramanathan, R., Ganesh, L.: Group preference aggregation methods employed in ahp: An evaluation and an intrinsic process for deriving members' weightages. European Journal of Operational Research, 79, 249–265 (1994)
7. Saaty, T.L.: The Analytic Hierarchy Process. McGraw Hill, New York (1980)
8. Takane,Y., Shibayama, T.: Principal component analysis with external information on both subject and variables. Psychometrika 56(1), 97-120 (1991)

Solicited Session SS19 - *New advanced statistical methods for data science*

Session of the SIS Group SDS organized by Rosanna Verde
Chair: Silvia Salini

A unified framework for two-dimensional clustering on preference-approvals: an analysis of Eurobarometer data

Una nuova procedura per il clustering bidimensionale di preference-approvals: un'analisi dei dati Eurobarometer.

Alessandro Albano, Mariangela Sciandra and Antonella Plaia

Abstract This paper presents a novel approach to two-dimensional clustering within a preference-approvals framework. The proposed method employs the ordinal unfolding technique to estimate the coordinates of a set of *m* individuals and *n* alternatives in a *p*-dimensional space based on their proximity values. The main contribution of the approach is the production of a graphical representation that simultaneously captures both individuals and alternatives. To explore the implications of the proposed method, we utilize Eurobarometer data poll. Our goal is to identify common patterns of preferences among individuals and alternatives. The analysis reveals the existence of two main clusters of countries with similar preferences, with some countries having mixed or unique preferences.

Abstract *Questo articolo presenta un nuovo approccio al clustering bidimensionale nell'ambito dei preference-approval. Il metodo proposto utilizza la tecnica dell'unfolding per stimare le coordinate di un insieme di m individui e di un insieme di n alternative in uno spazio p-dimensionale basato sui loro valori di prossimità La principale innovazione dell'approccio è la produzione di una rappresentazione grafica che cattura contemporaneamente le relazioni sia tra gli individui che gli oggetti. Viene proposta un'applicazione del metodo a dati reali, provenientei sito Eurobarometer. La nostra analisi ha rivelato l'esistenza di due cluster di paesi con preferenze simili, con alcuni paesi che hanno preferenze miste o uniche.*

Key words: unfolding, non-metric multidimensional scaling, preference-approvals, eurobarometer

Alessandro Albano · Mariangela Sciandra · Antonella Plaia
Department of Economics, Business and Statistics - University of Palermo
e-mail: alessandro.albano@unipa.it e-mail: mariangela.sciandra@unipa.it
e-mail: antonella.plaia@unipa.it

1 Introduction

Preference-approvals data, which captures individuals' preferences over a set of alternatives, is commonly used in social science research to study various phenomena, from voting behaviour to consumer preferences. However, analyzing such data can be challenging due to its high dimensionality and complex structure. This paper introduces a new method for two-dimensional clustering within the framework of preference-approvals. The proposed approach utilizes the ordinal unfolding technique to estimate the coordinates of a set of individuals and alternatives in a p-dimensional space based on their proximity values. The key contribution of this method is its ability to produce a graphical representation that captures both individuals and objects, allowing for the exploration of patterns and relationships in preference-approvals data.
To demonstrate the usefulness of our method, we apply it to Eurobarometer data on social values. Our goal is to identify common patterns of preferences among individuals and alternatives and gain insights into public opinion on democracy and citizenship. We analyze the resulting clusters to identify key factors shaping public opinion on these issues.
The rest of the paper is organized as follows: Section 2 explains the proposed method in detail. In Section 3, we present a case study involving the application of the proposed approach to Eurobarometer data. Finally, Section 4 offers concluding remarks and suggestions for future research.

2 Notation

Suppose a set of voters $V = \{v_1, \ldots, v_m\}$, with $m \geq 2$, are asked to order n different alternatives, the ranking π is a mapping function from the set of alternatives $X = \{x_1, \ldots, x_n\}$ to the set of ranks $\pi = \{P_\pi(x_1), \ldots, P_\pi(x_i), \ldots, P_\pi(x_n)\}$, where $P_\pi : X \longrightarrow \{1, \ldots, n\}$ assigns the rank of each alternative. If the n alternatives are ranked in n different ranks, a complete (full) ranking or linear order is achieved. In certain cases, some alternatives could receive the same rank, and then a tied ranking or a weak order is obtained.
In the framework of preference-approval modelling, each preference ranking, π, is paired with an approval vector, A. For any given set X of alternatives, we define approvals by partitioning X into G, the set of good alternatives, and $U = X \backslash G$ the set of unacceptable alternatives, where G and U can be empty sets.
We represent a voter's preference-approval profile by a top-down order of alternatives with a horizontal bar: alternatives above the bar are approved, and those below are rejected.

$$\begin{array}{c} x_2 \\ x_1 \\ x_3 \\ \hline x_4 \end{array}$$

The previous representation indicates that the voter's three top-ranked alternatives ordered as: $x_2 \succ x_1 \succ x_3$ are approved and and the voter's bottom-ranked alternative x_4 is disapproved. The preference-approval profile is codified as follows:

$$(\pi_1, A_1) = (2,1,3,4)\&(1,1,1,0).$$

A *profile* ω is a vector of preference-approvals $\omega = \big((\pi_1,A_1),\ldots,(\pi_m,A_m)\big) \in \mathscr{R}(X)^m$, where m individuals and n alternatives are present. Denote with $\omega^{(k)} = (\pi_k, A_k)$ the preference-approval of the voter $v_k \in V$.

3 Method

The aim is to build an $N \times N$ augmented preference-approval distance matrix Δ, with $N = n+m$, which can be analyzed using a standard MDS procedure. This extends the approach of D'Ambrosio et al. [1]. This augmented distance matrix is a block matrix. The first block contains the within-individual dissimilarities, which correspond to the distances between the individuals' evaluations of each alternative. The second block includes the within-objects dissimilarities, which correspond to the distances between the options based on the individual's preferences. Finally, the third block contains the between-sets dissimilarities, which correspond to the distances between the sets of alternatives and the set of individuals. To achieve this goal, the following steps are taken:

1. *Build the $m \times m$ symmetric matrix* **D***, containing the within-individual dissimilarities:*
   First, consider two preference-approvals ,$\big(\omega^{(1)} = (\pi_1,A_1),\ \omega^{(2)} = (\pi_2,A_2)\big) \in \mathscr{R}(X)^m$, and two generic alternatives ,$x_i, x_j \in X$. Then, introduce the preference discordance p_{ij} and the approval discordance a_{ij}, as defined by Albano et al. [2]:

$$\begin{aligned} p_{ij} &= \frac{1}{2}\cdot|O_{\pi_1}(x_i,x_j) - O_{\pi_2}(x_i,x_j)|, \\ a_{ij} &= \frac{1}{2}\cdot\big(|I_{A_1}(x_i) - I_{A_2}(x_i)| + |I_{A_1}(x_j) - I_{A_2}(x_j)|\big), \end{aligned} \tag{1}$$

   Here, O_{π_1}, O_{π_2} are the score matrices defined by Emond and Mason [3]. Next, let $\lambda \in [0,1]$, and $r > 0$. The distance for preference-approvals is given by the mapping ,$D^r_\lambda : \mathscr{R}(X) \times \mathscr{R}(X) \longrightarrow [0,1]$:

$$D^r_\lambda\big(\omega^{(1)},\omega^{(2)}\big) = \frac{2}{n\cdot(n-1)} \cdot \sum_{\substack{i,j=1 \\ i<j}}^{n} \big(\lambda\cdot p^r_{ij} + (1-\lambda)\cdot a^r_{ij}\big)^{\frac{1}{r}}. \tag{2}$$

   To find the elements of the **D** matrix, compute $d_{rs} = D^r_\lambda(\omega^{(r)},\omega^{(s)})$ for all $r,s \in 1,\ldots,m$.

2. *Generate the $n \times n$ symmetric matrix N containing the within-objects dissimilarities:*
   Building on the previous step, we compute two indices that quantify the distance between two alternatives x_i and x_j in terms of preference ($\rho_{ij}^{v_k}$) and approvals ($\alpha_{ij}^{v_k}$), respectively, for each voter $v_k \in V$. These indices represent the discordances between x_i and x_j for voter v_k. The formal definition of these discordances is given below.

$$\begin{aligned} \rho_{ij}^{v_k} &= \frac{1}{n-1} \cdot |P_{\pi_k}(x_i) - P_{\pi_k}(x_j)|, \\ \alpha_{ij}^{v_k} &= |I_{A_k}(x_i) - I_{A_k}(x_j)|, \end{aligned} \tag{3}$$

   where $\rho_{ij}^{v_k}$, $a_{ij}^{v_k} \in [0,1]$. Taking into account the preference and approval discordances introduced in Eqs. (3) and the family of weighted means, we introduce a global measure of discordance between pairs of alternatives.
   Given a profile $\omega = \big((\pi_1, A_1), \dots, (\pi_m, A_m)\big) \in \mathscr{R}(X)^m$ and $\lambda \in [0,1]$, the mapping $\delta_\lambda : X \times X \longrightarrow [0,1]$ is defined as

$$\delta_\lambda(x_i, x_j) = \frac{1}{m} \cdot \sum_{v_k \in V} \left(\lambda \cdot \rho_{ij}^{v_k} + (1-\lambda) \cdot \alpha_{ij}^{v_k} \right). \tag{4}$$

   Thus, the elements of the $\mathbf{B}$ matrix are $b_{i,j} = \delta_\lambda(x_i, x_j)$
3. *Generate the $m \times n$ matrix $\mathbf{C}$, containing the between-sets dissimilarities:*
   - To calculate the distance between an individual and an alternative, it is necessary to express the alternative in its preference-approval form. In particular, the generic alternative x_i will have a preference-approval form in which x_i is ranked first and all other alternatives are ranked last, and x_i is the only approved object. The m-dimensional vector containing the preference-approval form of each alternatives is called O.
     Consider the following set of items $\{x_1, x_2, x_3, x_4\}$ their preference approval form will be $x_1 = \big((1,3,3,3)\&(1,0,0,0)\big)$, $x_2 = \big((3,1,3,3)\&(0,1,0,0)\big)$, $x_3 = \big((3,3,1,3)\&(0,0,1,0)\big)$ and $x_4 = \big((3,3,3,1)\&(0,0,0,1)\big)$.
     Each element of O, denoted by $O^{(i)}$, $i = 1, \dots, n$, is a tied preference-approval representing object i.
   - Find the elements of the $\mathbf{C}$ matrix as $c_{ri} = D_\lambda^r(\omega^{(r)}, O^{(i)})$, with $r = 1, \dots, m$, $\quad i = 1, \dots, n$.
4. *Define the normalizing constants* as
   $\psi = \sqrt{\frac{m^2}{\sum_r \sum_s d_{rs}^2}}$, $\beta = \sqrt{\frac{n^2}{\sum_i \sum_j b_{ij}^2}}$ and $\xi = \sqrt{\frac{mn}{\sum_r \sum_i c_{ri}^2}}$
5. *Generate the preference-approval distance matrix:*

$$\Delta = \left[\begin{array}{c|c} \psi \mathbf{D} & \xi \mathbf{C} \\ \hline \xi \mathbf{C}^T & \beta \mathbf{B} \end{array} \right] \tag{5}$$

6. *Perform a non-metric MDS analysis on the Δ matrix.*

The block-wise matrix defined in Equation 5 is normalized in such a way that the sum of the squared elements is equal to N^2, to avoid trivial solutions in MDS.

4 Case study: Eurobarometer

To showcase the usefulness of the proposed approach, this section features a case study that employs data sourced from the Eurobarometer website[1]. The dataset includes 27 rows, each corresponding to an EU member state, and 9 columns which pertain to social values, such as: x_1: Equality between women and men, x_2: Fight against discrimination, x_3: Tolerance and respect for diversity, x_4: Solidarity among EU States, x_5: Solidarity between the EU and poor countries, x_6: Protection of human rights, x_7: Freedom of religion, x_8: Freedom of movement, and x_9: Freedom of speech. The process of acquiring preference-approvals involves ranking the alternatives according to their popularity within each country. Acceptable options are determined by identifying the alternatives that received more votes than the national average. Figure 1 illustrates the identification of mainly two clusters of countries

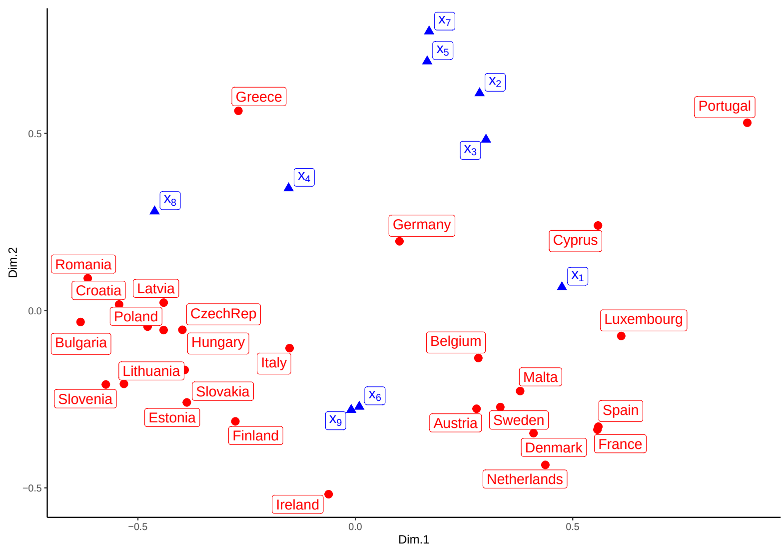


Fig. 1 Multidimensional scaling plot of the Eurobarometer data.

with similar preferences. The first cluster includes countries from Romania to Italy

[1] https://europa.eu/eurobarometer/surveys/detail/2612.

and is located on the left side of the graph, while the second cluster comprises countries from Austria to Luxembourg and is located on the right side. Countries such as Germany and Ireland are located between the two clusters and have borderline ratings, indicating their preference for alternatives is somewhat mixed. Moreover, some countries, such as Portugal and Greece, have ratings that are different from all the others, suggesting that their preferences are unique.
Regarding the alternatives, x_6 and x_9 are close to both clusters. This means that, these alternatives are considered highly important by the countries in both clusters. Additionally, the countries in the left cluster have a preference for x_8, while those in the right cluster lean towards x_1. Overall, the identified clusters and their respective preferences provide valuable insight into how different groups of EU countries express their views on the nine proposed alternatives.

5 Conclusion

In this paper, we proposed an approach for constructing an augmented preference-approval distance matrix to be analyzed by MDS to build a graphical representation that simultaneously captures individual and alternative relationships. This approach extends the method proposed by D'Ambrosio et al. [1] by incorporating preference-approvals dissimilarities into the distance matrix. We demonstrated the usefulness of our approach through a case study using Eurobarometer data on social values across EU member states. Our analysis revealed the existence of two main clusters of countries with similar preferences, with some countries having mixed or unique tendencies. The identified clusters and their respective preferences provide valuable insight into how different groups of EU countries express their views on the nine proposed alternatives. The insights gained from our analysis can inform policymakers and decision-makers about the values and priorities of EU member states, facilitating the creation of policies that align with the preferences of different groups of member states.

References

1. D'Ambrosio A., Vera J. F., and Heiser W. J.: Avoiding Degeneracies in Ordinal Unfolding Using Kemeny-Equivalent Dissimilarities for Two-Way Two-Mode Preference Rank Data. Multivariate Behavioral Research, 57(4): 679-699 (2022)
2. Albano A., García-Lapresta J. L., Plaia A., and Sciandra M.: A family of distances between preference-approvals. Annals of Operation Research. 323: 1-29 (2023)
3. Emond E. J. and Mason D. W.: A new rank correlation coefficient with application to the consensus ranking problem. Journal of Multi-Criteria Decision Analysis, 11(1): 17-28 (2002)

Pandemic Data Quality Modelling: A Bayesian Approach

Modellazione della qualità dei dati pandemici: un approccio bayesiano

Luisa Ferrari, Giancarlo Manzi, Alessandra Micheletti, Federica Nicolussi and Silvia Salini

Abstract When dealing with pandemics like COVID-19, it is crucial for policymakers to constantly monitor the emergency. Correct data reporting is a hard task during pandemics, and errors affect the overall mortality, resulting in excess deaths in official statistics. In this work, we provide tools for evaluating the quality of pandemic mortality data. We accomplish this through a spatio-temporal Bayesian approach accounting for the bias implicitly contained in the data.

Abstract *Quando si affrontano pandemie come il COVID-19, è fondamentale che si monitori costantemente lo stato della pandemia. Tuttavia, una corretta raccolta dei dati è un compito difficile in questi casi e gli errori influiscono sulla valutazione della mortalità complessiva, traducendosi in un eccesso di mortalità nelle statistiche ufficiali. In questo lavoro, si forniscono strumenti per valutare la qualità dei dati sulla mortalità pandemica attraverso un approccio spazio-temporale bayesiano.*

Key words: Pandemics, Bayesian analysis, variance models, time-space models

---

Luisa Ferrari
Department of Statistical Sciences ”Paolo Fortunati”, University of Bologna, e-mail: luisa.ferrari5@unibo.it

Giancarlo Manzi
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: Giancarlo.Manzi@unimi.it

Alessandra Micheletti
Department of Environmental Science and Policy, University of Milan, e-mail: alessandra.micheletti@unimi.it

Federica Nicolussi
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy, e-mail: federica.nicolussi@polimi.it

Silvia Salini
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: Silvia.Salini@unimi.it

1 Introduction

The COVID-19 pandemic brought the world close to a halt in 2020 and 2021 and killed almost seven million people as of early 2023. In similar contexts, actions of surveillance with limited error is crucial, especially in mortality monitoring [3, 6, 9]. In this paper, we consider the bias between excess mortality and the official Italian COVID-19 data in the first 2020 outbreak for evaluating data quality in a space-time context. To model this bias we use a Bayesian framework where two different quality measures ought to be evaluated: (i) the share in the population dying because of a particular infectious disease without being officially reported, and (ii) the coverage of the epidemic by the health systems, which can be considered an adequate indicator of their quality and a proxy for the efficacy of the crisis response.

2 Data

In order to evaluate the data quality on COVID-19 mortality we considered two data sources: (i) official data series on pandemic mortality (ii) national or supranational statistical institute data on population mortality. Data (i) are weekly provincial (EU NUTS-3 level) COVID-19 deaths from February 24th to May 11th, 2020. While data about new cases were regularly published, the number of weekly COVID-19 new deaths was not officially available at this level. Nevertheless, we reconstructed the time series of COVID-19 at a NUTS-3 level indirectly, using other official sources like regional authorities' daily bulletins on provincial deaths and other information sources [4]. Bulletins were in general in a "pdf" format so we were able to scrape data from these documents and to retrieve the data of interest for the majority of the Italian provinces. Regarding data (ii), each year the Italian National Statistical Istitute (ISTAT) provides a weekly record of deaths reported in each municipality in Italy. Here we use a 5-year window consisting of the period 2015-2019 to represent the stable mortality level. The excess mortality is then found by subtracting this stable level from the COVID-19 2020 deaths data, in each province and week of the year.

3 Proposed metrics

The aim of the metrics to be defined is to provide an estimate for the under-reporting mortality bias that the official data has been subject to. Two different metrics with different interpretations for the policy-makers are proposed.

Let D_{ij} be the officially reported total number of deaths in province i and week j which exceeds the average of the previous 5 years, assuming that they can all be imputed to the COVID-19 emergency. Let $\hat{D}_{ij}$ be an estimate for D_{ij}. Let Y_{ij} be the officially reported number of COVID-19-related deaths in province i and week j.

Finally, let POP_i be the average population in province i along the considered period. The additive bias is built as the difference between the *actual mortality* D_{ij}/POP_i and the *official mortality* Y_{ij}/POP_i. This bias is defined as "additive" because it must be added to the official mortality to get the unbiased value:

$$b_{ij}^A = 1000 \cdot \frac{\hat{D}_{ij} - Y_{ij}}{POP_i}$$

In terms of interpretation, $b^{(A)}$ defines the share in the population that died because of COVID-19 without being officially reported, so large values represent a negative scenario. Its trend over time and space (but not its magnitude) is a rough proxy for the part of the pandemic that was concealed and undetected by the public administration.

The ratio between Y_{ij} and D_{ij} assesses the probability of a COVID-19-related death being officially reported. In order to transform it into a bias metric, the complementary probability of not being reported is considered instead and called $b_{ij}^{(M)}$. This is equivalent to the additive bias divided by the excess mortality rate.

$$b_{ij}^M = 1000(1 - \frac{Y_{ij}}{\hat{D}_{ij}})$$

A large bias indicates a bad situation. Regarding its meaning, b^M measures the promptness of the health system to react to the pandemic, thus it is an adequate indicator of its quality and a proxy for the efficacy of the crisis response.

4 Model

Our spatio-temporal model for the two above metrics model resembles the one by Franco-Villoria et al. [5] with the temporal component being a Gaussian random walk model of order 1. For the spatial component we adopted an ICAR model [1] [2], but instead of considering the traditional adjacency matrix $\mathbf{M}$ with only non-negative entries and a null diagonal, and a corresponding diagonal matrix $\mathbf{D}$ where $d_{p,q} = \sum_q m_{p,q}$ to take into account the geographical boundaries, we considered an adjacency matrix with smartphone location data, which actually estimates the average commuting of individuals between two provinces, no matter their actual geographical location:

$$\mathbf{u} \sim N_I\left(\mathbf{0}; \Sigma_u = (\mathbf{I_J} - \mathbf{D^{-1}M})^{-1}\mathbf{D}^{-1}\right).$$

The covariance matrix of the interaction term is defined as the Kronecker product of the covariance matrix of the two main effects, following the work of Knorr-Held [7]. Finally, as for the prior specification we reparametrize (as in Franco-Villoria et al.

[5]) the original variances $\sigma_u^2, \sigma_v^2, \sigma_w^2$ into a total residual variance V, the proportion ψ of this V given by the interaction term, and the proportion ϕ of main effects variance imputable to the spatial effect. The prior specification is then chosen on this new set of parameters. Specifically, the INLA default prior on variance parameters on σ_ε^2, a Uniform on ϕ, and a Penalized Complexity (PC) prior [8] on ψ with base model $\phi_0 = 0$, and a PC prior on V with base model $V_0 = 0$.

5 Results

Figure 1 shows the posterior mean of the spatial random effects over the provinces of Italy. For b_A on the left, provinces in the North of Italy experienced a larger share of underreported deaths with respect to the overall population. However, the spatial distribution completely changes for the b_M metric, as most of the Northern provinces show small values, while the highest effects are found in the Southern and North-Eastern provinces. These figures display how the two indices measure very different quantities, with b_M being much more consistent with the literature on the spatial distribution of health system quality indicators in Italy. With respect to the temporal pattern, the two metrics also show differences. The average temporal trend for b_A, shown in Figure 2, green curve, starts with an increasing part, up to the sixth week in the considered period, followed by a steady fall in the remaining weeks. This is a reasonable result, as it is expected that the indicator b_A performed the worst at the peak of the "official" epidemic evolution, plus a delay due to the fact that deaths are considered instead of cases. Hence, this confirms the assumption that b_A is related to the level of stress of the health system, rather than to the quality of its response to a certain amount of stress.
The results for b_M are again completely different as the posterior means, shown in Figure 2, red curve, shows a steady decreasing trend. Finally, with a DTW clustering on b^M it was possible to detect four groups of provinces according to their performance in facing the emergency. Figure 3 shows the 4 different groups and their centroids, ordered by best (on the left) to worst (on the right) overall performance.

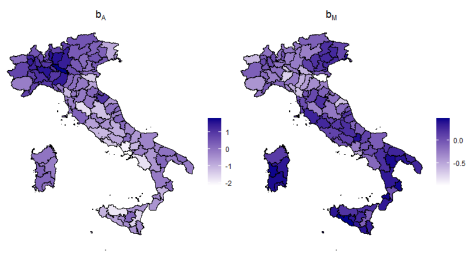


Fig. 1 Posterior mean of the spatial random effects on b_A and b_M

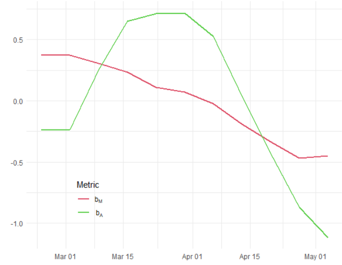


Fig. 2 Posterior mean of the temporal random effect on b_A and b_M

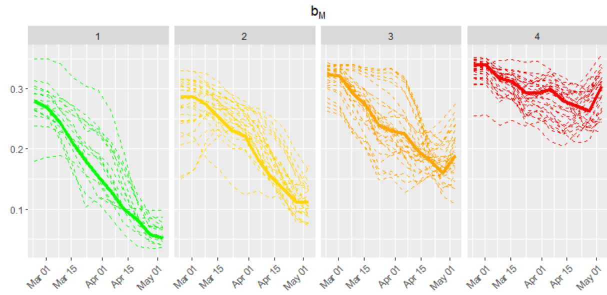


Fig. 3 Posterior mean of the fitted values divided in 4 clusters with corresponding centroids in the provinces of Aosta, Rimini, Catanzaro, and Cosenza

References

1. Besag J.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B. 36, 192–225 (1974)
2. Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43(1), 1–20 (1991)
3. Colombo, R.M., Garavello, M., Marcellini, F., Rossi, E.: An age and space structured SIR model describing the Covid-19 pandemic. J. Math. Ind. (2020) doi: 10.1186/s13362-020-00090-4
4. Ferrari, L., Gerardi, G., Manzi, G., Micheletti, A., Nicolussi, F., Biganzoli, E., Salini, S.: Modeling Provincial COVID-19 Epidemic Data Using an Adjusted Time-Dependent SIRD Model. Int. J. Env. Res. Pub. He. (2021) doi: 10.3390/ijerph18126563
5. Franco-Villoria, M., Ventrucci, M., Rue, H.: Variance partitioning in spatio-temporal disease mapping models. Stat. Methods Med. Res. 31(8), 1566–1578 (2022)
6. Kantner, M., Koprucki, T.: Beyond just "flattening the curve": Optimal control of epidemics with purely non-pharmaceutical interventions. J. Math. Ind. (2020) doi: 10.1186/s13362-020-00091-3
7. Knorr-Held, L.: Bayesian modelling of inseparable space-time variation in disease risk. Stat. Med. 19(17-18), 2555–2567 (2000)
8. Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H.: Penalising model component complexity: A principled, practical approach to constructing priors. Stat. Sci. 32, 1–28 (2017)
9. Wu, J., Tang, B., Bragazzi, N.L., Nah, K., McCarthy, Z.: Quantifying the role of social distancing, personal protection and case detection in mitigating COVID-19 outbreak in Ontario, Canada. J.Math.Ind. (2020) doi: 10.1186/s13362-020-00083-3

Explainable AI for Peer-to-Peer Credit Risk Management

Intelligenza Artificiale Spiegabile per il Management del Rischio di Credito Peer-to-Peer

Golnoosh Babaei, Paolo Pagnottoni and Thanh Thuy Do

Abstract Complex Artificial Intelligence (AI) models used to support decision-making, such as in peer-to-peer lending, often lack interpretable explanations. While Shapley values and the computationally efficient variant Kernel SHAP may be employed for this aim, the latter makes the assumption that the features are independent. We extend the Kernel SHAP method to be able to handle dependent features in the context of credit risk management for peer-to-peer lending. We demonstrate the effectiveness of our method by considering linear and non-linear models with varying degrees of feature dependence, showing that our approach yields more accurate approximations of true Shapley values.

Abstract *I modelli di Intelligenza Artificiale (AI) complessi utilizzati per supportare la presa di decisioni, come nel caso del "peer-to-peer lending", spesso mancano di spiegazioni interpretabili. Sebbene i valori di Shapley e la loro variante computazionalmente efficiente Kernel SHAP possano essere impiegati per tale scopo, quest'ultima fa l'assunzione che le caratteristiche (predittori) siano indipendenti. Estendiamo il metodo Kernel SHAP per gestire le predittori dipendenti nel contesto della gestione del rischio di credito per il "peer-to-peer lending". Dimostriamo l'efficacia del nostro metodo considerando modelli lineari e non lineari con diversi gradi di dipendenza dei predittori, mostrando che il nostro approccio produce approssimazioni più accurate dei veri valori di Shapley.*

Golnoosh Babaei
University of Pavia, Via S.Felice 5, 27100, Pavia (PV), e-mail: `golnoosh.babaei01@universitadipavia.it`

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

Thanh Thuy Do
Insubria University, Via Monte Generoso, 71, 21100 Varese (VA), e-mail: `ttdo@studenti.uninsubria.it`

Key words: Feature dependence; Shapley values; Machine Learning; Explainability.

1 Introduction

Artificial Intelligence (AI) model explainability has gained significant importance over recent times for supporting decision-making in various domains such as medicine and healthcare, credit scoring and fraud detection. However, these models' predictions are difficult to explain. Model explanation is crucial in two main aspects: a) a practical one, which raises the question as to whether the model is trustworthy enough to rely on its predictions; b) a legal one, which pertains the compliance with regulations on automated decisions and the General Data Protection Regulation. This is of particular relevance in the context of peer-to-peer lending, where the decisions of lenders are based on the creditworthiness of borrowers, and transparent and trustworthy explanations are necessary for establishing trust, and rendering transparent and explainable decisions to lenders and borrowers.

Over the past few decades, the practise of credit risk management has increased significantly, resulting in an increased precision of Forecasting Models. Many of these algorithms demonstrate a high degree of accuracy, yet their complexity and opaque nature render their outcomes difficult to interpret. The evaluation of credit risk is utilised to ascertain the borrowers' solvency and applicable interest rates in peer-to-peer lending. It can be challenging for lenders to have faith in the predictions generated by intricate ML models. Explainable Machine Learning (XAI) endeavours to render Machine Learning models more transparent and accessible. The use of XAI could promote an increased sense of trust in algorithms when employed in credit risk management and other areas, by providing transparent insight into the predictions yielded by the model. Game-theoretic Shapley values may solve this issue. Recent model-agnostic explanation methods made it possible to understand each predictor's contribution to the overall prediction.

In this paper, we propose an extension of the Kernel SHAP method to handle dependent features in the context of credit risk management for peer-to-peer lending in line with [1]. We demonstrate the effectiveness of their approach by using four predictive ML models: logistic regression (LR), generalized additive models (GAMs), XGBoost, and random forest (RF). We provide examples of linear and non-linear models with varying degrees of feature dependence, which show that their method yields more accurate approximations of the true Shapley values. As a result, our proposed method has practical applications in improving the interpretability and transparency of decision-making ML models for credit risk management in peer-to-peer lending.

2 Kernel SHAP for dependent features

The Kernel SHAP technique Kernel uses a weighted linear regression to compute the importance of each feature, which are namely also coefficents from a local linear regression. It may, however, produce inaccurate results if the features of a given model show high levels of dependence because predictions are made using non-representative data instances. In the context of classical machine learning, a predictive model, $f(x)$, can be trained using a training set of size n_{train} comprised of sets y $\left\{y^i, x^i\right\}_{i=1,\ldots,n_{\text{train}}}$ where $j = 1,\ldots,n_{\text{train}}$. This model attempts to closely approximate the response value y. To explain the prediction $f\left(x^*\right)$ for a particular feature vector $x = x^*$. The Kernel SHAP technique only uses the independence assumption $p\left(x_{\bar{S}} \mid x_{\mathscr{S}}\right) = p\left(x_{\bar{S}}\right)$ - see [1].

Here we investigate how taking into account for dependence in three different ways improves AI credit risk model accuracy and feature explainability, relative to a baseline of independence.

2.1 Multivariate Gaussian distribution

Assuming that the feature vector x is derived from a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, then the conditional distribution $p\left(x_{\mathscr{S}} \mid x_S = x^*_{\mathscr{S}}\right)$ is also multivariate Gaussian. By expressing $p(x)$ in terms of $p(x) = p\left(x_{\mathscr{S}}, x_{\mathscr{S}}\right) = \mathrm{N}_M(\mu, \Sigma)$ with $\mu = \left(\mu_{\mathscr{S}}, \mu_{\mathscr{S}}\right)^\top$ and

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{\mathscr{S}\mathscr{S}} \\ \Sigma_{\mathscr{S}\mathscr{S}} & \Sigma_{\bar{S}\bar{S}} \end{bmatrix}$$

gives $p\left(x_{\mathscr{S}}|_{\mathscr{S}} = x^*_{\mathscr{S}}\right) = \mathrm{N}_{|\overline{\mathscr{S}}}\left(\mu_{\overline{\mathscr{S}}|\mathscr{S}}, \Sigma_{\overline{\mathscr{S}}_{|\mathscr{S}}}\right)$, with

$$\mu_{\bar{S}|\mathscr{S}} = \mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\left(x^*_S - \mu_{\mathscr{S}}\right)$$

and

$$\Sigma_{\bar{S}|\mathscr{S}} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{\mathscr{S}S}^{-1}\Sigma_{S\bar{S}}.$$

2.2 Gaussian Copula

When the features are far from having a multivariate Gaussian distribution, an alternative approach is to model the marginal distributions with their empirical counterparts and model the dependence structure using a Gaussian copula. A d-dimensional copula is a multivariate distribution, C, characterized by uniformly distributed marginal probabilities $\mathrm{U}(0,1)$ over the unit interval of $[0,1]$. Sklar's theorem states that for each multivariate distribution F with univariate distributions $F_1, F_2, \ldots, F_d$ can be written as

$$F(x_1, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)),$$

for some appropriate d-dimensional copula C. In fact, the copula from (12) has the expression

$$C(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_d^{-1}(u_d)\right)$$

where the F_j^{-1} s are the inverse distribution functions of the marginals. Assuming a Gaussian copula, the following methodology can be employed to generate samples from $p\left(x_{\mathscr{S}} \mid x_{\mathscr{S}} = x_{\mathscr{S}}^*\right)$.

2.3 Empirical conditional distribution

If the dependence structure and marginal distributions of x are significantly different from the Gaussian, it is anticipated that neither of the two aforementioned methods will be effective. In such cases, we suggest a non-parametric methodology. The kernel estimator, a classical approach for non-parametric density estimation, has been subject to further refinement and advancement in the decades following its initial introduction, with various works representing some developments in the field. The kernel estimator is severely hampered by the issue of the curse of dimensionality. Rapidly impedes its application in multivariate circumstances. Additionally, there are only a few techniques available for the non-parametric estimation of conditional densities, particularly when either $x_{\mathscr{S}}$ or $x_{\overline{\mathscr{S}}}$ is not one-dimensional. Ultimately, the majority of kernel estimation techniques produce a non-parametric density estimate, while it is necessary to be capable of producing samples from the estimated distribution. Consequently, we have formulated an empirical conditional method to approximately sample from$p\left(x_{\overline{\mathscr{S}}} \mid x_{\mathscr{S}}^*\right)$. It can be seen that our empirical conditional distribution approach (with $K = n_{\text{train}}$) is equivalent to the Nadaraya-Watson estimator [35], by considering the estimation of $E\left[f(x) \mid x_{\mathscr{S}} = x_{\mathscr{S}}^*\right]$ as a regression problem with response $f\left(x_{\mathscr{S}}^i, x_{\mathscr{S}}^*\right)$ and covariates $x_{\mathscr{S}}^i, i = 1, \ldots, n_{\text{train}}$.

3 Empirical Findings

We test the proposed method on a set of four predictive models, so to compare the accuracy and prediction explanations coming from different ML models and feature dependence setups. We therefore select one simple predictive model, i.e. logistic regression, and three more complex ones. The three machine learning (ML) models considered are Generalized Additive Models (GAMs), Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

One of the mostly used dataset in the P2P Lending Credit Scoring literature is provided by the Lending Club (LC) platform. This dataset includes information of the borrowers (individuals) and their requested loans, and we consider information from 2007 to the fourth quarter of 2018, with a total number of loans in this dataset is 2260701. To motivate our study, we provide an illustrative example of the Shapley value prediction explanation for ten randomly selected observations in Figure 1. As one may notice, the observations having higher predicted probability of default are associated with generally higher Shapley values of both loan amount and annual income.

We then perform random subsampling and obtain the distribution of the Shapley values for each of the 100 repetitions we consider. We present in Figure 2 the distribution of AUC values for the random sub-sampling applied to test data. The figure shows that the models are quite comparable in terms of forecasting performances, with the distribution of the AUC of XGBoost being skewed slightly to the right, hence being a little preferable in terms of predictive performance. This is confirmed by Table 1, which highlights that the highest average AUC is that of XGBoost, though the least dispersed predictions are those made by GAM. In our working paper, we have carried out analyses on the methods to take into account for dependence when producing prediction explanations through Shapley values, and have found that these increase the accuracy and explanation of the model features.

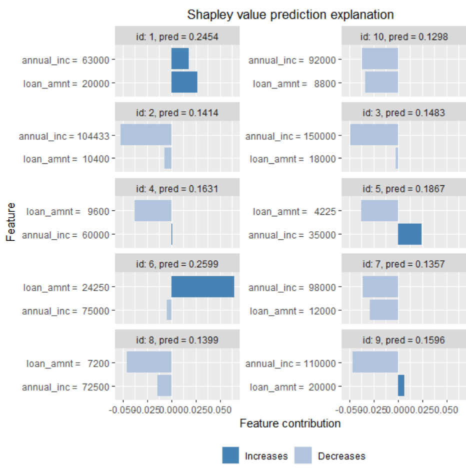


Fig. 1: Shapley Values of the variables "Loan Amount" and "Annual Income" for ten randomly selected observations.

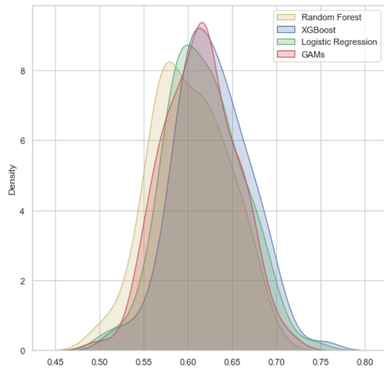


Fig. 2: Distribution of AUC values for the random sub-sampling applied to test data.

Model	Average AUC	STD AUC
XGBoost	0.6269518	0.04201579
Logistic Regression	0.6176347	0.04274917
GAM	0.6126129	0.04024374
Random forest	0.6014294	0.04299568

Table 1: Descriptive statistics on the distribution of AUC values for the 100 models applied to the test dataset.

References

1. Aas, K., Jullum, M. and Loland, A.: Explaining individual predictions when features are dependent: More accurate approximations to shapley values. Artificial Intelligence, 298, 103502 (2021)

Tackling misclassification in surveys about undeclared work via the EM algorithm

Far fronte agli errori di classificazione nei sondaggi sul lavoro sommerso tramite l'algoritmo EM

Maria Felice Arezzo, Giuseppina Guagnano and Domenico Vitale

Abstract For the development of adequate policy measures which deal with undeclared work, it is important to knowledge its extent and structure. Direct surveys of individuals aim to fill this gap. However, in view of the sensitivity of the subject, collected information may be affected by misclassification errors. In this work, we aim to estimate the individual propensity to work off-the-book, tackling misclassification errors. To this end, we used an approach designed for modeling presence-only ecological data via logistic regression and based on the expectation-maximization (EM) algorithm.
Abstract *Lo sviluppo di adeguate misure di contrasto al lavoro nero richiede la conoscenza della sua dimensione e struttura. Le indagini campionarie dirette mirano a colmare questo divario. Tuttavia, data la natura sensibile dell'argomento, le informazioni raccolte possono essere affette da errori di classificazione. In questo lavoro ci proponiamo di stimare la propensione individuale a lavorare in nero e di affrontare gli errori di classificazione. A tal fine, abbiamo utilizzato una procedura sviluppata per la modellazione di dati ecologici di sola presenza tramite regressione logistica e basata sull'algoritmo di expectation maximization (EM).*

Key words: Undeclared work, Expectation-Maximization, Eurobarometer survey, Misclassification error, Logit

Maria Felice Arezzo
Department of Methods and Models for Economics, Territory and Finance (MEMOTEF), Faculty of Economics, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy, e-mail: mariafelice.arezzo@uniroma1.it

Giuseppina Guagnano
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy, e-mail: giuseppina.guagnano@uniroma1.it

Domenico Vitale
MEMOTEF Department, Sapienza University of Rome, Via del Castro Laurenziano, 9, 00161 Rome, Italy, e-mail: domenico.vitale@uniroma1.it

1 Introduction

Shadow economy is the part of an economy which is not declared for tax and that typically involves exchange of goods and services which are paid for in cash. One important component of shadow economy is the undeclared work. Knowledge of the extent and structure of undeclared work within an economy is crucial for the development of adequate policy in agreement of the EU employment goals set out in the Lisbon strategy.

To this end, the Eurobarometer survey no. 402 on undeclared work conducted in 2013 is a first attempt to measure undeclared work on an EU wide basis and in a cross-nationally comparable way using the same methodology, questionnaire concept and definition in all countries. Nevertheless, in view of the sensitivity of the subject, only a low number of respondents (4%) reported to have carried out undeclared work while the vast majority declared they have not. Part of the latter responses are likely to be misclassified.

With the aim to estimate the individual propensity to work off-the-book and tackling misclassification errors, we used the approach proposed by [3] for modeling presence-only ecological data. The approach is based on the expectation-maximization (EM) algorithm to estimate the parameters of the logistic model where the dependent binary variable measures the involvement in undeclared work.

2 Materials and Methods

2.1 Data

We used the data from special Eurobarometer survey no. 402 on undeclared work conducted in 2013. The survey is nationally representative of the adults aged 15 years or older living in one of the EU-28 countries. Besides the demographics, it comprises a rich set of questions on undeclared work. The total number of interviews is n = 27,563. The sample sizes vary from a minimum of 500 in Malta to a maximum of 1 499 in Germany (see https://europa.eu/eurobarometer/screen/home for more details and materials).

2.2 Model and Computation

Let Y be a true binary variable. The probability that an individual participated in undeclared work ($y = 1$) or not ($y = 0$) conditional on a set of covariates x, can be modeled via its logit

$$\text{logit}(\mathbb{P}(Y=1|x) = \eta(x) \quad \Rightarrow \quad \mathbb{P}(Y=1|x) = \frac{e^{\eta(x)}}{1+e^{\eta(x)}} \tag{1}$$

where $\eta(x)$ can be linear in x, as in logistic regression. With *uncontaminated* data, these models can be fitted using standard methods.

In our case-study, however, the observed dependent binary variable representative of Y, can be affected by misclassification errors. In particular, it is very likely that someone declares to work on the book when they actually do not. The other way around, although theoretically possible, is very unlikely to happen for undeclared work.

Let S a binary variable indicating whether someone declares to work off-the-book ($s=1$) or not ($s=0$). We assume that there is some unknown distribution $\mathbb{P}(Y,X,S)$ such that (y_i, x_i, s_i) for $i = 1,...,n$ is an i.i.d. sample drawn from it and data (x_i, s_i) is observed. We assume that misclassification errors may occur only when $s = 0$, that is we assume $y = 1$ when $s = 1$, but when $s = 0$, y can be either 1 or 0.

In this perspective, for $s_i = 0$, we can think of these y_i as missing data and use an iterative procedure based on the EM algorithm to impute the unknown y_i at each iteration and then fits a logit model using these imputed y_i. The procedure can be summarized in the following steps (for a complete exposure of the theoretical background, the reader can refer to [3]):

- Step 1: Initialize the value of y_i for $s_i = 0$ with π, the expected *true* probability of working off-the-book:

$$y_i^{(0)} = \pi \quad \text{for} \quad s_i = 0. \tag{2}$$

- Step 2: A Maximization step is then performed, which fits a maximum-likelihood logistic model to the current values of $\hat{y}_i^{(k)}$

$$\text{logit}(\mathbb{P}(\hat{y}^{(k)} = 1|x, s=1)) = \hat{\eta}(x)^{(k)}, \tag{3}$$

  and applies a case-control adjustment to account for unequal sampling rates of declared and undeclared cases of working off-the-book:

$$\hat{\eta}(x)^{(k)}_{adj} = \hat{\eta}(x)^{(k)} - log\Big(\frac{n_1 + \pi n_0}{\pi n_0}\Big), \tag{4}$$

  where n_1 and n_0 are the number of individuals having declared to work off-the-book ($s = 1$) and not ($s = 0$).
- Step 3: An Expectation step then applies the model to update the estimates of the (missing) y_i data:

$$\hat{y}_i^{(k)} = \frac{e^{\hat{\eta}^{(k)}_{adj}}}{1+e^{\hat{\eta}^{(k)}_{adj}}} \quad \text{for } s_i = 0 \quad \text{and} \quad \hat{y}_i^{(k)} = 1 \quad \text{for } s_i = 1. \tag{5}$$

- Step 4: Iterate Step 2 and 3 k times until convergence.

We need to specify the π parameter both in the initialization step and in the case-control adjustment. According to previous findings [1], we put it equal to 0.3.

3 Results and Discussion

In Table 1 we summarize the main results. First of all, we note that the signs of the estimated coefficients are coherent both with each other, and with respect to previous results (see, for example, [1] and [2]). In particular, the typical undeclared worker is a young male, unemployed, having financial problems most of the time and a low level of tax morality. In this regard note that the Tax moral covariate is obtained constructing a composite indicator, which takes on higher values the higher the morality. Furthermore, the propensity to work off-the-book increases when the risk of detection is very small, and the expected sanction is prison (which may be considered unrealistic) or simply the payment of all taxes due. For this covariate, however, only the EM estimates result statistically significant, due to the lower estimated coefficients and larger standard errors. In fact, the naive model always shows higher standard errors than the EM model.

A second important result is that parameter estimates from the EM model are always larger, in absolute values, than those from the standard logit model (the only exception refers to Urban covariate, but the parameter estimate is not significant in both model). This suggests that ignoring the misclassification errors, as in the standard logit procedure, may produce inconsistent parameter estimates. The main difference lies in the intercept estimates, both in magnitude and in sign. This is particularly relevant because of the possible implications. In fact, in the EM model, the positive and larger estimate implies a larger probability, for the base individual, of being an undeclared worker. As a consequence, this larger estimate in general allows to predict a case ($y_h = 1$) much more often than in the naive model, thus converting some of the (presumable) false negative answers into positive ones.

4 Conclusions

In this article we propose a first attempt to estimate the individual propensity to work off-the-book using the approach of modeling the presence-only ecological data [3] and the EM algorithm. The main advantage of the proposed procedure is a correction of the parameter estimates from a standard logit model and in particular that corresponding to the intercept. Actually, we obtained a very larger estimate for this parameter, thus increasing the occasions in which we can predict a case, even when we observe a negative answer for the response variable. The detection of false negative answers is particularly important for the development of adequate policy measures and we believe that our work may contribute to fill an important gap of the economic literature, giving the possibility of detecting a possible false regular

Table 1 Summary fits of the Logit model estimated with standard procedure (naive) and through the EM algorithm with $\pi = 0.3$

Regressors	EM Estimate	 Std error	 Pr(> $\|z\|$)	Naive Estimate	 Std error	 Pr(> $\|z\|$)
Intercept	$+3.268$	0.154	0.000	-0.680	0.236	0.004
Female	-0.988	0.038	0.000	-0.623	0.064	0.000
Age	-0.043	0.002	0.000	-0.021	0.003	0.000
Tax morale	$+0.917$	0.016	0.000	$+0.369$	0.018	0.000
Urban	-0.059	0.039	0.134	-0.086	0.065	0.190
Occupation (Ref. Cat.: Unemployed)						
Self employed	-0.840	0.087	0.000	-0.036	0.126	0.774
Employed	-1.437	0.065	0.000	-0.612	0.092	0.000
Inactive	-1.224	0.076	0.000	-0.532	0.111	0.000
Retired	-1.716	0.083	0.000	-1.034	0.141	0.000
Financial problems (Ref. Cat.: Most of the time)						
Occasional	-0.982	0.058	0.000	-0.503	0.087	0.000
None	-1.662	0.060	0.000	-0.940	0.090	0.000
Detection risk (Ref. Cat,: Very small)						
Very high	-1.290	0.073	0.000	-0.890	0.130	0.000
Fairly high	-1.379	0.058	0.000	-0.877	0.092	0.000
Fairly small	-0.631	0.052	0.000	-0.370	0.079	0.000
Expected sanctions (Ref. Cat.: Tax or social security contributions)						
Due tax plus fine	-0.194	0.045	0.000	-0.106	0.074	0.154
Prison or other	$+0.179$	0.061	0.003	$+0.078$	0.096	0.419

worker. We are also confident that the proposed procedure gives more reliable estimates and our future work will be devoted to ascertain a possible evidence of this, through simulation studies.

References

1. Arezzo, M.F., Arima S., Guagnano G.: A two-part measurement error model to estimate participation in undeclared work and related earnings. Stat Model (2023) doi:10.1177/1471082X221145240
2. Arezzo, M.F., Williams, C.C., Horodnic, I.A. and Guagnano, G.: Revisiting tax morale: evaluating the acceptability of business- and individual-level non-compliance on participation in undeclared work. Int J Manpow (2023) https://doi.org/10.1108/IJM-11-2022-0543
3. Ward, G., Hastie, T., Barry, S., Elith, J., Leathwick, J.R.: Presence–only data and the EM algorithm. Biometrics, 65(2), 554–563 (2009) doi:10.1111/j.1541-0420.2008.01116.x

Solicited Session SS20 - *Tourism, territory and data analysis*

Session SISTUR organized by Fabrizio Antolini
Chair: Fabrizio Antolini

1. *Tourism, sustainability, and territorial impact: an input-output analysis* (Garau G. and El Meligi A.K.)
2. *The Impact of Big Data in Tourism* (Ciuffreda R., Choedon C. and Simonetti B.)
3. *Neural network-based prediction of domestic tourists' length of stay in Italy* (Antolini F. and Cesarini S.)
4. *The management of cultural heritage in contexts of undertourism: a model for assessing the economic sustainability of public-private partnerships* (Calabrò F.)

Tourism, sustainability, and territorial impact: an input-output analysis

Turismo, sostenibilità e impatto territoriale: un'analisi input-output

Giorgio Garau and Andrea Karim El Meligi

Abstract In the article, the combination of the concept of Tourism Carrying Capacity (TCC) with intersectoral analysis using a Local Social Accounting Matrix (LSAM) allows for the estimation of employment impact related to crowding perception. The empirical analysis is based on a dataset constructed from a survey conducted in the Asinara National Park. Furthermore, using the Leontief inverse matrix, backward multipliers can be calculated to measure the increase in total production resulting from changes in final demand. The study concludes with a discussion on the potential improvement of impact estimation, considering both estimates of tourism expenditure by tourism type and the transition from an input-output table based on total flows to one based on domestic flows.

Abstract *Nell'articolo la combinazione del concetto di Capacità di Carico Turistico (CCT) con un'analisi intersettoriale, che utilizza una Matrice di Contabilità Sociale Locale (MCSL), consente di stimare l'impatto sull'occupazione legato alla percezione di affollamento. Il test empirico si avvale di una base di dati costruita a partire da un sondaggio condotto nel Parco Nazionale dell'Asinara. In seguito, a partire dalla matrice inversa di Leontief, si possono calcolare i moltiplicatori a monte, che misurano l'aumento della produzione totale derivante dai cambiamenti nella domanda finale. Il lavoro si conclude con una discussione sul potenziale miglioramento delle stime di impatto, considerando sia le stime della spesa turistica per tipologia di turismo che il passaggio da una tavola input-output basata sui flussi totali ad una tavola basata sui flussi domestici.*

Key words: Input-Output, Tourism Carrying Capacity, Impact analysis.

Giorgio Garau

Department of Humanities and Social Sciences, University of Sassari, Via Roma 151, Italy; e-mail: giorgio@uniss.it.

Andrea Karim El Meligi

European Commission, Joint Research Centre. Seville; e-mail: andreak.elmeligi@gmail.com.

1 Introduction

Input-output analysis, developed by Leontief, represent the economy through interrelationships between productive sectors, by means of matrix M. A second matrix, A, contains the technical coefficients and represents the technique used by companies, grouped in productive sectors, during a given period. The last matrix is the Leontief inverse matrix: $Z = (I - A)^{-1}$, used to calculate backward multipliers and measure the increase in total production (X) resulting from a variation in final demand (FD) in a sector: $X = Z * FD$.[2]
If we want to improve the role of tourism as an engine of the economy and its contribution to integrated economic development, we can only design policies of sustainable tourism that do not compromise the future development of the territory. In our work, we will adopt the perspective of Tourism Carrying Capacity (TCC) to define the environmental and economic impact of a visitor flow on a small island (Asinara), our case study. We will then combine the TCC with a simplified intersectoral analysis framework, using the Local Social Accounting Matrix (LSAM). This study evaluates the employment impact of the crowding perception by using data resulting from a survey conducted in the Asinara National Park, whose results are presented in Corbau [2] and in Carboni [1]. Studies on sustainable tourism policies must be local based and adapted to specific areas, considering the dimensions evaluated at local level. This proposal is the subject of section 2. Section 3 is then dedicated to the implementation of a National Accounting scheme and finally, conclusions are outlined in section 4.

2 Sustainable tourism policies during the pandemic

To build the Tourism Carrying Capacity (TCC), we start with the physical carrying capacity (available beach area in km^2) and apply correction factors (which account for the tourist impact, beach type, and geomorphology) to arrive at the real carrying capacity (RCC). Subsequently, using additional correction factors (related to management capabilities), we determine the effective carrying capacity (ECC). As shown in table 2, the transition from RCC to ECC depends on beach management. If challenges such as a lack of parking or beach services are addressed, the ECC can increase and reach the RCC. In this regard, improving the governance of tourism development strategies (involving various institutions with decision-making authority) is crucial.
During the pandemic, distancing measures were introduced, which defined the "social" distances to be maintained between beach umbrellas (approximately 10 m^2

[2] In the short to medium term, it is reasonable to assume that the technique available to companies (not the existing one), i.e., the mix of techniques used by companies within the sector, remains unchanged. It is said that the technical coefficients of production are fixed over time. In the short to medium term, supply does not change, but the effects of demand shocks (e.g., an increase in tourist spending or tourist arrivals) can be measured. Production bottlenecks, which hinder the deployment of demand effects, can also be identified.

per umbrella for two people). We drew inspiration from this to reflect on an additional correction factor for the Tourism Carrying Capacity (TCC), which considers the tourist's perception of congestion in the area, known as the Pf (perceived factor). Through a questionnaire, beach visitors were asked to express their perception of crowding based on images. Only 2% of respondents reported a high perception of crowding, 20% considered the level of crowding to be moderate, while the remaining 78% perceived the crowding to be nearly absent.
The results of the survey, based on the association between the pictures with different level of crowding and the answers provided in the questionnaires, were used to transform the subjective measure of the crowding perception into a new correction factor of the carrying capacity[3].

Table 1 Perceived crowding, average propensity, perceived factor and available area.

Response item	***Pc***	***c***	***Pf***	*Available area m^2*
High	0.02	0.98	0.98	8
Medium	0.20	0.80	0.80	6
Low or Null	0.39	0.61	0.61	4

Source: Own elaboration

Our measure is conceived to provide a specific correction factor for RCC and ECC, depending on the available area per visitor in the beach of our case study. Table 2 resumes the adjusted carrying capacities when considering the crowding perception, in this way, the new indices compute the potential reduction of encounters per season[4].

Table 2 Adjusted carrying capacities with the crowding perception factor, seasonal results.

A(m^2)	*Au(m^2)*	*Rotation factor*	*PCC*	*Correction factor*	*RCC*	*Management. capacity*	*ECC*	*Perceived factor*	*RCC adjusted*	*ECC adjusted*	*ΔRCC*	*ΔECC*
916	8	3	41,907	0,69	29,240	0,43	12,573	0,98	28,655	12,322	-585	-251
		4	55,876		38,987		16,764		38,207	16,429	-780	-335
	6	3	55,876		38,987		16,764	0,80	31,190	13,411	-7,797	-3,353
		4	74,501		51,983		22,353		41,586	17,882	-10,397	-4,471
	4	3	83,814		58,481		25,147	0,61	35,673	15,340	-22,808	-9,807
		4	111,752		77,974		33,529		47,564	20,453	-30,410	-13,076

Source: Carboni [1] and own elaboration

When looking at an available area of 8 m^2 per visitor and a rotation factor of 3, only slight decreases are observed in the RCC and ECC adjusted, after applying the related Perceived factor; while when the available area required per tourist is reduced to 4 m^2,

---

[3] The following expression shows a normalized measure related to the Perceived crowding $c = 1 - Pc$, then translated into an access propensity to the beach c in Table 1 and finally associated to an available area as a Perceived factor (*Pf*) in Table 2.

[4] The content of the Table 2 is the following: A= Available area for tourist use; Au= Area required per tourist; Rf = Rotation factor corresponding to the number of visits per day; PCC = Physical carrying capacity; Cf = Correction factor, RCC = Real carrying capacity, Mc = Management capacity; ECC = Effective carrying capacity.

ΔRCC and ΔECC decrease dramatically, passing from 58,481 to 35,673 visitors in the first case, and from 25,147 to 15,340 visitors, in the latter.

3 The national accounts perspective

In this section, a macroeconomic analysis is presented by using a SAM scheme developed at a local level, based on four municipalities representing a potential gravitational area of tourists visiting the Asinara National Park. The accounting scheme presented in this study includes more blocks than those normally considered in an Input-output(I-O) framework and, moreover, is built at the municipal level. This makes the Local Social Accounting Matrix (LSAM) an important and rare tool when considering that only recently national statistical offices has started to publish national accounting matrices, while there are currently very few official statistics that offer the same detailed information at the regional level. This framework is an adaptable tool that presents transactions able to identify interdependencies among Activities, Primary factors, and Institutional sectors, being therefore supporting modeling in the task of measuring the employment impact on specific sectors. Moreover, this framework assesses not only the impacts on the main economic variables (as for instance, GDP and output), but it is also able to capture the income distribution variation among the Institutional sectors.

4 The LSAM-based Model

The macro-economic analysis is based on the input-output theory established by Leontief, the system of national accounts (SNA) by Stone and the extended Input-Output model proposed by Miyazawa and further developed by Pyatt [4]. By proposing the SAM/SNA framework as a reference scheme, the analysis also considers along with the direct and indirect effects, the induced ones, by considering the block of the primary and secondary distribution. The results of the static model will be expressed in terms of employment generated by industry.
By using as starting point the 2014 SAM of Sardinia (as in Garau [3]) and local specific dataset (e.g., the number of employees at sectoral level), the regional coefficients have been downscaled in order to obtain an estimation of both the Supply and Use tables.

Table 3 Aggregated framework of the LSAM, in million euro.

	Commodities	***Industries***	***Value added***	***Taxes less Subsidies***	***Institutional Sectors***	***Capital Formation***	***Rest of Italy and of World***	***Total***
Commodities		5,323		-	3,198	792	4,612	**13,925**
Industries	9,401			-		-	-	**9,401**
Value added		3,979		-		-	14	**3,993**
Taxes less Subsidies		99		-	318	34	1	**452**
Institutional Sectors		-	3,994	447	3,889	-	1800	**8,510**
Capital Formation		-	-	-	861	-	-35	**826**
Rest of Italy and of World	4,524		-1	5	244	-	-	**4,772**
Total	**13,925**	**9,401**	**3,993**	**452**	**8,51**	**826**	**4,77**	

Source: Own elaboration

5 Building scenarios and improving impact estimation

Given the possibility of approximately 48 million (sustainable) visitors[5], in 2022 we are moving towards a much higher number of arrivals compared to the 2019 figure, which was around 18 million (official data). Sardinia can sustain more than double this number. However, the presence of the unrecorded visitors (estimated to be at least 50% of the official figure) and the uneven distribution of tourists along the entire coastline suggest the need to refine the initial estimate of sustainable arrivals (including a census of different beach types). Subsequently, it would be appropriate to estimate correction factors for different types of beaches (pocket beaches and long beaches) to determine where actions can be taken to mitigate the effects of various correction factors.

Moving from Input-Output Tables (IOTs) based on total flows to those using domestic flows (domestic production inputs) is crucial to estimate the impacts on local businesses activated by tourism demand. This transition is important because it would help us understand that what we need to focus on in terms of policy is not just increasing visitor numbers, which negatively affect service congestion and the perception of tourists among the resident population. Rather, it is about understanding where the impacts of an increase in visitor numbers end up. In other words, the IO

[5] In Sardinia, out of 1897 km of coastline, 459 km are categorized as low coast, with 153 km experiencing erosion. If we consider an average width of 10 m for the first type of beach and 5 m for the second type, we obtain: a. low coast not experiencing erosion (10 m width): 306,000 * 10 = 3,060,000 m^2 available; b. low coast experiencing erosion (5 m width): 153,000 * 5 = 765,000 m^2 available. In total, there are 3,825,000 m^2 of beach available, which, considering the surface areas in table 3 (8 m^2 to ensure non-congestion), can accommodate 478,125 tourists per day (47,812,500 per season).

model allows for a simple (because linear) initial estimation of how a change in a component of final demand (consumption or exports - selling services to non-residents) is transmitted to the production system that provides those services (accommodation and catering, primarily). The distinction between domestic and total flows precisely highlights whether the activated flows come from the regional production system or from outside this system, whether they activate regional production or involve the import of inputs[6].

References

1. Carboni, D. and Pungetti, G.: L'importanza della capacità di carico turistica per una governance condivisa e per uno sviluppo sostenibile delle isole mediterranee, *Geotema*, XXI (2018)
2. Corbau, C. et al.: Tourism analysis at Asinara Island (Italy): Carrying capacity and web evaluations in two pocket beaches, *Ocean & Coastal Management* (2019)
3. Garau G., Carboni D. and El Meligi A.K.: Economic and environmental impact of the tourism carrying capacity: A local-based approach. J Hospitality Tourism Res, 10963480211031426 (2021)
4. Pyatt G. and Jeffery R.: Accounting and fixed price multipliers in a social accounting matrix framework, *The Economic Journal*, 89, 356 (1979)

[6] Another improvement regarding the representativeness of these scenarios relates to constructing expenditure vectors for different categories of tourists to measure the impact of different demand segments (luxury tourism, nature-based tourism, cultural tourism, sports tourism, etc.) and develop targeted hospitality policies.

The Impact of Big Data in Tourism

L'Impatto dei Big Data nel Turismo

Raffaela Ciuffreda, Choedon Choedon and Biagio Simonetti

Abstract In the era of digital transformation, Big Data has emerged as a valuable resource for global travel by providing significant challenges and opportunities for the tourism sector. This study explores the benefits and drawbacks of big data in the tourist industry. Big Data offers valuable information on predicting tourist demand, enabling better decision-making, and providing the service in more efficiently and effectively. This can result in improved productivity, statistical analysis and targeted marketing campaigns along with customized suggestions to enhance customer relationship management. Big Data revolutionizes destination management, offering insights into tourist flows and assisting manage overcrowding. However, risks to data security and privacy, concerns with data quality, resource needs, and ethical dilemmas are some of the challenges, though. Sticking to a balance between data-driven insights and human judgment is crucial to maintain a personalized and authentic travel experience. Despite these difficulties, leveraging Big Data can drive growth and deliver exceptional experiences in the dynamic world of tourism.

Abstract *Nel corso della Digital Transformation, i Big Data giocano un ruolo cruciale nel settore turistico, fornendo contestualmente significanti sfide e opportunità. Attraverso questo studio vengono esplorati vantaggi e vantaggi dell'introduzione dei Big Data in relazione al loro utilizzo nella travel industy. L'applicazione di grandi moli di dati consente di prevedere la domanda, di migliorare i processi decisionali e di fornire servizi più efficaci ed efficienti. Conseguentemente, è possibile incrementare la produttività, condurre analisi statistiche e campagne di marketing mirate, migliorare il Custumer Relationship Management. L'utilizzo dei*

Raffaela Ciuffreda
University of Sannio, Italy; e-mail: ciuffreda@unisannio.it

Choedon Choedon
University of Sannio, Italy; e-mail: choedon@unisannio.it

Biaggio Simonetti
University of Sannio, Italy; e-mail: simonetti@unisannio.it

Big Data consente di rivoluzionare il Destination Management attraverso informazioni sui flussi turistici che consentono di gestire adeguatamente i turisti in arrivo nelle principali destinazioni. D'altronde, non bisogna trascurare i rischi legati alla privacy, alla qualità dei dati, al crescente fabbisogno di risorse specializzate ed altre incognite nelle quali è possibile imbattersi. Per fornire un'esperienza di viaggio autentica e personalizzata è determinante utilizzare in modo combinato i risultati ottenuti dai dati raccolti e le intuizioni e i giudizi dell'uomo. Nonostante le molteplici difficoltà, i Big Data sono uno strumento utile per appoggiare la crescita del turismo e offrire esperienze uniche e dinamiche.

Key words: Big Data, Tourism, traveller's satisfaction.

1 Introduction

Data has grown to be a significant resource for many businesses in the age of the Internet, including the travel and tourist industry. Big Data, which is a term for vast, intricate, and ever-expanding datasets, presents innovative opportunities to comprehend travellers' behaviour, customize travel experiences, and boost operational effectiveness. The influence of big data on the travel and tourism sector will be examined in this article.

Big Data's capacity to offer insights on travellers' behaviour is one of the industry's key benefits. Businesses can better comprehend their clients by analysing massive volumes of data gathered from a variety of sources, including social media, smartphone applications, and online booking systems. Companies may adjust their services to satisfy consumer needs through utilizing such data to determine patterns and trends in travel preferences. Big Data also enables the personalization of travel experiences. By leveraging data on customer preferences, past behaviours, and demographic information, travel companies can create customized offerings that resonate with individual travellers. Hotels, for example, can utilize data analytics to pinpoint well-liked tourist spots, offer individualized itineraries, and promote extra services that can improve the visitor experience.

Airlines, for instance, can utilize data analytics to offer tailored flight suggestions based on the traveller's history and preferences, such as seat choices or meal selections. Similar to this, travel agents may make use of data to create custom holiday packages that take into account preferences for certain locations, activities, and lodging types. The possibility of repeat business is increased, and client satisfaction is improved by this level of personalisation.

The tourist sector could significantly improve operational effectiveness through the utilization of big data analytics. Businesses are able to identify bottlenecks, inefficiencies, and emphasizes for improvement in their operations by analysing massive datasets.

Another valuable application of Big Data in tourism is predictive analytics and forecasting. By analysing historical and real-time data, businesses can predict future trends, demand patterns, and market fluctuations. This information allows them to make proactive decisions and adapt their strategies accordingly. Predictive analytics, for instance, may be used by tourist boards to anticipate the number of visitors during peak seasons, allowing them to allocate resources and schedule infrastructure upgrades appropriately. Similar to this, travel companies may utilize data to predict client preferences and create customized marketing efforts to lure in potential customers.

Customer relationship management (CRM) in the tourist sector heavily relies on big data. Companies can gain more about consumer preferences, behaviour, and satisfaction levels by analysing customer data. This information helps businesses in building strong relationships with their customers by offering personalized recommendations, loyalty programs, and targeted marketing campaigns. For example, a hotel chain can leverage customer data to offer personalized promotions, upgrades, and special amenities to its loyal guests, thereby fostering customer loyalty and repeat bookings.

Big Data has revolutionized destination management by providing valuable insights into tourist flows, preferences, and behaviour patterns. Destination management organizations (DMOs) can leverage Big Data to understand the visitor experience, identify popular attractions, and develop strategies to manage tourist flows and mitigate overcrowding. By analysing data from various sources, such as social media platforms and tourism surveys, DMOs can gain real-time insights into visitor satisfaction, sentiment, and feedback, allowing them to make informed decisions for destination development and marketing initiatives.

2 Disadvantages of Big Data Analysis in Tourism

The potential risk to data security and privacy is one of the main issues with big data analysis. Businesses need to make sure that consumer data is safeguarded and managed in compliance with privacy regulations as they gather and evaluate enormous datasets. There is a risk of data breaches, unauthorized access, and misuse of personal information, which can harm customer trust and result in legal consequences for businesses.

The availability of trustworthy and high-quality data is essential for big data research. However, ensuring data quality can be challenging, as data may come from multiple sources and in different formats. Inaccurate or incomplete data can lead to biased analysis and poor decision making. Businesses need to invest in data cleansing, validation, and quality control processes to ensure the accuracy and reliability of the data used for analysis.

Big Data analytics implementation in the tourist sector requires skilled workers and adequate resources. Companies have to invest funds on hiring managers or training data scientists, data analysts, and IT infrastructure that can handle massive datasets. It can be costly and time-consuming to acquire the required resources and experience, especially for smaller organizations with limited budgets.

Big Data analysis raises ethical concerns regarding the collection, storage, and use of customer data. Companies need to ensure transparency and obtain proper consent from customers regarding data collection and usage. Additionally, businesses should be cautious about potential biases in data analysis that may lead to discriminatory practices or exclusion of certain customer groups.

While Big Data analysis provides valuable insights, there is a risk of overreliance on data at the expense of human judgment and intuition. Data should be used as a tool to inform decision making rather than replacing human expertise. It is essential to strike a balance between data-driven insights and human creativity in the tourism industry to maintain a personalized and authentic travel experience.

3 Conclusion

Big Data has entirely transformed the tourism sector, offering businesses with valuable insights into traveller behaviour, enabling personalization of travel experiences, enhancing operational efficiency, facilitating predictive analytics, and strengthening customer relationships. By harnessing the power of Big Data, companies can stay ahead of the competition, deliver exceptional customer experiences, and drive growth in the dynamic and ever-evolving world of tourism.

Big Data's importance in the tourism industry is projected to increase as technology advances, providing opportunities and influencing the way individuals travel in the future. However, it is important to address the potential disadvantages, such as data privacy concerns, data quality issues, resource requirements, ethical considerations, and the need for human expertise. By carefully navigating these challenges, businesses can leverage the power of Big Data analysis to gain a competitive edge and deliver exceptional experiences in the ever-evolving landscape of tourism.

References

1. Buhalis, D., and Law, R.: Tourism Analytics: Big Data, Artificial Intelligence, and Robotics. Springer (2015)
2. Gretzel, U., Sigala, M., Xiang, Z., and Koo, C.: Smart tourism: foundations and developments. Electronic Markets, 25(3), 179-188 (2015)
3. Neuhofer, B., Buhalis, D., & Ladkin, A.: Technology as a catalyst of change: Enablers and barriers of the tourist experience and their consequences. Springer (2015)
4. Xiang, Z., Du, Q., Ma, Y., and Fan, W.: A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65 (2017)
5. Zehrer, A., and Raich, F.: Big data in tourism: Changing the way we understand the tourist experience. Journal of Business Research, 78, 257-260 (2017)
6. Sigala, M.: Tourism and technology: Converging worlds. Routledge (2017)
7. Eggers, F., and O'Dwyer, M.: Smart tourism destinations: State-of-the-art, smartness maturity and an integrated planning and management framework. Journal of Destination Marketing & Management, 12, 93-105 (2019)
8. Liang, S., Schuckert, M., Law, R., and Chen, C.: Be a smart tourist: Conceptualizing and developing a smart tourist behavior model. Journal of Travel Research, 57(8), 1059-1079 (2018)
9. Xiang, Z., Du, Q., Ma, Y. and Fan, W.: Big data analytics in tourism and hospitality: A literature review. Journal of Travel Research, 57(3), 1-23 (2017)
10. Song, H., Witt, S. F. and Li, G.: Big data analytics in tourism and hospitality: A systematic literature review. Journal of Travel Research, 58(5), 849-865 (2019)

Neural network-based prediction of domestic tourists' length of stay in Italy

La previsione della durata del soggiorno dei turisti domestici in Italia mediante l'utilizzo delle reti neurali

Fabrizio Antolini and Samuele Cesarini

Abstract This study proposes a predictive approach to estimate the duration of domestic tourists' stays in Italy, utilizing microdata from the 2019 *Travel and Holiday* Survey provided by the Italian National Institute of Statistics (ISTAT). The main objective is to develop a predictive model using neural networks and deep learning methodology, leveraging the most relevant predictive variables identified in the scientific literature. Deep learning methodology offers significant advantages in detecting and interpreting complex patterns in data, overcoming the limitations of traditional models. The aim is to achieve an accurate and reliable predictive model, providing valuable insights for the tourism sector and tourism development policies. This study contributes to the understanding and prediction of the duration of domestic tourists' stays in Italy, fully exploiting the benefits of deep learning methodology. The obtained results will have a significant impact on the tourism sector, enabling the optimization of hospitality and tourism promotion strategies in the country. By combining the microdata provided by ISTAT with the deep learning approach, complex relationships between the predictive variables and the duration of the tourist stay can be identified providing a robust foundation for informed decisions in the field of tourism.

Abstract *Questo studio propone un approccio predittivo per stimare la durata del soggiorno turistico domestico in Italia, basandosi sui microdati dell'indagine Viaggi e Vacanze del 2019 forniti dall'ISTAT. L'obiettivo principale è sviluppare un modello predittivo utilizzando le reti neurali e la metodologia del deep learning, sfruttando le variabili predittive più rilevanti emerse dalla letteratura scientifica. La metodologia*

Fabrizio Antolini
University of Teramo; e-mail: fantolini@unite.it

Samuele Cesarini
University of Teramo; e-mail: scesarini@unite.it

di deep learning offre notevoli vantaggi nel rilevare e interpretare pattern complessi nei dati, superando le limitazioni dei modelli tradizionali. L'obiettivo è ottenere un modello predittivo accurato e affidabile, fornendo informazioni utili per il settore turistico e per le politiche di sviluppo turistico. Questo studio contribuisce alla comprensione e alla previsione della durata del soggiorno turistico dei visitatori domestici in Italia, sfruttando appieno i vantaggi della metodologia del deep learning. I risultati ottenuti avranno un impatto significativo nel settore del turismo, consentendo l'ottimizzazione delle strategie di accoglienza e promozione turistica del paese. L'uso combinato dei microdati forniti dall'ISTAT e dell'approccio di deep learning permetterà di identificare relazioni complesse tra le variabili predittive e la durata del soggiorno turistico fornendo una base solida per decisioni informate nel campo del turismo.

Keywords: Predictive modelling; Deep learning; Domestic tourism; Microdata analysis

1 Introduction

Tourism is a vital economic sector, and accurately predicting the duration of domestic tourists' stays plays a crucial role in tourism planning and development [5,8]. Different authors have demonstrated the importance of the length of stay on income generated by tourists at the destination [1,3,11,13]. Researchers have employed various statistical models to examine the factors influencing the length of tourist stays. However, in the existing literature, a common usage emerges regarding the explanatory variables employed to model the tourism length of stay [2]. Notably, numerous covariates, encompassing socio-demographic factors like tourists' country of origin, age, accommodation preferences, travel timing, and visited geographical areas, have exhibited statistical significance. Nevertheless, adopting a broader perspective, Gómez-Déniz and Pérez-Rodríguez [4] contend that the conventional models commonly employed may not be suited for this purpose. They underscore that data concerning tourism length of stay exhibits both overdispersion and bimodality or multimodality. As a result, employing traditional econometric methods such as OLS, count data models, or duration models could potentially yield misleading outcomes.

Moreover, only very few microeconomic studies on tourist demand have been focused on the analysis of this key element. This paper presents a novel approach to estimate the duration of domestic tourists' stays in Italy using microdata obtained from the 2019 *Viaggi e Vacanze* Survey conducted by the Italian National Institute of Statistics (ISTAT). The main objective is to develop a predictive model based on neural networks and deep learning methodology, incorporating the most relevant predictive variables identified in the scientific literature. By harnessing the power of deep learning, which excels in uncovering complex patterns, this study aims to overcome the limitations of traditional models and deliver an accurate and reliable predictive model. The paper is structured as follows. The next section provides a general overview of the microdata used and the variables included in the model. This is followed by a brief description of the neural network employed. Section 3 presents the results, with a particular focus on the root mean square (RSME) error and the predictive capability of the model. Section 4 concludes by presenting the main findings and discussing the practical and theoretical implications for tourism policy makers.

2 Data and Methodology

2.1 Data

According to Lin et al.'s study [8], explanatory variables that influence tourist behaviour can be categorized into four main groups: sociodemographic, economic, travel-related, and psychological. However, the selection of these variables is contingent upon the availability of relevant information. While micro-data often provide insights into sociodemographic, travel-related, and psychological variables, they may not consistently capture economic variables. Furthermore, researchers frequently encounter challenges in obtaining economic data due to non-response issues among the selected units. In cases where income information is not readily available, such as in the present study, researchers may resort to using proxy variables, such as occupation and education level, as suggested by Marcussen [10]. Based on the considerations and the relevant literature, the following categories and variables were utilized: *Sociodemographic factors* - Nationality; Sex; Age; Marital status; Region of origin; *Economic factors* - Occupation and Education; *Travel-related factors* - Region of destination; Month of travel; Accommodation; Transport; Transport organization; Accommodation organization; Group of travel party; Principal activities; *Psychological factors* - Travel motivation. The statistical survey from which the variables were obtained is the "*Viaggi e Vacanze*" conducted by the National Institute of Statistics. The available micro-data consists of three separate databases: individuals, excursions, and trips. The database utilized for this research specifically pertains to the trips archive. The reference year is 2019, with 4,393 records and 140 variables. By considering only residents who engaged in domestic tourist trips, a total of 3,001 observations were obtained. With the use of expansion coefficients (COEV) recorded for each record it was possible the estimation of the entire Italian domestic tourist population size. Consequently, a dataset containing 48 million and 410 thousand trips was reconstructed. In the neural network’s architecture, however, categorical variables need to be transformed into a numerical representation. This can be done using various techniques such as *one-hot encoding*. One-hot encoding converts each category into a binary vector representation, where each element indicates the presence or absence of that category. After the dataset is encoded as described earlier, the number of rows used in the model remains unchanged, while the number of columns (variables) increases from 16 to 128. By converting categorical variables into numerical representations, neural networks can effectively handle and learn complex relationships and patterns of “categorical” inputs.

2.2 Methodology

Neural networks, or artificial neural networks (*ANNs*), represent a class of *machine learning* models that draw inspiration from the structural and functional characteristics of the human brain [7,12]. ANNs find extensive application in various tasks, including pattern recognition, classification, regression, and decision-making. At their fundamental level, neural networks comprise interconnected nodes known as artificial neurons or units, organized in layers. The initial data is received by the input layer and subsequently transmitted through one or more hidden layers, ultimately reaching the output layer where predictions or results are provided by the network.

Each connection between neurons possesses a weight, determining the strength and significance of the signal transmitted between neurons. During the training process, these weights are adjusted based on observed data to minimize discrepancies between the network's predictions and the actual outcomes [6]. In a feedforward neural network, information flows in a unidirectional way, passing through the hidden layers until it reaches the output layer. This sequential progression enables the network to generate predictions based on the provided input data. A typical artificial neuron with *n* input dendrites can be represented by the formula that follows:

$$y(x) = f(\sum_{i=1}^{n} w_i x_i).$$

The *w* weights allow each of the *n* inputs, (*x*), to contribute a greater or lesser amount to the sum of input signals. The net total is used by the activation function *f(x)*, and the resulting signal, *y(x)*, is the output axon. However, the learning process in neural networks typically employs a technique called *backpropagation*. This algorithm involves propagating prediction errors from the output layer back through the network, subsequently adjusting the weights to reduce overall error. This iterative process continues until the network's performance reaches a satisfactory level [14].

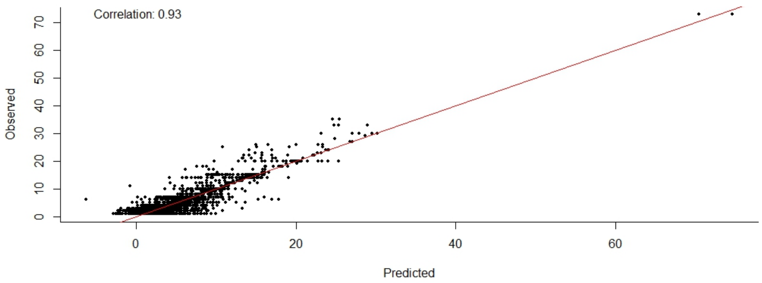
Deep learning encompasses the utilization of neural networks with multiple hidden layers, typically two or more, to address complex problems (*Multilayer Perceptron*). Deep learning architectures leverage the hierarchical representation learning capabilities of neural networks, enabling the capture and comprehension of intricate patterns and relationships within data. By incorporating additional hidden layers, deep neural networks can acquire progressively abstract and high-level features from input data. The design and configuration of neural networks, including the determination of layer count, unit count per layer, and activation functions employed, depend on the specific problem at hand. Careful selection of architecture and fine-tuning of hyperparameters are crucial for achieving optimal performance [6].

Below are presented the results provided by a *feedforward* single-layer network, (*single hidden layer*), and a multilayer perceptron neural (MPL) network with *two hidden layers* employing the *backpropagation algorithm*. To assess the predictive performance of the two models, two randomized samples, namely the train set and the test set, were constructed, containing 75% and 25% of the total observations, respectively. The train set was used to train the networks, while the test set was employed to make predictions for unseen data instances. How suggested in Lantz [6] since this is a numerical prediction problem rather than a classification problem, it is not possible to use a confusion matrix to evaluate the accuracy of the model. Instead, it is necessary to measure the correlation between the model's predictions on the test set and the actual recorded values. This provides an indication of the strength of the linear association between the two variables. Therefore, the predictive capacity of the models was evaluated by visually examining a scatter plot depicting the predicted values against the corresponding actual values. Finally, a comparison between the two models was conducted by assessing the standard deviation of the generated residuals.

3 Results

The first ANN employed a feedforward algorithm to train a neural network to predict the domestic Italian touristic length of stay. The architecture of this model consists of a single hidden layer comprising 16 units. The output layer of the network utilizes a *linear activation function* [15]. During the training process, 75% of the observations were randomly selected from the entire dataset to form the training set. The remaining 25% of the observations were retained as the test set, which was utilized to assess the model's predictive performance. The high correlation coefficient of 0.93 between the predicted and actual values indicates a strong positive linear relationship (Fig.1). This suggests that the model successfully captures the relevant features and patterns related to the duration of stay in the domestic Italian tourism context.

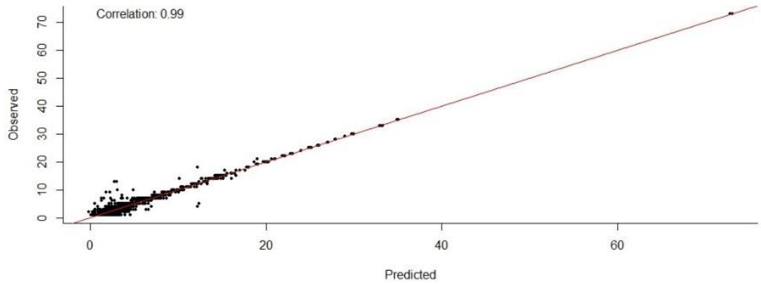
Fig. 1 F*eedforward single-layer network* values vs actual values



Source: Elaboration of authors

The relatively low *RMSE* value of approximately 1.97 suggests that, on average, the model's predictions deviate by a small margin from the actual duration of stay. This indicates that the model captures the underlying patterns and trends in the data effectively. The mean value of the target variable in the test set, which is 5.33, serves as a reference point for understanding the average duration of stay observed in domestic Italian tourism within the test data. As explained above, to improve the predictive capabilities of the model, it is possible to employ the backpropagation algorithm by adding an additional hidden layer. The scatter plot below (Fig. 2) depicts the relationship between the actual values recorded in the test set and the values calculated by a neural network trained on two hidden layers consisting of 16 and 8 neurons, respectively. In this architecture, information flows backward between the neurons to minimize the error between the prediction and the actual value.

Fig. 2 *Backpropagation two hidden layers MPL network* values vs actual values



Source: Elaboration of authors

As can be observed graphically, the performance of the model significantly improves, leading to a very high correlation (0.99) between the predicted values of tourist lengths of stay and the actual values. The effectiveness of the model is further supported by a dramatically reduced standard deviation of the residuals, which stands at 0.65. Overall, these results suggest that the proposed machine learning models shows promise in accurately predicting the duration of stay in domestic Italian tourism, indicating its potential usefulness in the tourism industry for decision-making and planning purposes.

4 Conclusions and Discussion

This study proposes a predictive approach using neural networks and deep learning methodology to estimate the duration of domestic tourists' stays in Italy. By leveraging microdata from the 2019 Travel and Holiday Survey provided by the Italian National Institute of Statistics (ISTAT), the model aims to provide valuable insights for the tourism sector and tourism development policies. The results demonstrate the effectiveness of deep learning in capturing complex patterns and achieving a strong positive linear relationship (correlation coefficient of 0.93) between predicted and actual values. The model's low RMSE value (approximately 1.97) indicates its ability to accurately predict the duration of stay in domestic Italian tourism. By incorporating the backpropagation algorithm and adding an additional hidden layer, the model's performance further improves, resulting in a very high correlation (0.99) and a significantly reduced standard deviation of the residuals (0.65). The findings of this study have practical implications for policymakers in the tourism industry. The accurate prediction of the duration of tourists' stays can inform and guide decision-making processes aimed at increasing the length of stay in various destinations. By utilizing the insights provided by the predictive model, policymakers can develop targeted strategies and interventions to optimize hospitality services, enhance tourism promotion efforts, and create engaging experiences that encourage tourists to extend their visits. These results offer valuable information for formulating effective policies and initiatives aimed at maximizing the economic benefits derived from tourism.

References

1. Aguiló, E., Rosselló, J. and Vila, M.: Length of stay and daily tourist expenditure: A joint analysis. Tourism Management Perspectives, 21, 10-17 (2017) https://doi.org/10.1016/j.tmp.2016.10.008.
2. Almeida, A., Machado, L. P., and Xu, C.: Factors explaining length of stay: Lessons to be learnt from Madeira Island. Annals of Tourism Research Empirical Insights, 2(1), 100014 (2021) https://doi.org/10.1016/j.annale.2021.100014.
3. Antolini, F., Cesarini, S. and Simonetti, B.: Italian tourist expenses' determining factors: A machine learning approach. Quality & Quantity in Methods in the tourism and hospitality industry, in press (2023)
4. Gómez-Déniz, E. and Pérez-Rodríguez, J. V.: Modelling bimodality of length of tourist stay. Annals of Tourism Research, 75, 131-151 (2019) doi:10.1016/j.annals.2019.01.006.
5. Gössling, S., Scott, D. and Hall, C. M.: Global trends in length of stay: Implications for destination management and climate change. Journal of sustainable tourism, 26(12), 2087-2101 (2018) doi:10.1080/09669582.2018.152977.
6. Lantz, B.: Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd. ISBN: 9781788295864 (2019)
7. LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning. nature, 521(7553), 436-444 (2015) 10.1038/nature14539.
8. Li, K. X., Jin, M. and Shi, W.: Tourism as an important impetus to promoting economic growth: A critical review. Tourism Management Perspectives, 26, 135-142 (2018) https://doi.org/10.1016/j.tmp.2017.10.002.
9. Lin, V. S., Qin, Y., Li, G. and Wu, J.: Determinants of Chinese households' tourism consumption: Evidence from China Family Panel Studies. International Journal of Tourism Research, 23(4), 542-554 (2020) doi:10.1002/jtr.2425.
10. Marcussen, C. H.: Determinants of Tourist Spending in Cross-Sectional Studies and at Danish Destinations. Tourism Economics, 17(4), 833–855 (2011) doi:10.5367/te.2011.0068.
11. Marrocu, E., Paci, R. and Zara, A.: Micro-economic determinants of tourist expenditure: A quantile regression approach. Tourism Management, 50, 13-30 (2015) doi:10.1016/j.tourman.2015.01.006.
12. Kriegeskorte, N. and Golan, T.: Neural network models and deep learning. Current Biology, 29(7), R231-R236 (2019) https://doi.org/10.1016/j.cub.2019.02.034.
13. Park, S., Woo, M. and Nicolau, J. L.: Determinant factors of tourist expenses. Journal of Travel Research, 59(2), 267-280 (2020) https://doi.org/10.1177/0047287519829257.
14. Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors. nature, 323(6088), 533-536 (2016) https://doi.org/10.7551/mitpress/4943.003.0042.
15. Sharma, S., Sharma, S. and Athaiya, A.: Activation functions in neural networks. Towards Data Sci, 6(12), 310-316 (2017)

The management of cultural heritage in contexts of undertourism: a model for assessing the economic sustainability of public-private partnerships

La gestione del patrimonio culturale in contesti di undertourism: un modello di valutazione della sostenibilità economica dei partenariati pubblico-privati

Francesco Calabrò

Abstract The article briefly illustrates a model for assessing the economic sustainability of cultural heritage management models which envisage forms of public-private partnership. The model was created, in particular, to support public decision-makers in territorial contexts in which there are limited tourist flows.
Abstract *L'articolo illustra sinteticamente un modello per la valutazione della sostenibilità economica dei modelli di gestione del patrimonio culturale che prevedono forme di partenariato pubblico-privato. Il modello nasce, in particolare, per supportare i decisori pubblici in contesti territoriali nei quali sono presenti flussi turistici limitati.*

Key words: management models, economic sustainability, economic evaluation, undertourism

1 The economic context of reference

The sustainability of cultural heritage management is today one of the most difficult challenges for the purpose of transmitting to future generations the set of knowledge, values and traditions inherent in it [11, 8].

Fabrizio Calabrò
Mediterranea University of Reggio Calabria, e-mail: francesco.calabro@unirc.it

Beyond the transitory inebriation from PNRR, the condition of progressive contraction of available public resources, far from being one of the well-known cyclical phases, presents itself more and more as a structural condition to be reckoned with, also in future prospects, as highlighted by many authors [12, 13]. This contraction, as has been evident for some time now and similarly to what is happening for other sectors, also concerns the public resources available for the conservation and enhancement of cultural heritage [18].

This is the reality which needs to be acknowledged and which makes the search for innovative solutions indispensable, which allow the pursuit of the objectives of conservation and enhancement of the heritage in any case: the Codice dei Beni Culturali e del Paesaggio [7] has implicitly acknowledged this, as transpired in particular by the procedures envisaged for the involvement of private subjects in the management of assets.

Below is a summary illustration of an evaluation model, still in an experimental phase, of a monetary nature, the SostEc Model [5]: the model was created with the aim of providing the public an easy-to-use tool, to identify in advance the type of manager, in the hypothesis of indirect management of cultural assets envisaged by art. 115 of the Code.

1.1 Heritage management

The forms of cultural heritage management envisaged by article 115 are essentially two, direct or indirect: in the economy of this contribution, we will focus in particular on the indirect form.

The indirect form of management is implemented by resorting to the institution of the concession of valorisation activities to third parties, through public tender procedures, on the basis of the comparative evaluation of specific projects [4].

The purpose of the assignment in indirect management, as specified by paragraph 4 of the article, is to ensure a better level of enhancement of the cultural heritage [2, 3].

The same paragraph 4 introduces a fundamental concept: the choice between direct and indirect management is not made arbitrarily, but through a comparative assessment in terms of economic-financial sustainability and effectiveness, on the basis of previously defined objectives.

This contribution aims to deepen the economic-estimative and evaluation tools capable of verifying the feasibility and economic sustainability of innovative solutions for the management of cultural heritage, with particular reference to the architectural and landscape heritage.

The need to guarantee the economic-financial balance in the processes of valorisation of the cultural heritage is even more felt in the presence of forms of public-private partnership. The approach used starts from a basic observation: in relation to managers of different nature (profit or non-profit), different cost structures can be associated, while the revenue structure for the specific asset in question can be considered invariable.

1.2 Metrics: a small contribution

Speaking of entrusting the management to private subjects, it may be useful first of all to understand what a business is and what it is not, even if it falls within the field of cultural activities carried out by private subjects.
To do this, it is possible to resort to the aid of monetary valuation techniques, usually used in company analyses, such as for example the Cost-Revenue Analysis and the Break-even Analysis [15, 16].

At the end of the evaluation, what will be most relevant will not be the distinction of the juridical nature of the subjects, between profit and non-profit, but of the economic nature of the activities: that is, between activities endowed with the capacity to generate financial flows so conspicuous as to generate income and allow adequate returns on investments and/or management costs and, on the other hand, assets generating cash flows capable of covering only a portion of management costs, as in the case of contexts of undertourism.

In fact, to the purpose of this Article, the most relevant distinction is related to the nature of the activity, not of the subject (Table 1): there are, in fact, numerous cases of subject with not-for-profit nature but that carry out economic activities. In such cases does not change the cost structure between a not-for-profit or profit subject, but the tax regime to which they are subject, in addition, of course, the prohibition for not-for-profit organizations to distribute profits [9, 17].

Table 1 Nature of the subjects by type of activity

TYPE OF ACTIVITY	NATURE OF MANAGING ENTITY
Public services	Public
Activities of public interest without economic relevance	Private not-for-profit
Economic activities of public interest with significant revenues, such as to be regarded as economic activities	Private not-for-profit Private for profit

The moment in which the juridical nature of the subjects is detected is only in the stage of estimating management costs, i.e. if we are dealing with profit subjects, for which adequate levels of profit are also conventionally calculated in this section, and subjects for which this it is not needed.

The further distinction, then, between non-profit entities, is the one that once again looks at the economic nature of the activities: that is, whether they have economic significance, thus determining a cost structure very similar to that of businesses, or if, on the other hand, they are voluntary in nature and, as such, determine a profoundly different structure of management costs (especially for human resources).

In the latter case, the case in which, by virtue of the reduction in costs for human resources, the revenues are still able to cover the other management costs, in particular those for the extraordinary maintenance of the assets, appears to be of particular relevance.

It is evident that this is a case of particular relevance for the purposes of sustainability and the transmissibility of heritage to future generations [14, 6]: it is precisely the coverage (also) of extraordinary maintenance costs that allows the conservation of assets over time.

In the latter case, however, very widespread in contexts characterized by very limited levels of tourist presences, it is the local communities that are called directly into question: if they recognize the value of the property and the tangible and intangible effects that it can generate its enjoyment (usability), then they must directly take charge of the related services through direct and voluntary commitment. In the absence of this commitment and with the current (and future) budget constraints of public entities, assets are inevitably destined to disappear.

2 The evaluation model

The SostEc model, subject of publications to which we refer for a more complete illustration, uses a simple algorithm applicable in the initial phase of the decision-making process linked to the enhancement of an asset, especially of an architectural nature. It makes it possible to identify, as a preliminary step, the type of manager most suitable for the specific asset in question, in relation to its ability to generate revenue streams.

The model uses two variables:
v1: nature of the activities carried out and of the subjects involved (public; private for profit; private non-profit with profit and non-profit activities);
v2: attractiveness of the good (intrinsic attractiveness; extrinsic attractiveness);
The obtainable results allow us to understand:
r1: the composition of investment capital (public; private; mixed public-private)
r2: the type of manager
r3: the entity of the charges that can be charged to the manager, guaranteeing the economic sustainability of the activity and the maintenance of the asset for future generations.

In practice, the SostEc Model allows for the verification of the economic sustainability of the management models, i.e., whether the conditions exist, first of all economic sustainability, for entrusting the management to private entities and which public-private partnership model is truly sustainable and more suitable for the present case.

The SostEc Model uses a typical tool in these cases, the cash flow analysis, capable of verifying with relative ease whether the expected revenue flow is equal to or higher than the estimated costs or not, in relation to the different cost structures of the different types of managers.

Basing on the nature of the activities and of the managing entity, it is possible to hypothesize the following three models (Table 2):

MANAGING MODEL	NATURE OF ENTITY AND ACTIVITIES
Model P - Profit	Entity profit, activities profit
Model NP - Not-for-profit	Entity not-for-profit, activities not-for-profit
Model M - Mixed	Entity not-for-profit, activities profit

Table 2 Managing models by type of subject

3 Profitability and public-private partnership forms

In order to understand whether private parties may be interested in some form of partnership, it is necessary to first examine whether equilibrium conditions can be applied in the budget, and, where appropriate, if there is a management surplus sufficient to ensure sufficient profitability for a possible private capital investment.

From this verification will depend in general whether the project is feasible or not, but also what type of private entity can be involved as a partner.

In relation to the capacity of the asset in question to generate revenue, in theory it can be hypothesized five different conditions of profitability (Table 3):

Band A. High profitability
Band B. Medium to high profitability
Band C. Average profitability
Band D. Lower-middle profitability
Band E. Low profitability

	INVESTMENT COSTS		MANAGING COSTS	
Band A. High profitability				
Band B. Medium to high profitability	1 - μ	μ		
Band C. Average profitability				
Band D. Lower-middle profitability				
Band E. Low profitability			1 - ε	ε

Private for profit

Private non-profit, activities non-profit

Public

Table 3 Distribution of investment and managing costs between public and private entities

Then, there is the case of insufficient profitability or nothing (sixth profitability assumptions), which implies the absence of the minimum conditions for any form of public-private partnership and entrusts exclusively to public bodies the responsibility to make available the particular case of asset. This assumption, however, is less and less feasible in reality, due to the progressive decline of available resources in the delivery of public services [1, 10].

4 Conclusions

The model is susceptible to improvements: an aspect which will certainly be dedicated in depth, in terms of the cost structure, is how the incidence of human resources varies, for example, if the profit subject resorts to the new forms of contractualization envisaged by the innovations in the field.

Further insights, then, will concern a more precise distinction between those that are work-related activities and voluntary activities.

In fact, it is appropriate to specify that this is a field that presents various risks in terms of social sustainability: it is necessary to clearly distinguish what is voluntary, free commitment from what is work activity. The valorisation of cultural resources must constitute an opportunity for development for the territories and not a further terrain for the precariousness of work or, worse, an incentive for forms of irregular work.

References

1. Aas, C., Ladkin, A., and Fletcher, J.: Stakeholder collaboration and heritage management. Annals of tourism research, 32(1), 28-48 (2005)
2. Barile S., Montella M. and Saviano M.: “Enhancement, Value and Viability of Cultural Heritage. Towards a Service-Based Systems Approach”, in: Gummesson E., Mele C., Polese F. (eds), The 2011 Naples Forum on Service. ServiceDominant Logic, Network & Systems Theory and Service Science: integrating three perspectives for a new service agenda, Giannini Editore, Napoli (2011)
3. Barile S. and Saviano M.: Dalla Gestione del Patrimonio di Beni Culturali al Governo del Sistema dei Beni Culturali in: Golinelli, G.M., (ed), Patrimonio culturale e creazione di valore, Verso nuovi percorsi, Cedam, Padova, pp. 97-148 (20129
4. Bilancia, P.: La valorizzazione dei beni culturali. Modelli giuridici di gestione integrata (Vol. 642). FrancoAngeli (2006)
5. Calabrò, F.: Integrated programming for the enhancement of minor historical centres. The SOSTEC model for the verification of the economic feasibility for the enhancement of unused public buildings. La programmazione integrata per la valorizzazione dei centri storici minori. Il Modello SOSTEC per la verifica della fattibilità economica per la valorizzazione degli immobili pubblici inutilizzati. ArcHistoR, 13(7), pp. 1509–1523 (2020) DOI: 10.14633/AHR280
6. Council of Europe: Framework Convention on the Value of Cultural Heritage for Society, Council of Europe Treaty Series - No. 199, Faro, Portugal (2003)
7. Decreto Legislativo 22 gennaio 2004, n. 42, Codice dei beni culturali e del paesaggio
8. Drury, P. and McPherson, A.: Conservation principles: policies and guidance for the sustainable management of the historic environment (2008)
9. Francesconi, A.: Comunicare il valore dell'azienda non profit. Wolters Kluwer Italia (2007)
10. Franch, M.: Le frontiere manageriali per la valorizzazione della cultura e dell’arte. Cultura, arte e management: frontiere e connessioni, 95-107 (2010)
11. Fusco Girard, L., Cerreta, M., De Toro, P. and Forte, F.: The human sustainable city: Values, approaches and evaluative tools (pp. 65-93). London: Routledge (2007)

12. Gualerzi, D.: The Coming of Age of Information Technologies and the Path of Transformational Growth.: A long run perspective on the 2000s recession. Routledge. pp.: 20-22 (2009)
13. Hagemann H., Seiter. S.: "Growth, Productivity, and Employment: Consequences of the New Information and Communication Technologies in Germany and the US" in: Growth, Distribution, and Effective Demand: Alternatives to Economic Orthodoxy: Essays in Honor of Edward J. Nell. M. E. Sharpe Inc., New York (USA). p. 98 (2004)
14. ICOMOS: International Cultural Tourism Charter. Managing Tourism at Places of Heritage Significance. Adopted by ICOMOS at the 12th General Assembly in Mexico (1999)
15. Morano P. and Tajani F.: The Break-Even Analysis applied to urban renewal investments: a model to evaluate the share of social housing financially sustainable for private investors. HABITAT INTERNATIONAL, vol. 59, p. 10-20 (2017) ISSN: 0197-3975, doi: 10.1016/j.habitatint.2016.11.004
16. Nesticò, A. and Maselli, G.: Declining discount rate estimate in the long-term economic evaluation of environmental projects. Journal of Environmental Accounting and Management, Vol. 8, Issue 1, pp. 93-110 (2020) https://doi.org/10.5890/JEAM.2020.03.007
17. Propersi, A.: Il sistema di rendicontazione negli enti non profit. Dal bilancio d'esercizio al bilancio di missione. Vita e pensiero (2004)
18. Sutherland, A.: Fiscal crises and aggregate demand: can high public debt reverse the effects of fiscal policy? Journal of public economics, 65(2), 147-162 (1997)

Solicited Session SS21 - *Statistical methods and models for land monitoring with spatio-temporal data*

Organizer and Chair: Maurizio Carpita

1. *Geo-referenced data and complex networks for measuring road accident risk* (Cantaluppi G., Clemente C., Della Corte F. and Zappa D.)
2. *A comparison of geospatial models for car crash risk* (Cantaluppi G., Giardino G. and Zappa D.)
3. *Geostatistical modelling of livestock-related $PM_{2.5}$ pollution and scenario analysis for policymakers - Work in progress* (Fassò A., Rodeschini J., Fusta Moro A. and Finazzi F.)
4. *Functional clustering methods for space-time big data from mobile phone networks* (Perazzini S., Metulini R. and Carpita M.)

Geo-referenced data and complex networks for measuring road accident risk

Dati geo referenziati e reti complesse per la misurazione del rischio di incidente stradale

Gabriele Cantaluppi, Gian Paolo Clemente, Francesco Della Corte and Diego Zappa

Abstract The assessment of risk related to car crashes in road networks is a relevant topic for both social impact and the related political/administrative decisions. To this end, we show how the spatial objects and the information concerning the structure of the roads along with the crash history can be used to map the risk related to any road of a network. In particular, we follow a combined approach. On the one hand, a statistical model is developed in order to assess the risk on the basis of a set of features related to the characteristics of the streets. On the other hand, from the spatial object we build a weighted network, where the assessed risk of each segment is used as a weight. We study the topology structure of the graph and we show how classical network indicators can provide meaningful insights about the risk of an area.

Abstract *La misurazione del rischio di incidente stradale è quanto mai rilevante sia per le implicazioni sociali che le scelte amministrative e politiche che da queste possono derivare. Con questo contributo si vuole mostrare come combinando le informazioni inerenti la rete stradale e la geolocalizzazione di incidenti sia possibile mappare il rischio per l'intera rete stradale. A tale proposito modelli statistici affiancati da indicatori di complessità di un network sono risultati efficaci al raggiungimento dello scopo.*

Gabriele Cantaluppi
Department of Statistical Sciences - Università Cattolica del Sacro Cuore e-mail: `gabriele.cantaluppi@unicatt.it`

Gian Paolo Clemente
Department of Mathematics for Economic, Financial and Actuarial Sciences, e-mail: `gianpaolo.clemente@unicatt.it`

Francesco Della Corte
Department of Mathematics for Economic, Financial and Actuarial Sciences e-mail: `francesco.dellacorte1@unicatt.it`

Diego Zappa
Department of Statistical Sciences - Università Cattolica del Sacro Cuore e-mail: `diego.zappa@unicatt.it`

Gabriele Cantaluppi, Gian Paolo Clemente, Francesco Della Corte and Diego Zappa

Key words: Car accident risk, Complex Networks, Penalized regression, Geospatial modelling

1 Introduction

Worldwide, road accidents represent one of the leading causes of death (see, e.g. [1]) and they are one of the most serious public health issues in the world. As reported by the World Health Organization [2], car crashes are responsible for more than 1 million deaths each year (17 deaths per 100,000 people) and are the eighth leading cause of death for all age groups. In 2019, road injuries leave the overall list of top causes of death, but remain in 7th place for the same list regarding only low-income countries and in 10th place in the analogous lists in both lower and upper-income countries . For these reasons, the quantification of road risk assumes not only scientific but also social relevance. In middle-high-income nations, such as Italy, roads are safer than the global average, but car crashes have still a very high impact on society and in particular on families touched by these events.

In this paper we will focus on road accidents in Italy but we provide a methodology that can be suitably extended to any road network. To give a detailed view of the situation in this country, according to the latest statistics published by the Italian national statistical institute in the Annual report on Road Casualties of 2021 [3], accidents and injuries decrease in January and February and increase substantially in March-June if compared to 2020, thus returning to levels very close to the pre-pandemic period also in the second half of the year. In 2021, there were 2,875 deaths in road accidents in Italy (+20% from the previous year), 204,728 injuries (+28.6%) and 151,875 road accidents (+28.4%). Hopefully they were decreasing in comparison with 2019 (-9.4% fatalities, -15.2% injuries and -11.8% accidents). There were 2,397 fatalities within 24 hours of the accident event while there were 478 fatalities from the second to the 30th day after the event.

We contribute to the existing literature presenting a computationally efficient method to map the risk of accidents at the road levels of any areas. To this end, we consider the full complexity of road networks and their many features, the local spatial dependence structure and, depending on the availability of the data, also time dependence. To achieve this result, we have merged information from open access databases (OSM, [4]), proprietary databases with additional road features and the traffic at the street level, public datasets on car crashes location and social/demographic features of areas at the census level.

Fig. 1 shows the datasets we joint to fit the model.

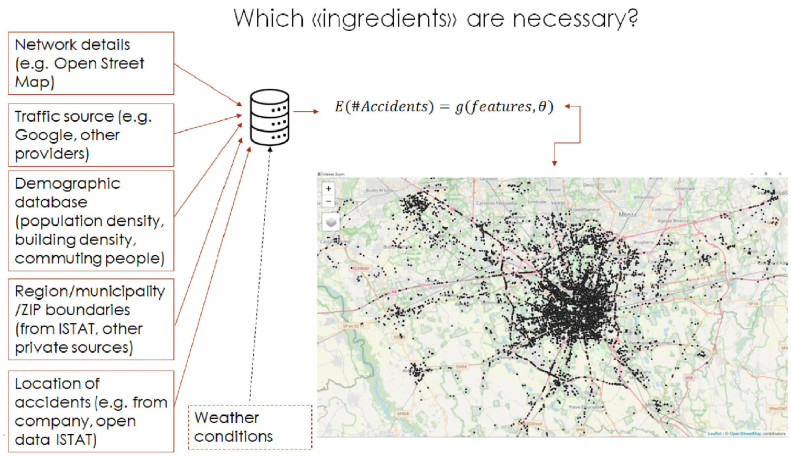


Fig. 1 Databases used to estimate the risk of a road network. On the right the location of accidents in the municipality of Milan

2 The model

The most common approach to model this kind of outcomes is the conditionally autoregressive (CAR) model where the dependent variable conditional on its neighbors follows a Poisson distribution with average λ equals to

$$\log(\lambda) = \mu + \psi + \theta. \tag{1}$$

μ includes the linear combination of fixed effect as it follows

$$\mu = \mathbf{x}^T \boldsymbol{\beta}. \tag{2}$$

$\mathbf{x}$ is the profile of the region/link/road and $\boldsymbol{\beta}$ is a vector of regression coefficients. To consider the role coped with spatial covariates, ψ and θ are respectively a spatially structured random effect and an unstructured effect, i.e. a location specific component. For identification purposes, the above model is typically treated within a Bayesian framework.

The main drawbacks of the above models are threefold. First the metric used to compute the distance between spatial objects is often the Euclidean distance or the adjacency matrix. We propose to use the real distance to join two links according to the the minimum directed shortest path connecting them considering also the direction (i.e. including also the real street navigation from one point to another one). Second, in many applications, accident locations are often clustered into areas also for a matter of computational burdens. That implies one cannot estimate the risk at a street location, where the dimension of the dataset generally contains thousands of records. Thanks to efficient procedures, we are able to propose a method that works at the street level. Few examples are available in the literature (e.g. [5], [6]). Third, to estimate accurately the risk of accidents, we exploit a very highly detailed set of the link characteristics. On the one side the availability of a large set of covariates adds computational burdens because of the high dimensional information set and because some characteristics are available in some places but not in others. On the other side what sounds interesting is that this framework may address a substantial simplification of the standard CAR model or its generalization. Since the spatial random effects are usually added as surrogates of unobserved factors, as the number of explanatory variables increases the less will be the covered by this component to capture the heterogeneity across the regions.

To deal with the issues listed above we propose to recover a mix of two spatial autoregressive models (SLM and SLX, see e.g. [7] for details and further references) from the econometric literature. Among the many SLM formulation, we join the CAR model and the spatial lagged model to consider the influence of the neighboring network structure. Then, we use

$$\log(\lambda) = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_{-\mathbf{1}}\boldsymbol{\Delta} \tag{3}$$

where ρ modules the spatial dependence among occurrences. This component is quite interesting to consider areas that are more likely prone to accidents. $\mathbf{X}_{-1}$ denotes the matrix $\mathbf{X}$, where the column regarding the intercept has been removed. $\mathbf{W}$ is the matrix of distances. $\boldsymbol{\Delta}$ is a vector of coefficients.

3 A case study

Fig. 2 shows the outcome of the fitted risk to the municipality of Milan (city and province). This is just an example because we were able to fit the model to the whole Italian road network. some details are necessary: results are affected by a bit of bias in the data, that is somehow unavoidable in this type of research. For instance, an important feature, represented by the number of road crossings, is not directly available in the databases and, therefore, it is computed as the number of segments that have in common one coordinate with that road. This method represents an approximation of the true crossings but it returns an estimate quite close to reality. Moreover, coordinates of accidents are not always strictly in line with its segment. Misalignments are due to proxies implicit into the reverse geocoding algorithms or to errors in the registration of accident locations. To include traffic we have used the Vehicle Miles Traveled (VMT) as an offset. It is defined as the product of the length of the segment and the volume of traffic and it has the advantage

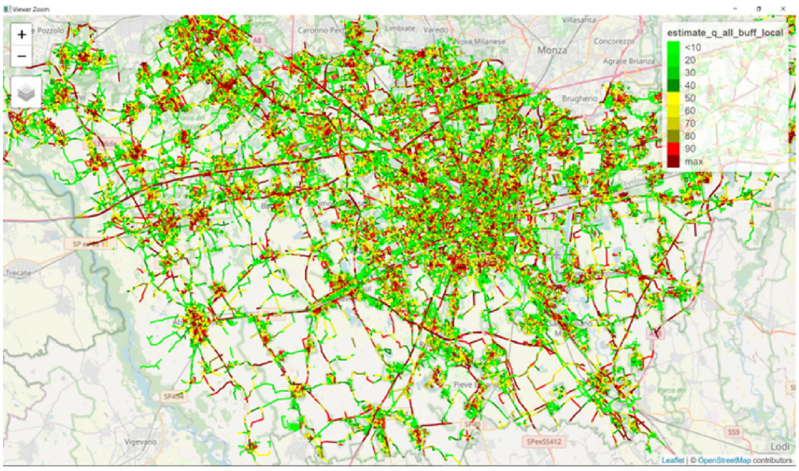


Fig. 2 Ropd network of Milan municipality. Green lines denote roads with low risk of accidents. Red lines denote roads highlky expose dto the risk of accident

to provide a measure of exposure at risk in each road. In Fig. 2 the roads of Milan municipality is coloured according to the decile of the estimated risk of crashes.

To include spatial dependence at the municipality level, we have split the domain into disjoint subregions. For each subset all the adjacent subregions are considered. We apply a model fitting data considering the region of interest incremented by a spatial buffer. The aim is indeed to get the model that assures both the best adaptation to the peculiar details of an area and, at the same time, considers details and characteristics observed not too far from the domain of interest. Furthermore the spatial object and the accident risk assessed by the model for each road are then converted in a directed and weighted graph. In particular, we focus on a ”junction graph”, where each segment is an arc and nodes are given by junctions (or by termination of closed streets). Each arc is then weighted according to the risk of the segment detected at previous step. Focusing on network topological indicators, we observe a significant correlation between the risk associated to a node and the node betweenness measured on the network. Therefore, the centrality of a node in the topological structure appears related to the risk measured by the model. Additionally, by means of the Louvain methodology, we detect communities in the area. The communities depend on both the arc density and on the weights. The split of the area into clusters can be used by insurance companies to measure the propensity to get an accidents in the neighbour of a point, and then to fine tune the cost of premiums to be paid to drive a car.

References

1. Nantulya, V.M. and Reich, M.R.: The Neglected Epidemic: Road Traffic Injuries in Developing Countries. BMJ, 324, 1139-1141. http://dx.doi.org/10.1136/bmj.324.7346.1139 (2002)
2. Worl Health Organization: Global status report on road safety 2018. ISBN 978-92-4-156568-4. Available at https://apps.who.int/iris/rest/bitstreams/1164010/retrieve
3. ISTAT: Road accidents. (2021) Available at https://www.istat.it/it/files//2022/07/REPORT_INCIDENTI_STRADALI_2021_EN.pdf
4. Open street Map (2023) https://www.openstreetmap.org/#map=5/43.769/9.404
5. Borgoni, R, Gilardi, A, Zappa, D. : Assessing the Risk of Car Crashes in Road Networks, Social Indicators Research (2020) doi 10.1007/s11205-020-02295-x
6. Gilardi, A, Mateu J, Borgoni, R, Lovelace, R, : Multivariate Hierarchical Analysis of Car Crashes Data Considering a Spatial Network Lattice, Journal of the Royal Statistical Society Series A: Statistics in Society, 185(3)1150–1177, (2022) doi 10.1111/rssa.12823
7. Halleck Vega, S. and Elhorst, J.P. : The SLX model. Journal of regional science, 55: 339-363 (2015)

A comparison of geospatial models for car crash risk

Confronto tra modelli geospaziali per la valutazione del rischio di incidente stradale

Gabriele Cantaluppi, Giorgio Giardino and Diego Zappa

Abstract The assessment of risk related to car crashes in road networks is a topic relevant for both social impact as well as for the related political/administrative decisions. By means of Monte Carlo simulation we study the performance of different models aimed at evaluating the number of crash occurrences that can be expected at each road street segment level in a network of streets in a certain time window.
Abstract *La misurazione del rischio di incidente stradale è essenziale sia per le implicazioni sociali che per le scelte amministrative e politiche che da queste possono derivare. Si presentano i risultati di una simulazione Monte Carlo per confrontare alcuni modelli geospaziali che possono essere utilizzati per stimare il numero di incidenti a livello di segmento di una rete stradale in un generico periodo temporale.*

Key words: Car crash risk, generalized linear model, INLA, geospatial modelling

1 Introduction

Aim of the contribution is to describe different methods that can be used to evaluate the number of crash occurrences that can be expected in a year at each road street segment level in a network of streets. Section 2 briefly presents models and methods for geospatial analysis of car accidents and Section 3 compares models by means of a simulation study.

Gabriele Cantaluppi
Faculty of Economics, Department of Statistical Science, Università Cattolica del Sacro Cuore, Largo A. Gemelli, 1, 20123 Milan, Italy, e-mail: gabriele.cantaluppi@unicatt.it

Giorgio Giardino
Generali Italia, e-mail: giorgio.giardino1997@gmail.com

Diego Zappa
Faculty of Banking Finance and Insurance Sciences, Department of Statistical Science, Università Cattolica del Sacro Cuore, Largo A. Gemelli, 1, 20123 Milan, Italy, e-mail: diego.zappa@unicatt.it

2 Models

Let us consider a network of n road segments and let Y_i be the response variable describing the number of crashes occurred at road segment i, $i = 1,2,\ldots,n$. Y_i is a count variable that can take values $0,1,2,\ldots$ [2].

Let X_i be a vector of $1+p$ covariates, the first one corresponding to the constant term. The covariates can be metric, count or dummy variables. In case of dummy variables they can describe the type of street segment (highway, traffic circle, one-way street, etc.) or denote the presence of some signals on the street (speed limits, traffic-lights, etc.).

The adjacency structure of the n road segments is resumed by the so-called adjacency matrix W, an $n \times n$ binary matrix with normalized rows (row totals are all equal to 1)

$$w_{is} = I(i,s)/n_i, \quad i,s = 1,2,\ldots,n \tag{1}$$

where $I(i,s)$ is the indicator function that takes value 1 if segments i and s are directly linked, that is if they share a common vertex; $n_i = \sum_s I(i,s)$ is the number of neighbours of segment i, that is the number of segments that are directly linked to segment i.

Outcome Y_i can be described by means of a classical Poisson model

$$Y_i|\lambda_i \sim Poisson(\lambda_i) \tag{2}$$

with link function

$$\log(\lambda_i) = X_i\beta \tag{3}$$

describing the crashes rate λ_i on segment i, as well by means of three-level hierarchical models [6].

In case of hierarchical models we assume, at the first stage of the hierarchy, that

$$Y_i|\lambda_i \sim Poisson(\lambda_i) \tag{4}$$

where λ_i represents the crashes rate on segment i. The link function for λ is assumed at the second stage of the hierarchical model as

$$\log(\lambda_i) = X_i\beta + \theta_i + \phi_i \tag{5}$$

where θ_i is an error component following a Normal distribution and ϕ_i is a spatially structured random effect. The two random effects θ_i and ϕ_i represent the unstructured and structured spatial components and are defined differently in different models as discussed below. The third stage of the hierarchical model concerns the specification of prior and hyperprior distributions [8, 6]. An uncertain $N(0,1000)$ prior is assigned to the $p+1$ parameters in β. LogGamma priors with parameters $(0.01,0.01)$ and logit-Beta priors with parameters $(1,1)$ are considered for precision parameters and spatial autocorrelation parameters, respectively.

The models here considered in order to take into account some possible spatial structure features are described in the following paragraphs, see also [8].

Spatial lag model for spatial effects The spatial lag model for spatial effects, known also as spatial autoregressive model (SLM-SAR) has link[1] functions for all segments $\lambda = (\lambda_1, \ldots, \lambda_n)$ that can be jointly modeled as

$$\log(\lambda) = (I_n - \rho W)^{-1}(X\beta + \theta) \tag{6}$$

where X is the matrix of covariates, with rows X_i, $i = 1, 2, \ldots, n$, corresponding to all row segments, ρ is a spatial autocorrelation parameter, and $\theta \sim \mathcal{N}(0, \sigma^2 I_n)$ is a zero mean Gaussian noise with covariance matrix $\sigma^2 I_n$.

Spatially lagged X model The link functions of the spatially lagged X model model (SLX) for all segments $\lambda = (\lambda_1, \ldots, \lambda_n)$ can be jointly modeled as

$$\log(\lambda) = X\beta + WX\gamma + \theta \tag{7}$$

where WX is the lagged matrix of covariates, that introduces a dependence of segment i from the covariate values of road segments that are neighbour of i. Extensions of this model are the spatial Durbin model and the spatial Durbin error model, [8].

Spatial Durbin model The link functions of the spatial Durbin model (SDM) for all segments $\lambda = (\lambda_1, \ldots, \lambda_n)$ can be jointly modeled as

$$\log(\lambda) = (I_n - \rho W)^{-1}(X\beta + WX\gamma + \theta) \tag{8}$$

it allows for spatial correlation of the outcome as well as it considers the lagged matrix of covariates as regressors like the SLX model.

Conditional AutoRegressive model This model is introduced as premise and for simulating the following ICAR. As [10] note, the spatially structured random effect in model (5) are usually surrogates of unobserved factors that vary smoothly over the space. In our case two neighbouring roads segments are supposed to be more dependent than two more distant segments. The link functions of a conditional autoregressive model (CAR) for all segments can be jointly modeled as

$$\lambda = \mathcal{N}(X\beta + \theta, \Sigma = (I_n - \rho C)^{-1} M) \tag{9}$$

where C is an adjacency matrix (we will assume[2] $C = W$) and M a diagonal matrix with road segments conditional variances describing the random effect related to segment i given all its neighbours random effects $\phi_i | (\phi_s, s \in \partial_i)$, with ∂_i denoting the indices of the road segments that are neighbour of i; usually $m_{ii} = \sigma^2 / n_i$ in order to let the variance of the ith segment decrease as its neighbour segments number n_i increases. Σ is a positive definite matrix when CM is symmetric, $m_{ii} > 0$, $i = 1, 2, \ldots, n$ and the spatial autocorrelation parameter ρ satisfies $-l_{min}^{-1} < \rho < l_{max}^{-1}$, with l_{min} and l_{max} the lowest and highest eigenvalues of matrix C [1, 9, 3, 4, 5].

[1] For this model and for the next two models we do not consider spatially structured random effect.

[2] See [4] for possible generalizations.

It is customary to use this specification with $\rho = 0.999$ in order to simulate the ICAR model, described in the following paragraph [10].

Intrinsic Conditional AutoRegressive spatial model By setting $\rho = 1$ in (9) it is possible to obtain the strongest relationship among road segments random effects, but matrix $(I_n - \rho W)$ becomes singular and the distribution function improper; in this setting the CAR model is named as intrinsic CAR (ICAR).

3 Monte Carlo Simulation

To assess the ability of models presented in Section 2 in taking into account different possible spatial structure features, we conducted a Monte Carlo simulation. Specifically, we generated geo-spatial data from each of previous model specifications (glm, SAR, SLX, SDM and ICAR) and assessed by means of model selection procedures which models (always in the same set: glm, SAR, SLX, SDM and ICAR) had the best performance in fitting data.

All models were estimated in a Bayesian setup, by having recourse to Integrated Nested Laplace Approximations (INLA) [7]; goodness of fit assessment was performed by using the deviance information criterion (DIC), the saturated DIC, the conditional predictive ordinate (CPO), as $-\sum_{i=1}^{n} \log(CPO_i)$ and marginal Likelihood (MML) as described in [7], mean squared error $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and the classical deviance, that does not include the DIC penalization term.

3.1 Simulation design

We considered lattices, defining 'junction graphs', where each segment is an arc corresponding to a street segment, and nodes are given by junctions or intersections between streets. To assess the influence of the density of segments in the lattice, we assumed three different sizes of the lattice, with 11, 17 and 21 segments on each of the four external sides of the greed, that correspond to 220, 544 and 840 total number n of road segments, respectively.

Fixed effects X_i for segment $i = 1, 2, \ldots, n$ were defined as follows. Its latitude and longitude covariates, with associated parameters β_i 0.04 and 0.08. Four metric covariates corresponding to conditions that increase the probability of crash occurrence, like average speed held by vehicles and level of traffic; they were generated, for simplicity, as i.i.d. uniform pseudo-random variables $U(0,1)$, with parameters β_i $0.06, 0.07, 0.08$ and 0.09. Four dummy variables, corresponding to factors that decrease the probability of crash occurrence, like the presence of traffic lights, speed bumps or tutors; these were generated as i.i.d. Bernoulli pseudo-random variables all with probability 0.55 and with parameters β_i $-0.05, -0.10, -0.15$ and -0.20. β_0 was set equal to -2.

With these parameter settings, the logarithm of feasible expected occurrence crash rate, also motivated by the application of the above mentioned models to real data, is

$$\log(E(\lambda_i)) = -2 + (0.06+0.09)\cdot 0.5 + (0.06+0.07+0.08+0.09)\cdot 0.5 + \\ -(0.05+0.10+0.15+0.20)\cdot 0.55 = -0.7,$$

and the expected rate results $\lambda_i = \exp(-0.7) = 0.4965$, with the following probabilities of crashes, according to a Poisson distribution

x_i	0	1	2	3	4	5
f_i	0.61	0.30	0.08	0.01	0.00	0.00

The unstructured random spatial component θ_i were generated as i.i.d. normal pseudo-random variables with zero mean and variance 0.01.

The structured spatial components ϕ with multivariate normal pseudo-random variables with zero mean and variance equal to 0.04.

Table 1 reports results of model assessment. Since DIC and saturated DIC values were equal in foremost cases, only DIC results appear in the table. Simulation were replicated $1,000$ times. Each column of the table corresponds to a simulation condition; the header reports models used to simulate data and a goodness of fit measures; parameter posterior distributions were estimated with all methods considered in Section 2; figures in each column of the table report the frequency distribution of best models in the 1000 trials.

Observe that the sample sizes (lattices with at most 840 road segments) considered in simulation are small with respect to cases encountered in real applications; simulation with larger sizes of lattice will be considered in future work.

It emerges that ICAR is always selected as best estimator, according to MML criterion. According to DIC and WAIC the classical glm estimator, possibly taking into account the presence of unstructured random effects, performs fine in most situations. The underlying model is characterized by the lowest effective number of parameters, and this fact has a clear impact on the penalization term in DIC. A more thoroughly analysis of performance indicator values for the models under competition shows that coefficient of variations computed at each simulation trial and for each criterion get values lower than 0.02 for DIC in all situations. Therefore, according to this preliminary results, we conclude that the models estimation methods, we have considered, appear to have very similar assessment selection criterion values. More simulation studies need to be performed, also in terms of much different effects of spatially lagged components. The aim is to establish if the use of more sophisticated setup than the classical glm is worth to fit real data, that are generally characterized by much larger lattice size than those considered in this preliminary study.

Table 1 Simulation results of model evaluation. 1000 replications on a lattice with 840 road segments.

data	glm						glm(with unstructured error)						SAR($\rho = 0.2$)						SAR($\rho = 0.5$)					
	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE
glm	445	0	774	419	0	0	112	0	780	127	0	0	106	0	779	98	0	0	71	0	912	52	0	0
glm u.e.	313	3	22	322	0	0	655	34	50	671	0	2	547	41	26	542	0	4	389	17	2	318	0	3
SAR	31	7	43	76	0	4	17	16	16	41	0	13	68	29	48	155	0	14	209	53	45	385	0	44
SLX	27	533	17	22	0	484	41	640	13	27	0	554	64	497	21	55	0	408	143	202	21	93	0	159
SDM	0	448	1	1	0	502	0	307	0	0	0	429	0	430	2	2	0	571	0	713	0	1	0	780
ICAR	184	9	143	160	1000	10	175	3	141	134	1000	2	215	3	124	148	1000	3	188	15	20	151	1000	14
max CV	0.01	0.06	0.51	0.01	0.29	0.18	0.01	0.08	0.79	0.01	0.28	0.24	0.01	0.09	0.83	0.01	0.31	0.24	0.01	0.09	1.44	0.01	0.39	0.22
data	SLX						SDM($\rho = 0.2$)						SDM($\rho = 0.5$)						ICAR					
segm	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE	DIC	Dev	CPO	WAIC	MLL	MSE
glm	89	0	771	87	0	0	84	0	776	68	0	0	41	0	937	23	0	0	72	0	703	86	0	0
glm u.e.	663	16	25	668	0	2	525	26	17	506	0	4	346	18	2	276	0	2	625	27	39	654	0	3
SAR	37	18	39	83	0	8	127	22	90	234	0	19	276	57	44	468	0	38	9	26	12	25	0	14
SLX	36	666	14	36	0	560	52	484	9	43	0	391	147	145	8	89	0	107	41	598	13	33	0	482
SDM	0	298	0	0	0	428	0	459	0	0	0	578	0	765	0	1	0	838	0	344	0	1	0	496
ICAR	175	2	151	126	1000	2	212	9	108	149	1000	8	190	15	9	143	1000	15	253	5	233	201	1000	5
max CV	0.01	0.08	0.79	0.01	0.28	0.23	0.01	0.08	0.96	0.01	0.31	0.22	0.02	0.10	1.41	0.01	0.39	0.25	0.01	0.09	0.76	0.01	0.30	0.27

References

1. Besag J.: Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society: Series B (Methodological). 36(2): 192-225 (1974)
2. Borgoni R., Gilardi A. and Zappa D.: Assessing the Risk of Car Crashes in Road Networks. Social Indicator Research. 156: 429-447 (2021)
3. Cressie N. and Wikle C.K.: Statistics for Spatio-Temporal Data. John Wiley and Sons Inc., New York (2011)
4. Donegan C.: Building spatial conditional autoregressive (CAR) models in the Stan programming language. OSF Preprints (2022)
5. Donegan C.: geostan: An R package for Bayesian spatial analysis. Journal of Statistical Software. 7(79): 4716 (2022)
6. Gilardi A., Mateu J., Borgoni R. and Lovelace R.: Multivariate Hierarchical Analysis of Car Crashes Data Considering a Spatial Network Lattice. Journal of the Royal Statistical Society Series A: Statistics in Society. 185(3): 1150-1177 (2022)
7. Gòmez-Rubio V.: Bayesian Inference with INLA. Chapman & Hall/CRC Press. Boca Raton, FL (2020)
8. Gòmez-Rubio V., Bivand R.S. and Rue H.: Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximation. Mathematics. 9(17) (2021)
9. Lee D.: CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. Journal of Statistical Software. 55(13): 1-24 (2013)
10. Wang C., Schifano E.D. and Yan J.: Geographical Ratings with Spatial Random Effects in a Two-Part Model. Variance. 13(1): 141-160 (2020)

Geostatistical modelling of livestock-related $PM_{2.5}$ pollution and scenario analysis for policymakers - Work in progress

Modelli geostatistici per modellare il $PM_{2.5}$ da allevamenti intensivi e analisi di scenario

Alessandro Fassò, Jacopo Rodeschini, Alessandro Fusta Moro and Francesco Finazzi

Abstract A relevant part of the secondary $PM_{2.5}$ derives from chemical reactions between ammonia and other air chemicals. In Lombardy, livestock and fertilisers are widely acknowledged to be responsible for approximately 97% of all ammonia emissions and the region is characterised by both high $PM_{2.5}$ concentrations and high levels of ammonia emission, the latter due to high livestock density. The main objective of our study is to highlight the relationship between ammonia emissions and $PM_{2.5}$ concentrations and to predict the reduction of $PM_{2.5}$ concentrations in response to the reduction of ammonia emissions, using scenario analysis. Scenario analysis will be carried out on high-resolution maps of the Lombardy region showing the different $PM_{2.5}$ reductions across the area, from 2016 to 2020.

Abstract *Una parte rilevante del $PM_{2.5}$ è di origine secondaria, dovuta da reazioni chimiche tra ammoniaca e altri composti atmosferici. In Lombardia, bestiame e fertilizzanti sono responsabili di circa il 97% di tutte le emissioni di ammoniaca. La regione è caratterizzata sia da alte concentrazioni di $PM_{2.5}$ sia da alti livelli di emissioni di ammoniaca, quest'ultime dovute dell'elevata densità di bestiame. L'obiettivo principale del nostro lavoro è evidenziare la relazione tra le emissioni di ammoniaca e fare analisi di scenario sulle concentrazioni di $PM_{2.5}$ ipotizzando delle riduzione di emissioni di ammoniaca. L'analisi di scenario sarà effettuata considerando il periodo dal 2016 al 2020 e le riduzioni delle concentrazioni del $PM_{2.5}$ saranno mostrate sulle mappe della regione.*

Key words: Geostatistical models, Ammonia scenario analyses, Lombardy air quality modelling.

Alessandro Fassò, Jacopo Rodeschini, Alessandro Fusta Moro and Francesco Finazzi
University of Bergamo, Via dei Caniana 2, 24127, Bergamo, Italy, e-mail: alessandro.fasso@unibg.it, jacopo.rodeschini@unibg.it, alessandro.fustamoro@gmail.com, francesco.finazzi@unibg.it

1 Introduction

In the Lombardy region, livestock and fertilisers are responsible for up to 97% of all ammonia emissions [1], which undergo chemical reactions in the atmosphere to form fine particulate matter, or $PM_{2.5}$. Due to its stability in the air and limited air circulation in the Po Valley, $PM_{2.5}$ often accumulates at high concentrations and adversely affects air quality. This occurs especially during the cold months when the low altitude of the mixing boundary layer combines with high levels of emissions from endothermic-powered vehicles and domestic heating systems. In the Po basin, the composition of atmospheric particulate matter in the winter has a very significant secondary component, normally exceeding 50%. Secondary particulate matter is formed in the atmosphere through complex chemical reactions. This percentage can be even higher during accumulation episodes in meteorological conditions unfavourable to dispersion. As a result of this high level of pollution, Lombardy has been found to have one of the highest $PM_{2.5}$-related mortality rates among all regions of Europe, with 100-150 premature deaths per 100.000 inhabitants, according to epidemiological studies [4]. In 2020, in Lombardy, the measured annual average concentrations in the different areas range between 9 - 28 $\mu g/m^3$.

In 2021, the World Health Organisation (WHO) published new guidelines aimed at reducing air pollution, based on new scientific evidence about the consequences of poor air quality levels on the health of the population. For the fine particulate matter, or $PM_{2.5}$, the goal is to have the annual mean below 35, 25, 15, and 10 $\mu g/m^3$ from the short to long term, respectively.

In light of the above discussion, the main objective of the AgrImOnIA project [5] (Agriculture Impact On Italian Air Quality) is to implement a scenario analysis assessing the reductions of $PM_{2.5}$ concentration due to ammonia reductions, using a geostatistical model. In the end, the reduction will be compared with respect to the WHO guidelines, and they will be useful to understand more about the role of the secondary PM. This paper develops preliminary data analysis and describes the method that will be used to achieve the project objectives.

In this context and in agreement with other environmental protection national laws, the Lombardy region has redacted a Regional Plan of Interventions for Air Quality (PRIA) in order to define several intervention policies able to reduce both pollutant and precursor emissions. The scenario analysis will take into account the Lombardy plan for the emission reductions, and in particular, assuming the full application of the PRIA commitments, the reduction in ammonia emissions in the year 2025 (compared to 2015) will be equal to 26%. Considering the constraints derived from data availability (2016-2020, inclusive) the idea is to train the model on the full period and then predict the $PM_{2.5}$ reductions achieved in the same period as a result of different paths where NH_3 emissions are gradually reduced, with full reduction (-26%) obtained at the end of 2020. The reduction will be achieved through the containment of ammonia emissions deriving from (*i*) livestock housing structures, (*ii*) storage of livestock manure (*iii*) distribution of livestock effluents and, (*iv*) promotion of conservation agriculture.

To achieve the AgrImOnIA project objectives, we will utilise geostatistics to take into account the spatiotemporal correlation. In particular, we will use the Hidden Dynamic Geostatistical Model (HDGM) [2]. In the end, regional maps of the reduction of $PM_{2.5}$ will be available in order to highlight the spatial dimension of the reductions over the Lombardy region, looking at yearly averages and at the number of days above the 25 $\mu g/m^3$ threshold (the annual average limit for the $PM_{2.5}$). In particular, we will consider maps with a 0.1×0.1 degree spatial resolution. These maps will be compared to the current literature, and in particular with the result provided by the Regional environment agencies.

2 Dataset

For this work, we use the AgrImOnIA dataset [5] (available from the Zenodo repository), a comprehensive daily dataset containing air pollutants concentrations [$\mu g/m^3$], livestock densities [$number/km^2$], weather conditions, emissions fluxes and land use characterisation. The AgrImOnIA dataset is related to the augmented Lombardy region which includes also a buffer of 0.3 degrees around the regional boundaries, see Figure 3a. The time period covered by the dataset spans from 2016 to 2021, inclusive. The localisations of the data are referred to the localisations of the stations belonging to the air quality monitoring network. In particular, air quality (AQ) data are sampled at 141 ground-level monitoring stations, irregularly located over the augmented Lombardy region. The dataset contains the daily concentrations of the following air pollutants: PM_{10} and $PM_{2.5}$, NO_x, NO_2, SO_2 and NH_3. Table 1 summarises the variables included in the AgrImOnIA dataset helpful to model the $PM_{2.5}$ in Lombardy. After removing some $PM_{2.5}$ stations due to lack of data, the resulting network includes 49 $PM_{2.5}$ stations. The ammonia emissions in the dataset are provided by the Copernicus Atmosphere Monitoring Service (CAMS) service covering the years 2016-2020. In particular, from the global emission inventories, we consider the sum of ammonia emissions from (*i*) agricultural waste burning, (*ii*) livestock manure management and (*iii*) agriculture soil. For scenario analysis purposes, we built a complementary dataset (published on Zenodo), with the same variables, on a regular grid with a spatial resolution of 0.1×0.1 degrees, covering the Lombardy region and its surrounding area. In this way, it will be possible to create maps of the Lombardy region.

3 Preliminary Analysis

The use of spatiotemporal models is justified by the spatiotemporal variogram [3] shown in Figure 1 highlighting the spatiotemporal correlation structure among $PM_{2.5}$ concentrations. Indeed, moving away in space and time, the observed variance increases according to the distance. The spatial correlation seems to stabilise

Table 1: Variables retrieved from AgrImOnIA dataset [5] useful for modelling air quality in the Lombardy region.

Dimension	**Variable**
Air quality (AQ)	$PM_{2.5}$(49 stations)
Weather (WE)	Temperature, Precipitation, Wind speed, Relative humidity, Boundary layer height, Surface pressure
Emission (EM)	NH_3 from livestock manure management, NH_3 from agriculture soil, NH_3 from agriculture waste burning, Total NO_x, Total SO_2
Livestock (LI)	Density of pigs, Density of bovines
Land (LA)	Land use, High vegetation index, Low vegetation index

at $\sim$ 90[km] while the temporal correlation at around 7 days. Apart from the region near the origin, the variogram lines are roughly parallel which is consistent with a separable auto-covariance structure. The temporal correlation is further investigated through the Partial Auto-Correlation Function (PACF) calculated for each station for the first 60 lags (days) and summarised through box plots displayed in Figure 2. The most significant auto-correlated component is within the first lag (1 day). As a result, it is appropriate to use a model with first-order Markovian dynamics. Additionally, the PACF shows that $PM_{2.5}$ is relatively stable in the atmosphere and shares a similar behaviour around augmented Lombardy. It is worth noting that, in the presence of high levels of correlation both in space and time, a geostatistical model with random effects (as the HDGM) can explain most of the variability even without any regressor. This fact could be problematic for modelling, especially in the scenario analysis framework where the coefficients determine the reduction of the response variable. Moreover, the spatiotemporal correlation modelled in the random effect makes it challenging to identify which regressors explain a large amount of the variability of the response variable.

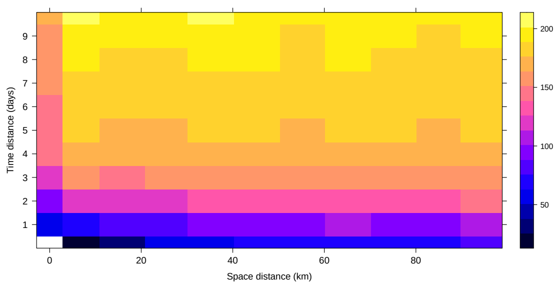


Fig. 1: Spatiotemporal variogram [3] computed on $PM_{2.5}$ daily observations at 49 stations from 2016 to 2021.

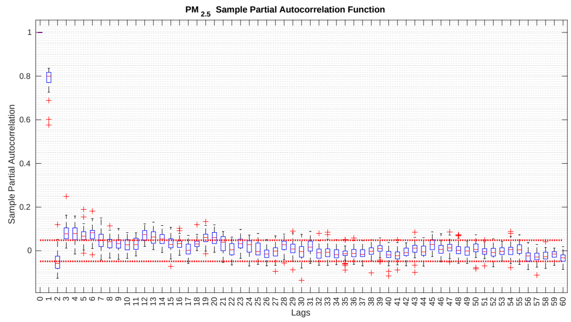


Fig. 2: Partial Auto-Correlation Function (PACF) calculated by stations for the first 60 lags (days) and summarised through box plots. The i-th box plot summarises the partial autocorrelation calculated for each station at i-th lag.

The NH_3 emissions in the limited region considered by our study are shown in Figure 3a. As expected, NH_3 emissions are linearly correlated with the livestock densities shown in Figure 3b. However, the air quality monitoring network is not specifically designed to monitor emissions from livestock and therefore many of the monitoring stations are located in areas with low emission levels, and no one is located where NH_3 emissions reach the peak, as shown in Figure 3a. So, it will be difficult to estimate the livestock impact on $PM_{2.5}$ concentration using NH_3 emissions. However, areas presenting high levels of ammonia emission are characterised by high concentrations of $PM_{2.5}$ (as Cremona and Brescia provinces), highlighting the need for knowledge about the impact of the agricultural sector on air quality.

4 Methods

To better understand the relationship between predictor variables and particulate matter, taking into account the spatial and temporal correlation, we will use the Hidden Dynamic Geostatistical Model (HDGM). The HDGM is a hierarchical geostatistical model [3, 2] defined as:

$$\begin{cases} y(s,t) = \boldsymbol{X}_b(s,t)\boldsymbol{\beta} + z(s,t) + \varepsilon(s,t) \\ z(s,t) = gz(s,t-1) + \eta(s,t) \end{cases} \tag{1}$$

where $y(s,t) = log(PM_{2.5})$; the fixed effects term $\boldsymbol{X}_b(s,t)\boldsymbol{\beta}$ accounts for all exogenous, regressive effects; the random effects term $z(s,t)$ covers all spatial and temporal dependence; the measurement error $\varepsilon(s,t) \sim N(0,\sigma_\varepsilon)$ is independent in space and time while the innovation term $\eta(s,t)$ is a zero-mean Gaussian process $GP(0,v\rho(\|s-s'\|,\theta))$ where v is a scale variance parameter and ρ is the exponential correlation function. In the end, the parameters will be estimated are

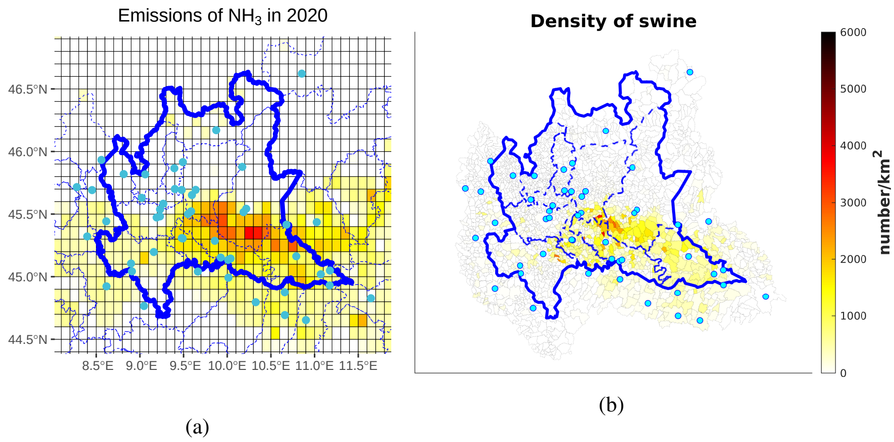


Fig. 3: Emission density (t/km^2) (3a) and swine density (3b) over the augmented Lombardy region. The AQ stations (cyan crosses) are spread randomly over the studied area. The continuous blue line represents the boundary of Lombardy while the dotted blue line represents the provinces.

$\boldsymbol{\Phi} = \{\boldsymbol{\beta}, g, \theta, v, \sigma_\varepsilon\}$. The parameters are estimated by the maximum likelihood (MLE) method using the EM algorithm.

Acknowledgements This research was funded by Fondazione Cariplo under the grant 2020-4066 "AgrImOnIA: the impact of agriculture on air quality and the COVID-19 pandemic" from the "Data Science for science and society" program.

References

1. ARPA Lombardia: INEMAR - Inventario Emissioni in Atmosfera: emissioni in Regione Lombardia nell'anno 2019 - versione in revisione pubblica (2022) https://www.inemar.eu/xwiki/bin/view/InemarDatiWeb/Aggiornamenti+dell%27inventario+2019
2. Calculli, C., Fassò, A., Finazzi, F. et al.: Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. Environmetrics. 26, 406–417 (2015)
3. Cressie, N., and Wikle, C. K.: Statistics for spatio-temporal data. John Wiley & Sons (2015)
4. European Environment Agency: Premature deaths due to exposure to fine particulate matter in Europe - 8th EAP (2022) https://www.eea.europa.eu/ims/health-impacts-of-exposure-to-. Cited 25 Apr 2023
5. Fassò, A., Rodeschini, J., Moro, A.F. et al.: Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. Sci Data (2023) doi: 10.1038/s41597-023-02034-0

Functional clustering methods for space-time big data from mobile phone networks

Metodi di clustering funzionali per big data spazio-temporali dalla rete di telefonia mobile

Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

Abstract Understanding people's movements in urban areas is critical for efficient urban planning and infrastructure management in smart cities. In this work, we analyze the crowding dynamics in six urban areas of the municipality of Brescia characterized by different land use using recent high technology data. Recurring behaviors in the days of a year are captured by means of functional data clustering. Two clustering methodologies are applied: the k-means Alignment, and the Model-Based Functional Data Clustering. The results of the two procedures are compared using the Rand index and by analyzing the clusters' similarities over a set of passive descriptive variables.

Abstract *Comprendere gli spostamenti delle persone nelle aree urbane è fondamentale per un'efficiente pianificazione urbana e gestione delle infrastrutture nelle smart cities. In questo lavoro si analizzano le dinamiche di affollamento in sei aree urbane del comune di Brescia caratterizzate da diverso uso del suolo utilizzando dati recenti ad alta tecnologia. I comportamenti ricorrenti nei giorni di un anno vengono catturati mediante clustering di dati funzionali. Vengono applicate due metodologie di clustering: il k-means Alignment, ed il Model-Based Functional Data Clustering. I risultati delle due procedure vengono confrontati utilizzando l'indice di Rand e analizzando le somiglianze dei cluster su un insieme di variabili descrittive passive.*

Selene Perazzini
University of Brescia, DMS Statlab, Department of Economics and Management, Contrada Santa Chiara, 50, Brescia, e-mail: selene.perazzini@unibs.it

Rodolfo Metulini
University of Bergamo, Department of Economics, Via Caniana, 2, Bergamo, e-mail: rodolfo.metulini@unibg.it

Maurizio Carpita
University of Brescia, DMS Statlab, Department of Economics and Management, Contrada Santa Chiara, 50, Brescia, e-mail: maurizio.carpita@unibs.it

Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

Key words: Functional data clustering, k-means alignment, model-based clustering, mobile phone density.

1 Introduction

Capturing recurring behaviors in the crowding dynamics in urban areas is critical for various local management purposes, especially in the context of smart cities. Nowadays, understanding people's dynamics is considered essential for the achievement of some of the Sustainable Development Goals set by the United Nations (e.g., goals 9 - Industry, innovation, and infrastructure, and 11 - Sustainable cities and communities). In this respect, mobile phone data are currently among the richest sources of information as they allow for a detailed representation of people at the small area level. For this reason, they have been extensively used for the analysis of crowding (e.g., [1, 2, 3, 4]). In this work, we use Mobile Phone Density (MPD) data to analyze people's presence in six areas of the Municipality of Brescia characterized by different land use (residential, productive, etc ...). In order to capture similarities in the crowding dynamics in the days of the year, clustering analysis can be used. Here, we focus on functional data clustering, as crowding during the day can be interpreted as a discrete realization of a continuous stochastic process. Some applications of functional cluster analyses in the context of traffic and crowdings in smart cities can be found in the literature (e.g., [5]). We apply two methodologies - a distance-based and a filtering method - to cluster the 365 days from May 2022 to April 2023 into 4 or 5 groups. The results of the two methodologies are compared using the Rand index. Moreover, we investigate similarities in the obtained clusters with respect to the predominant seasons and weekdays that constitute them.

The paper organizes as follows: Section 2 describes the data; Section 3 presents the adopted clustering methodologies; Section 4 shows results; Section 5 concludes.

2 Data

In this work, we use MPD data to capture the differences in the daily dynamics of traffic in the Municipality of Brescia. The MPD database at our disposal refers to the "Aree di CEnsimento" (ACEs) of the Municipality of Brescia observed from May 1st, 2022 to April 30th, 2023, and reports the average number of mobile phone SIM cards in an ACE during a 15-minute interval. The database has been released by Olivetti S.p.A. (www.olivetti.com) and has been kindly provided to us by the Municipality of Brescia.

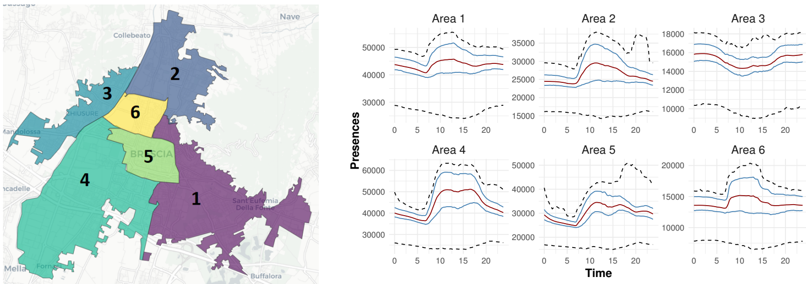


Fig. 1 Left: map of the six analyzed areas. Right: Summary of the set of curves of the six areas. The lines represent the mean (red), the first and third quartiles (blue), and the minimum and maximum (black dotted) presences observed. Note that axes values are allowed to vary among the areas.

We restrict our attention to the 14 ACEs constituting the city's main urban area and group them into 6 areas on the basis of spatial proximity and land use[1] as shown in the left map of Figure 1. Specifically, areas 1-3 are mainly occupied by sparse or discontinuous residential urban fabric; area 4 is predominantly dedicated to industrial, craft, commercial, and agricultural settlements; area 5 is a residential area with dense urban fabric; area 6 is equally divided between dense residential fabric and industrial, craft, and commercial settlements.

The 6 areas are analyzed separately. For each of them, MPD data have been collected in a 365×96 matrix where each row corresponds to one day, and each column to a 15-minute interval. Each row in the matrices can be viewed as a curve observed at 96 points in time[2]. As shown in the right plots of Figure 1, the observed curves describe substantially different patterns in the six areas.

3 Functional clustering methodology

Different clustering methods for functional data can be found in the literature. In this work, we focus on methodologies that estimate proper functional curves. For this reason, we neglect the "raw data" methods. The other clustering algorithms can be divided into two categories [6]: the "distance-based" methods, which are based on specific traditional distances adapted to functional data; the "filtering" methods, which approximate the curves into some basis functions and perform clustering using the basis expansion coefficients.

To analyze the daily traffic dynamics, we use two clustering algorithms (one per category) with the aim to compare the obtained clusters and check whether the

[1] For data on the land use, we referred to the dataset "Uso e copertura del suolo della regione Lombardia 2021" (DUSAF 7.0) released by the Lombardy Region and available at https://www.geoportale.regione.lombardia.it

[2] Data for the four intervals corresponding to March 26th, 2023 between 2 and 3 AM are missing and have been computed by linear interpolation.

methods lead to similar results. Specifically, we chose the distance-based method "k-means Alignment" (k-mA) by [7], and the filtering method "Model-Based Functional Data Clustering" (M-BC) by [8]. The analyses have been performed using, respectively, the `R` packages `fdacluster` and `funFEM`. A brief description of the two methods is provided in the following.

k-mA. The k-mA procedure decouples both the phase and amplitude variability and the between-cluster amplitude variability. Curves are clustered on the basis of a dissimilarity index measuring the dissimilarity between the functional data and a class of warping functions defining the transformation of the abscissa (i.e., the 96 15-minute intervals t). In this work, we apply the dissimilarity index $|\frac{f_1}{|f_1|} - \frac{f_2}{|f_2|}|$ where $|\cdot|$ indicates the norm and f_1 and f_2 are two functions, and the warping function $h(t) = mt + q$ with $m \in \mathbb{R}^+$ and $q \in \mathbb{R}$.

M-BC. The M-BC assumes that the set of observed curves $\{x_1, \ldots x_n\}$ is generated by an unknown stochastic process $X(t) = \sum_{r=1}^{p} \gamma_r(X) \upsilon_r(t)$ defined over a random vector $\{\gamma_1(X), \ldots, \gamma_p(X)\}$ and a set of basis functions $\{\upsilon_1, \ldots, \upsilon_p\}$ with p assumed known and fixed. The algorithm then clusters the $\{x_1, \ldots x_n\}$ curves into K homogenous groups based on a Discriminative Functional Mixture model by applying the Fisher-EM algorithm [9]. Since we have periodic data, we used Fourier basis functions. We chose $p = 9$ on the basis of the sum of the Root Mean Squared Errors computed between the process' realization x_i and the smoothed estimated curve evaluated at discrete points of time.

4 Results

To begin with, the number of clusters K has to be chosen. In order to avoid too small groups, a constraint of at least 20 days per cluster has been introduced. This choice reflects the number of Saturdays and Sundays in summer, which, according to [5], typically show different patterns in people's crowdings. Then, the optimal number of clusters has been evaluated for each methodology. For the k-mA method, K_{k-mA} is typically chosen as the lowest number of clusters associated with a considerable reduction of the total dissimilarity index with respect to a clustering solution with one less cluster. In the M-BC, K_{M-BC} is chosen on the basis of the AIC or BIC. We observed that, for most areas, $K_{M-BC} > K_{k-mA}$. However, AIC or BIC model selection criteria tend to overestimate the optimal number of clusters when the functional data are non-Normal [10]. To compensate for this effect and ease the comparison of results, we set the same number of clusters $K = min(K_{M-BC}, K_{k-mA})$ for both the k-mA and the M-BC analyses. Instead, we let K free to vary among the six areas to better capture their specific characteristics. As a result, we obtain 4 clusters for areas 1 and 2, and 5 for the others.

The Rand Index (RI) [11] has been computed. The index is an evaluation metric measuring the similarity between the cluster assignments by making pair-wise com-

Table 1 Number of units per cluster and Rand Index (RI). The numbers of clusters' units are reported per area and clustering methodology. Note that only 4 clusters have been constructed for Areas 1 and 2.

	Area 1		Area 2		Area 3		Area 4		Area 5		Area 6	
Cluster	m-kA	M-BC	m-kA	M-BC	m-kA	M-BC	m-kA	M-BC	m-kA	M-BC	m-kA	M-BC
1	115	165	166	164	124	151	108	174	105	132	91	145
2	114	83	92	90	97	74	77	63	93	72	82	68
3	78	79	61	73	61	62	64	44	84	68	75	63
4	58	38	46	38	49	58	61	44	51	54	65	60
5					34	20	55	40	32	39	52	29
RI	0.80		0.74		0.78		0.84		0.76		0.82	

parisons. The index is an extrinsic metric taking values from 0 to 1 and we use it to evaluate the similarity of the clusters constructed with the k-mA and the M-BC. The RI is shown in Table 1, along with the number of units per each cluster obtained with each method in the six areas. As one can notice, the index takes values between 0.74 and 0.84 and therefore indicates a high level of agreement between the two sets of clusters.

The composition in terms of weekdays and seasons of the obtained clusters has been analyzed. Overall we find reasonable agreement between the clusters constructed with the two methods. This is particularly the case when comparing the weekdays composition of the clusters obtained using the k-mA and the MB-C (see top charts of Figure 2). Some similarities also emerge when considering the cluster's composition in terms of seasons (bottom charts of Figure 2). It is worth noticing that the MB-C always tends to identify a cluster grouping almost only summer days. We do not observe the same result in the k-mA clusters.

5 Conclusion

Two functional data clustering methodologies have been used to capture recurring behaviors in the crowding dynamics data from mobile phone networks of six small urban areas of the municipality of Brescia: the k-means Alignment, and the Model-Based Functional Data Clustering. The results of the two procedures have been compared using the Rand index and by analyzing the clusters' similarities on a set of passive descriptive variables.

Acknowledgements This study has been developed for the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project "Sustainable Mobility Center (MOST)" 2022-2026, CUP D83C22000690001, Spoke N° 7 "CCAM, Connected networks and Smart Infrastructures".

Fig. 2 Composition of clusters obtained with k-mA and M-BC per area. Top charts (rows 1-2): weekdays. Bottom (rows 3-4): seasons.

References

1. Secchi, P., Vantini, S. and Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Statistical Methods & Applications. 24(2), 279-300 (2015)
2. Carpita, M., Manisera, M., and Zuccolotto, P.: Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia. In: Lombardo R., Camminatiello I., Simonacci V. eds. IES 2022: Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, Book of Short Papers of the 10th Scientific Conference of the SVQS, 575-581. PKE Press, Milano (2022)
3. Metulini, R., and Carpita, M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research. open access online first, 1-26. (2023) https://link.springer.com/article/10.1007/s10479-023-05195-8.
4. Perazzini, S., Metulini, R., Carpita, M. Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Submitted to journal.
5. Metulini, R., and Carpita, M.: A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Social Indicators Research. 156(2), 761-781 (2021)
6. Jacques, J. and Preda, C.: Functional Data Clustering: A Survey. Advances in Data Analysis and Classification. 8, 231-255 (2013)

7. Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V.: K-mean alignment for curve clustering. Computational Statistics & Data Analysis. 54(5), 1219-1233 (2010)
8. Bouveyron, C., Côme, E. and Jacques, J.: The discriminative functional mixture model for a comparative analysis of bike sharing systems. The Annals of Applied Statistics. 9(4), 1726-1760 (2015)
9. Bouveyron, C., and Brunet, C.: Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. Statistics and Computing, 22(1), 301–324 (2012)
10. Bouveyron C., Celeux G., Murphy T.B., Raftery A.E.: Model-Based Clustering and Classification for Data Science: With Applications in R, Cambridge University Press, UK (2019)
11. Rand W.M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66(336), 846-850 (1971)

Solicited Session SS22 - *Methodological developments and applications for the assessment of student competencies*

Organizer: Stefania Mignani
Chair: Fulvia Pennoni

1. *Modeling the main drivers of mathematical literacy of school-leaving students. Some evidence from the Invalsi tests* (Davino C., Palumbo F., Romano R. and Vistocco D.)
2. *Educational Data Mining: clustering students' performance over time* (Taraborrelli G. and Farnè M.)
3. *The nexus of cultural capital with participation in early childhood education* (Ripamonti E.)
4. *High- and Low-Performing students and future career: a gender and social issue* (Falzetti P. and Ricci R.)

Modeling the main drivers of mathematical literacy of school-leaving students. Some evidence from the Invalsi tests

Modelli statistici per l'analisi dei drivers dell'alfabetizzazione matematica degli studenti in uscita dalla scuola secondaria. Alcune considerazioni a partire dai test Invalsi

Cristina Davino, Francesco Palumbo, Rosaria Romano and Domenico Vistocco

Abstract The early diagnosis of gaps in students preparing to leave secondary school is important for their entry into the labour market or for their academic experience. The present paper iinvestigates a specific aspect of the skills of school-leavers, namely mathematical literacy. Quantile regression is proposed to explore the impact of students' characteristics and social context on mathematical literacy, taking into account the ubiquitous heterogeneity of the students' population. The proposed study grounds on Invalsi test results.

Abstract *La diagnosi preventiva delle lacunw degli studenti che si apprestano a lasciare la scuola secondaria è molto importante per l'accesso al mondo del lavoro o per la prosecuzione degli studi all'Università. Il presente lavoro si basa sullo studio di un aspetto specifico delle competenze degli studenti che hanno terminato la scuola, ovvero l'alfabetizzazione matematica. In particolare, lo studio parte dai risultati dei test Invalsi e propone l'uso della regressione quantile per valutare l'impatto che le caratteristiche degli studenti ed il contesto sociale hanno sulle competenze matematiche, tenendo conto dell'inevitabile eterogeneità della popolazione studentesca.*

Key words: Mathematical literacy, Quantile Regression, Heterogeneity, Invalsi tests

---

Cristina Davino
University of Naples Federico II, Dept. of Economics and Statistics, e-mail: cristina.davino@unina.it

Francesco Palumbo
University of Naples Federico II, Dept. of Political Science, e-mail: fpalumbo@unina.it

Rosaria Romano
University of Naples Federico II, Dept. of Economics and Statistics, e-mail: rosaroma@unina.it

Domenico Vitocco
University of Naples Federico II, Dept. of Political Science, e-mail: domenico.vistocco@unina.it

1 Introduction and reference framework

Mathematical literacy has received increasing attention in many countries over the last few years because of its importance in the social and working life of every individual. According to the OECD definition [11], “mathematical literacy is an individual’s capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual’s life as a constructive, concerned and reflective citizen”.

It is widely acknowledged that people can only tackle many of the challenges of modern life effectively only if they are mathematically literate in key areas (planning in personal finance, assessment of risk, and critical appraisal of the flood of statistical information from advertising, politicians and the press) [12]. Moreover, too many students leave school unable to function mathematically at the level needed in the modern labour market or to enter university with an acceptable background.

In educational research, exploring if and how individual characteristics and contextual factors relate to learning outcomes is considered of great interest in order to deal with inequality issues [3]. It is a matter of fact that the results achieved by each student are affected by different components: the outcomes of the learning-teaching process (i.e. students’performance), individual characteristics of the student (gender, field of study attended, regularity in studies, economic, social and cultural context of the family of origin, etc.) and the environment in which she/he lives (geographical area of residence, the economic-social-cultural context of the school, etc.) [1, 13, 2, 10].

In this framework, the aim of this paper is exploring the impact of student characteristics and social context on mathematical literacy highlighting possible heterogeneity related both to the territory and toto the living context. The goal is achieved exploiting potentialities of Quantile Regression [8, 4] and a clustering and modeling procedure proposed by Davino and Vistocco [5]. The proposed study is based on Invalsi test results [7].

2 Data, methods and main results

The proposed analysis deals with the wave 2021/2022 of the INVALSI tests to evaluate math ability at the end of upper secondary school. The INVALSI tests are national standardized tests aimed to collect data useful to calculate the added value of the education offered from the single schools and to identify excellencies or weaknesses of the Italian educational system. The mathematics test aims to detect the mastery of the discipline intended as an instrument of analysis and reflection on reality and not as a mechanical application of formulas and rules. The test is based on international comparative surveys (such as the PISA assessment by OECD), so it does not aim to detect specific knowledge acquired at school but to detect how students manage to use what they have learned at school to solve real problems.

The sample is composed by more than ten thousands of students and considers several factors to explain mathematical literacy measured throuh the ability math score (from now, *wle_math_score*), school (scientific lyceum, technical institute, other lyceum, professional institute), gender, age, place of birth (foreign, Italy), regularity of school career (latecomer, regular), geographical area (north, centre, south-islands), ESCS index (Economic, Social and Cultural Status).

Quantile Regression (QR) [8, 4] is proposed to evaluate if and how much the effect of these factors changes according to the different levels of the math score. Moreover, this paper exploits quantile regression to identify groups of units with a different dependence structure thus identifying the possible unobserved heterogeneity in the sample [5]. The best model for each group will be estimated and inferential procedures will allow us to test if group structures are statistically different.

QR gained popularity in applied economics by the end of the 90's, when people realized the importance of heterogeneity. A QR model for a given conditional quantile θ can be formulated as follows:

$$Q_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \mathbf{X}\hat{\beta}(\theta)$$

where $\mathbf{X}$ denotes the regressor matrix, $\mathbf{y}$ is the dependent variable, $0 < \theta < 1$ a generic quantile, $Q_\theta(.|.)$ the conditional quantile function and ε is the error term such that $Q_\theta(\varepsilon|\mathbf{X}) = 0$.

The clustering and modeling approach proposed by Davino and Vistocco [5] allows to identify a typology in a dependence model, namely to cluster units characterized by similar dependence structures. Moreover, the approach detects the best model for each group, and inferential procedures are available for testing differences among groups. The clustering and modeling approach is structured in the five steps detailed below pplied on the INVALSI data.

1. Identification of the global dependence structure

In the first step, a QR model is estimated on the whole sample, where θ is the quantile of interest and it ranges from 0 to 1. Although theoretically it is possible to extract infinite quantiles, in practice, a fairly accurate approximation of the whole quantile process can be obtained using a dense grid of equally spaced quantiles in the unit interval (0, 1). In particular, for the empirical analysis on the INVALSI dataset, we considered the quantiles varying from 0.01 to 0.99 with a step of 0.01. For each regressor, k coefficients are estimated, where the value of k is the number of estimated conditional quantiles.

2. Identification of the best model for each unit

In the second step, the best model for each unit i is identified detecting the quantile model able to better estimate the response variable, namely selecting the model which minimizes the absolute difference between the observed and the estimated values:

$$\theta_i^{best} :_{\theta=1,\dots,k} \arg\min\, |y_i - \hat{y}_i(\theta)|\,.$$

We denote with θ_i^{best} the quantile associated to the best model for the i-th unit.

3. Clustering units

The third step of the proposed strategy aims to group units, namely to identify a partition on the basis of the QR results. Units are grouped according to the best quantile they have been assigned in step 2 because it can be considered as an indicator of a similar dependence structure. Different criteria can be followed to identify the groups. The simplest solution consist of dividing the θ^{best} vector into classes with equal frequency or equal width. Once the groups are identified, a reference quantile for each group is computed considering the average of the best quantiles assigned to the units belonging to the group. For our case study, we partitioned θ^{best} into 10 (D) equally spaced parts, i.e. considering the 10 deciles. For each of the D groups, a reference quantile is computed. Such group reference quantile is exploited in the following step to estimate the models representing each group. Table 1 shows the approach for the INVALSI data: the first column refers to the 10 considered cut points for partitioning the k quantiles, the reference quantiles are on the second column. Here we considered the average of the θ^{best} belonging to the each group.

In order to find the best partition, the joint test of equality of slopes [9] is used to compare the D models (third column of Table 1): if the slopes of two contiguous models are not significantly different, that means the two groups share a similar dependence structure. This procedure allows us to identify the best partition of the units into 5 groups, shown in Table 1 using the horizontal lines grouping the rows. The cardinality of the five detected groups is reported in fifth column. Each group is characterised by a reference quantile (${}_g\theta^{best}$) obtained synthesizing the θ^{best} assigned to the units belonging to each group, and shown in last column of Table 1. In particular, we computed such quantiles as the average of the θ_{best} of each group. The comparison among the ${}_g\theta^{best}$ values provides hints about the presence of group differences and peculiarities.

Table 1 Identification of the best partition

quantile	${}_d\overline{\theta}^{best}$	p-value	group	n_g	${}_g\theta^{best}$
0.1	0.053	0.008	1	898	0.053
0.2	0.159	0.092	2	3548	0.305
0.3	0.264	0.102			
0.4	0.371	0.151			
0.5	0.470	0.006			
0.6	0.570	0.002	3	876	0.554
0.7	0.670	0.193			
0.8	0.770	0.000	4	1882	0.705
0.9	0.864	0.127	5	1946	0.903

4. Modeling groups

In the fourth step, QR is again executed on the total sample but considering only the 5 reference quantiles previously defined (last column in Table 1). Differences in the explaining capability of the regressors according to the group membership can be easily identified through the inspection of a single coefficient matrix (Table 2). Comparing OLS results (first column in Table 2) with QR results it is possible to appreciate the role played by each regressor for different levels of the math score. For most regressors, the sign of the coefficients does not change moving from low to high performing students and a trend in the size effect is evident.

Table 2 OLS and QR coefficients with group effects

Variable	OLS	G1 $\theta = 0.053$	G2 $\theta = 0.305$	G3 $\theta = 0.554$	G4 $\theta = 0.705$	G5 $\theta = 0.903$
(Intercept)	**213.67**	**145.82**	**205.22**	**222.53**	**216.38**	**248.64**
technical institute	**-30.08**	**-26.95**	**-28.45**	**-28.20**	**-31.85**	**-37.48**
other lyceum	**-32.99**	**-29.94**	**-31.23**	**-32.22**	**-34.11**	**-38.24**
professional institute	**-54.03**	**-49.31**	**-48.57**	**-52.42**	**-54.35**	**-64.48**
male	**11.03**	**6.10**	**8.73**	**11.98**	**14.36**	**16.57**
age	0.22	1.59	-0.13	-0.07	0.89	0.35
birth_Italy	3.31	3.33	2.38	-0.18	0.11	6.31
regular career	**12.45**	8.89	**12.30**	**14.12**	**15.67**	**12.45**
foreigner	-1.79	-2.76	**-4.08**	**-3.57**	-1.76	1.48
centre	**-14.54**	**-13.83**	**-15.01**	**-14.64**	**-15.34**	**-11.30**
south-islands	**-30.91**	**-26.91**	**-30.92**	**-31.52**	**-32.37**	**-31.12**
escs	**3.16**	1.31	**2.45**	**3.04**	**3.83**	**4.35**

5. Testing differences among groups
In the final step, the evaluation of the statistical significance of the differences among the coefficients related to each group can be carried out exploiting the classical inferential tools in the quantile regression framework. It is important to highlight that group coefficients can be compared because they have been estimated on the whole sample. Exploiting again the joint test of equality of slops proposed by Koenker and Machado [9], it results that groups are statistically different while comparing sequentially differences between coefficients obtained in each pair of models ((p-values are reported in Table 3), the biggest differences emerge between the groups associated with high math scores.

Table 3 p-values from separate testing on each slope coefficient

	G1 vs G2	G2 vs G3	G3 vs G4	G4 vs G5
technical institute	0.360	0.826	0.001	0.001
other lyceum	0.397	0.317	0.061	0.017
professional institute	0.733	0.007	0.136	0.000
male	0.044	0.000	0.002	0.075
age	0.316	0.958	0.377	0.749
birth_Italy	0.849	0.364	0.890	0.040
regular career	0.219	0.369	0.395	0.233
foreigner	0.604	0.764	0.178	0.063
centre	0.444	0.724	0.439	0.003
south-islands	0.002	0.497	0.303	0.358
escs	0.074	0.169	0.048	0.403

3 Concluding remarks

The proposed approach is mainly based on the explanatory power of QR to explore the entire conditional distribution of the response variable and it aims to discover groups in a dependence model and to identify the best model for each group. The approach can represent a valid tool to cluster units according to the dependence structure without a priori information but only using the observed similarities among units in terms of conditional quantile estimates. The final results are easy to interpret as the coefficients associated to each group follow the same interpretation of any linear model; furthermore, the best quantiles assigned to detected groups summarizes the distribution of the response variables through a set of locations characterizing different dependent models which appear in the data. Classical inferential procedures can be used to compare the models because the group effects are identified using the whole sample.

References

1. Baye, A., Monseur, C.: Gender differences in variability and extreme scores in an international context. Large-scale Assessments in Education, 4(1), 1-16 (2016)
2. Cascella, C., Giberti, C., Bolondi, G.: An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. Studies in Educational Evaluation, 64, 100819 (2020)
3. Costanzo, A., Desimoni, M.: Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data. Large-scale Assessments in Education 5(1), 1–25 (2017)
4. Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Applications. Wiley Series in Probability and Statistics (2013)
5. Davino, C., Vistocco, D.: Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on University outcome. Statistics & Its Interface 11, 541-556 (2018)
6. Furno, M., Vistocco, D.: Quantile Regression: Estimation and Simulation, vol 216, John Wiley & Sons (2018)
7. INVALSI: I Dati INVALSI per indagare e migliorare l'insegnamento della matematica (2021)
8. Koenker, R.W., Basset, G.: Regression quantiles. Econometrica, 46(1) (1978)
9. Koenker, R.,Machado, J.: Goodness of Fit and related inference processes for quantile regression. J. Am. Stat. Assoc. 94, 1296-1310 (1999)
10. Matteucci, M., Mignani, S.: Investigating gender differences in mathematics by performance levels in the Italian school system. Studies in Educational Evaluation, 70, 101022 (2021).
11. OECD: The PISA 2003 Assessment Framework : Mathematics, Reading, Science and Problem Solving Knowledge and Skills, PISA, Éditions OCDE, Paris (2004)
12. Steen, L. A., Turner, R., Burkhardt, H.: Developing mathematical literacy. *Modelling and applications in mathematics education*, pp. 285-294, Springer, Boston, MA. (2007)
13. Yang H.K., Gustafsson, J. E.: Determinants of country differences in effects of parental education on children's academic achievement. Large-scale Assessments in Education, 4(1), 1-13 (2016)

Educational Data Mining: clustering students' performance over time

Educational Data Mining: clusterizzazione delle performance degli studenti nel tempo

Gioia Taraborrelli and Matteo Farnè

Abstract This paper provides an application of four cluster analysis methods, i.e., hierarchical cluster analysis, k-means algorithm, factorial k-means (FKM) and reduced k-means (RKM), to explore educational data in the context of Educational Data Mining and Learning Analytics. A pilot analysis was carried out on a dataset regarding the performance of a class of high school students in three periods (which were treated as three separate datasets). The goal is to show how the composition and the number of groups vary in each period and which are the factors that influence the creation of groups. The partitions were compared in terms of reliability using the average silhouette width index (ASW). RKM and k-means generated the same results, which are the most suitable considering the ASW.
Abstract *Questo paper mostra un'applicazione di quattro metodi di cluster analysis, quali clustering gerarchico, l'algoritmo k-means, factorial k-means (FKM) e reduced k-means (RKM), utilizzati per esplorare i dati sull'apprendimento, nel contesto dell'Educational Data Mining e del Learning Analytics. È stata condotta un'analisi pilota su un dataset contenente le performance di una classe di studenti delle scuole superiori in tre periodi (trattati come tre dataset separati). L'obiettivo è mostrare come la composizione e il numero di gruppi variano in ogni periodo e quali sono i fattori che influenzano la creazione dei gruppi. Le partizioni sono state confrontate in termini di affidabilità utilizzando l'indice average silhouette width (ASW). RKM e k-means hanno generato la stessa partizione, la più credibile considerando l'ASW.*

Key words: Educational Data Mining, Learning Analytics, Longitudinal data, High dimension.

Gioia Taraborrelli
Department of Statistical Sciences, University of Bologna; e-mail: gioia.taraborrelli@studio.unibo.it

Matteo Farnè
Department of Statistical Sciences, University of Bologna; e-mail: matteo.farne@unibo.it

1 Introduction

The goal of the present work is to describe a statistical analysis of students' performance over time using the periodic assessments they receive during their school career.

According to the official website[1], Educational Data Mining (EDM) is defined as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale datasets that come from educational settings and using those methods to better understand students, and the settings which they learn in". On the other hand, Learning Analytics (LA) is defined as the process of analysing educational data which includes measurement, collection, analysis and reporting of data on students and the school context, to understand and optimize learning and learning environments [4]. For the purposes of both EDM and LA, there are ad hoc statistical methods that can be used to respond to the needs of both students and teachers [5].

In this paper, we first briefly recall in Section 2 the four clustering methods that we use, we describe our empirical analysis in detail in Section 3, and we draw some concluding remarks in Section 4.

2 Methodology: cluster analysis methods

In this section, the cluster analysis methods used for the empirical analysis are shown. We focus on four different clustering algorithms to analyse the empirical dataset.

Hierarchical clustering

Hierarchical cluster analysis methods can be divided into aggregative (bottom-up) and divisive (top-down); since the latter are little used, we mainly focus only on the former. Bottom-up methods generate a sequence of partitions of the n units. Starting with the first partition in n groups, partitions are derived in (n-1), (n-2), ... groups, until the one in which all units are grouped into a single group. A limitation of hierarchical methods is that once a merge has taken place, the groups can no longer be partitioned.

The hierarchical clustering algorithm first returns a partition of the n units into n clusters, so each unit is a stand-alone cluster; the matrix D will contain the distances or dissimilarities between the n units. In the next step, on the basis of the D matrix we identify the closest clusters, which are then aggregated to form a new cluster (so the number of clusters becomes n-1), and the distance of the new cluster from every other cluster is calculated according to some heuristic criterion. In the end, the D matrix is updated. This procedure is repeated until there is a single cluster containing all n units.

[1] Reference: https://educationaldatamining.org/

The described can be visualized by a graph called dendrogram, which shows how the groups are merged at each step.

In order to measure the distance between two clusters A and B, it is used in the case study the average linkage method: it calculates the average of all pairwise distances between the points belonging to cluster A and cluster B [1].

The k-means algorithm

Given a partition of n units in K groups, the sum of n x n pairwise dissimilarities between between the units is defined as the total dissimilarity, which can be decomposed into two components: dissimilarity into the groups and dissimilarity between groups. A partitive clustering method generates a single partition of the n units by fixing *a priori* a number of clusters equal to K (with K<n), given an optimality criterion: it suggests looking for groups with high internal cohesion and high external separation. Partitive methods propose a variety of strategies for identifying a narrow subset of partitions that could potentially be optimal. One of the best-known methods is the k-means algorithm: the dissimilarity measure is the squared Euclidean distance, and the best partition is the one where the units of each group are as close as possible to their centroid (Cerioli, Zani, 2007).

The factorial k-means algorithm

Vichi and Kiers [6] provided a theoretical framework for factorial k-means analysis for "two-way" data: this involves the simultaneous estimation of a discrete clustering model and a continuous factorial model, aimed at identifying the best partition of statistical units described by the best orthogonal linear combination of variables according to an alternated least squares algorithm.

In large datasets factorial k-means is a particularly recommended method, although in datasets containing dimensions with low variance there is a risk that it might focus primarily on these dimensions because they contribute little on the loss function. So, one should first remove these "misleading" dimensions from the data.

The reduced k-means algorithm

De Soete and Carroll [2] proposed reduced k-means clustering. It is an alternative to the k-means algorithm (which can be seen as its own special case) where each cluster is represented by a point, called "centroid", in a space of reduced dimensions. Points are projected onto the space containing the centroids, and thus a reduced-dimensional representation of objects and clusters is derived. In this method, a subspace of the original data space is found such that the actual points (in the original space) generate the smallest squared sums of the distances from the centroids (which are placed in the subspace instead).

Case study

These methods were applied to a dataset which contains the grades and hours of absence of a class composed by 17 students of a high school (specialized in classical studies, so there will be subjects as Greek and Latin) who are currently attending the second year of high school, with reference to the school years 2020/2021 (both first and second quarters) and 2021/2022 (first quarter only). The statistical software R was used to carry out computational routines. The algorithms described in Section 2 will be applied independently to the three datasets.

2 Dataset description and preliminary analysis

The reference period of analysis was the single quarter: t=1 corresponds to the first quarter of the first year, t=2 corresponds to the second quarter of the first year, t=3 corresponds to the first quarter of the second year.

The variables considered in the analysis are five for each quarter: the grade in Mathematics, in Italian, in Greek, in Latin and the hours of absence. The choice fell on these subjects as the former two are basic subjects, while the latter two are those that specifically characterize this high school. The hours of absence can be useful because they could affect the final evaluation.

Given the presence of a clear outlier by hours of absence, a preliminary analysis was carried out by the *trimcluster* package in R [3] to verify how the results varied in the presence and absence of the outlier in terms of average silhouette width (ASW): in t=1 and t=2 the partitions with k=3 groups had a systematically better value of Average Silhouette Width (ASW) without it, so the outlier was excluded.

This preliminary analysis was also useful to understand which could be the number of groups in each period. Thus, the clustering will be carried out without the outlier with respect to the "hours of absence", so the number of units is equal to 16.

3 Results

Table 1 *Summary of the results obtained with clustering methods for each period in terms of number of groups (k), size of the groups (n) and average silhouette width (ASW) of the partition.*

Methods	***t=1***	***t=2***	***t=3***
Hierarchical clustering	k=3 (n1=9; n2=4; n3=3) ASW=0.37	k=3 (n1=9; n2=4; n3=3) ASW=0.36	k=2 (n1=14; n2=2) ASW=0.34
K-means algorithm	k=3 (n1=9; n2=4; n3=3) ASW=0.37	k=3 (n1=9; n2=4; n3=3) ASW=0.36	k=2 (n1=7; n2=9) ASW=0.25
Factorial k-means	k=2 (n1=11; n2=5) ASW=-0.038	k=2 (n1=8; n2=8) ASW=0.108	k=2 (n1=10; n2=8) ASW=-0.034
Reduced k-means	k=3 (n1=9; n2=4; n3=3) ASW=0.373	k=3 (n1=9; n2=4; n3=3) ASW=0.368	k=2 (n1=9; n2=7) ASW=0.259

Hierarchical clustering, k-means algorithm, factorial k-means and RKM were applied to the dataset made up of 16 students in each period. The optimal number of groups was k=3 in t=1 and t=2, and k=2 in t=3 for hierarchical clustering, k-means algorithm and RKM; the optimal number of groups was k=2 in each period for FKM. These methods were compared in terms of average silhouette width. For FKM and RKM, there is one latent dimension behind the derived partition.

Hierarchical clustering results were the same in t=1 and t=2: group 1 comprises students with fairly good grades and many hours of absence, group 2 comprises students with high grades overall and few hours of absence and group 3 is made up of students with low grades overall and many hours of absence. In t=3, group 1 is made up of students with few hours of absence, higher grades in basic subjects, lower grades in Greek and Latin and group 2 is made up of students with many hours of absence and lower grades in basic subjects, higher grades in Greek and Latin ("new subjects").

Using the k-means algorithm, results were the same in t=1 and t=2 and equal to hierarchical clustering results. In t=3 group 1 comprises students with many hours of absence, lower grades in basic subjects, higher grades in new subjects and group 2 is made up of students with few hours of absence, higher grades in basic subjects and lower grades in new subjects.

Using FKM, in each period there are k=2 groups. In t=1 the latent dimension is the level of ability in basic subjects; group 1 comprises students with heterogeneous performance and group 2 is made up of students with higher grades in basic subjects. In t=2 the latent dimension is the grade point average; group 1 is made up of students with medium-low grades and group 2 is made up of students with good performance. In t=3 the latent dimension is the difficulty in Latin; group 1 is made up of students with very low grades in Latin (with two exceptions however) and group 2 is made up of students with failures only in Latin.

Lastly, RKM results were identical to k-means results (in t=3 group 1 and group 2 are inverted). The latent dimension in t=1 and t=2 is the grade point average and in t=3 is the level of ability in basic subjects.

4 Conclusions

The analysis of student performances relative to a sample of classical high school students generated reasonable results, despite the fact that the datasets were independently analyzed at each of the three available quarters.

Using four clustering methods, the results were quite similar: k-means and RKM generated the same partitions in each quarter; the results of hierarchical clustering were identical to those of k-means and RKM only in the first two quarters, probably because these are the two quarters of the first year, without variations in student group membership.

The results of FKM differ from the other methods by number of groups, group size and student membership: due to its greater complexity compared to the others, FKM is probably not the best algorithm for clustering this type of data, considering the presence of only 16 observations and 5 variables for each dataset.

In terms of data interpretation, the algorithms that produced the most acceptable results were k-means and RKM, even though the value of the ASW is not optimal. Since k-means is a particular case of RKM, the results of this last algorithm are the most acceptable, as a latent dimension is also produced.

A limitation of this analysis is that the results cannot be generalized. It is not guaranteed that another class of classical high school students presents the same difficulties in Latin as in the class examined. In light of this, we suggest considering the whole school year as the time unit and not the single quarter, given that the same memberships are observed in the first two datasets. Finally, for datasets of this type, it is advisable to use less complex methods, such as k-means or RKM, since there are few statistical units and few variables available.

References

1. Cerioli, A., Zani, S.: Analisi dei dati e data mining per le decisioni aziendali. Giuffrè Editore (2007)
2. De Soete, G., Carroll, J. D.: K-means clustering in a low-dimensional Euclidean space. In New approaches in classification and data analysis (eds). Diday, E, Lechevallier, Y, Schader, M, Bertrand, P & Burtschy, B), Springer, Heidelberg; 212-219 (1994)
3. Hennig, C.: trimcluster: Cluster Analysis with Trimming. R package version 0.1-5 (2015) https://CRAN.R-project.org/package=trimcluster
4. Lang, C., Siemens, G., Wise, A., Gasevic, D.: Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research, WIREs (2020)
5. Taraborrelli, G., Farnè, M.: Come sfruttare gli Educational Data? Un inquadramento di usi e metodologie di analisi. Induzioni 62/63, 27–39 (2022)
6. Vichi, M., Kiers, H. A. L.: Factorial k-means analysis for two-way data. Comp Stat Data Anal 37, 49–64 (2001)

The nexus of cultural capital with participation in early childhood education

Il nesso tra capitale culturale e partecipazione al Sistema Educativo 0-3

Enrico Ripamonti

Abstract Early childhood educational development (ECED) is increasingly recognized as a crucial period for human development. We studied at a local level the relation of cultural capital, economic capital, and participation in center-based ECED. We used a high-quality ISTAT database collecting information from 103 Italian provinces. Spatial and multivariable models were adopted for analysis. Preliminary results indicate that Central Italian provinces present the highest level of cultural capital, as well as the highest rate of participation in ECED. Our findings highlight the importance of investing in cultural capital as it may also foster the development and the rate of participation in center-based ECED. This may mitigate the effect on education of social disparities and economic inequalities.
Abstract *La prima infanzia è sempre più riconosciuta come un periodo cruciale per i successivi stadi dello sviluppo umano. In questo contributo studiamo a livello locale la relazione tra capitale culturale, economico e partecipazione agli asili nido. Utilizziamo una base di dati ISTAT che raccoglie informazioni da 103 province italiane. Per l'analisi sono stati utilizzati modelli spaziali e multivariabile. I risultati preliminari indicano che le province del Centro Italia presentano il più alto livello di capitale culturale, nonché il più alto tasso di partecipazione ai nidi. Sulla base di queste evidenze, sottolineiamo l'importanza di investire sul capitale culturale in quanto esso può anche favorire il tasso di partecipazione al Sistema Educativo 0-3. Ciò potrebbe mitigare l'effetto sull'istruzione delle disparità sociali e delle disuguaglianze economiche.*

Key words: cultural capital, economic capital, childcare, center-based care, early childhood education

Enrico Ripamonti
University of Brescia; e-mail: enrico.ripamonti@unibs.it

1 Introduction

There is an increasing need for macro-level indicators of early childhood educational development (ECED). From the perspective of the theory of evolutionary and dynamic systems it is essential to keep into account the macro-level correlates of individual behavior, such as the historical, social, political, and economic context experienced by children and their families. In particular, the very early phase of education is critical for mitigating the effects of social and economic inequalities [2]. The present investigation is conducted in Italy, namely one of the countries in the European Union (EU) with the deepest roots in ECED [3]. Italy has a very long pedagogic tradition on nursery and ECED. However, the EU objective of national coverage of 33% of the 0-3 population attending full-time services is still not achieved, as well as the objective of coverage of 75% of the municipalities with public ECED programs. Moreover, the Covid-19 crisis led to an increasing gap in terms of economic inequalities in Italy [1]. In this scenario, center-based ECED is still more important to mitigate the effect of socioeconomic disparities. In the spirit of the Italian Constitutional Law, the presence of these services guarantees all children equal opportunities for the development of their cognitive, emotional, and social skills in an inclusive educational environment. The recent Next Generation EU Recovery Plan led to an investment of 4.60 billion euros in the creation of services for early childhood and primary schools. This should allow increasing participation in these services for 228,000 children.

2 Objectives

We focused on aggregated-level measures of cultural and economic capital and on their possible role in predicting participation in center-based ECED. In light of the recent Covid-19 crisis, it has been highlighted the importance of ECED for children and young people's development and to mitigate the effect of inequalities. It is thus important to focus on those factors that, at a regional or local level, may favor participation in center-based ECED, including the putative role of cultural capital and economic capital [4,5]. Research is scarce in this respect. This study aimed to investigate at a local level the relation of cultural capital, economic capital, and the rate of participation in center-based ECED in Italy.

3 Approach

We adopted a place-based approach and a large high-quality national database. We used the publicly available BES database (Benessere Equo e Sostenibile in Italia, Istituto Nazionale di Statistica, ISTAT, 2021), which collects aggregated data from 2004 to 2019 on 103 Italian provinces. From this database, we selected a proxy of participation in early childhood education. ECED participation is measured in terms of the percentage of children aged 0-3 years old who have used center-based day-care services offered by public or private structures present in the municipality. According to the International Standard Classification of Education (ISCED) developed by UNESCO in 1976 (revised in 1997 and 2011), early childhood education (level 0) comprises two sublevels: ECED (educational content designed for children aged 0-3, namely level 01, the level considered herein); pre-primary education (referring to children from the age of 3 years up to the start of primary education, level 02). Data on cultural capital and economic capital were obtained from the yearly reports on the Quality of Life in Italy. We operationalized cultural capital in terms of living cultural capital and participation in cultural or culturally oriented activities. Spatial and multivariable models (Structural Equations Models, SEM adopting the LISREL terminology and notation) were used for data analysis.

4 Preliminary results

We created average indicators of the standardized variables taken as proxies of participation in ECED, cultural capital, and economic capital. Figure 1 shows the boxplots for these indicators.

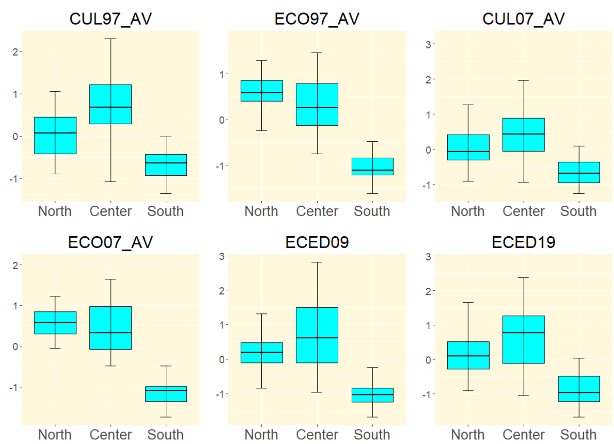


Fig. 1 Boxplots for the distribution of standardized scores of cultural capital (average indicator), economic capital (average indicator), and participation in center-based ECED in Italy in 2009 and 2019, by geographical area.

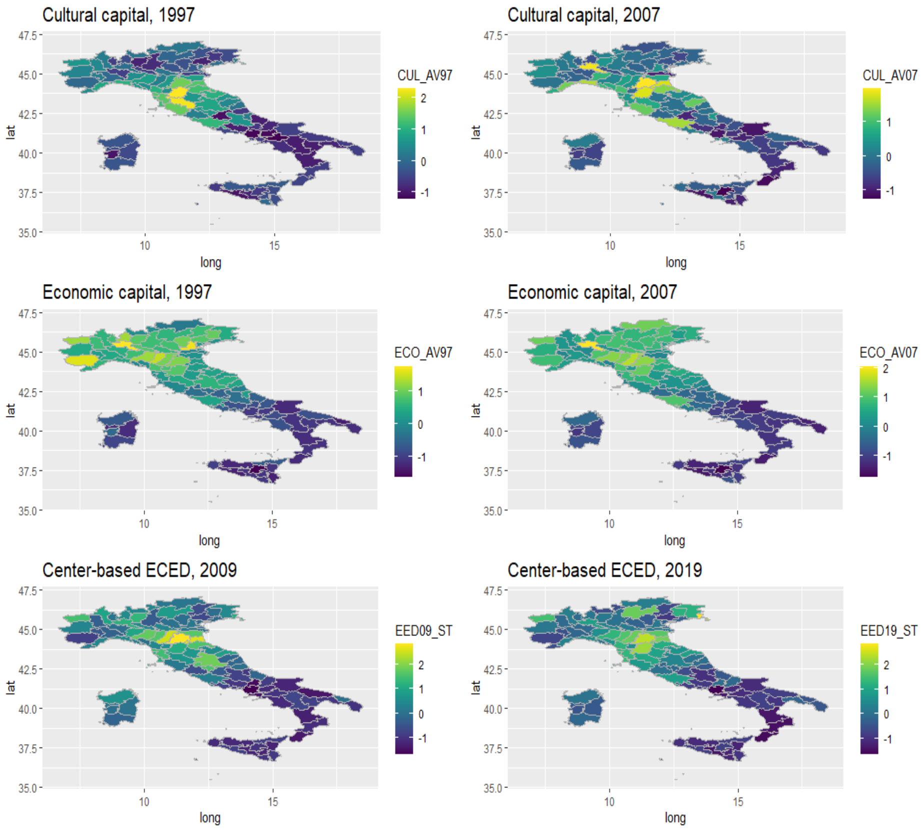


Fig. 2 Choropleth maps illustrating the distribution of standardized scores across Italian provinces for cultural capital in 1997 and 2007, economic capital in 1997 and 2007, participation in center-based ECED in 2009 and 2019.

We used choropleth maps (Figure 2) to contextualize the analysis at a fine-grained geographical level. Maps showed a decreasing trend from Northern provinces to Southern provinces in terms of the economic capital. As to cultural capital and center-based ECED, Central provinces obtained the highest standard scores. Across Southern provinces, we found very low standard scores, except for Sardinian provinces, which, notoriously have a better ECED system than other Southern provinces. Our data showed a significant correlation of cultural capital with the rate of participation in ECED, along with a historical effect played by the economic capital. We also estimated SEM in the entire sample of the Italian provinces. The path diagram of the model under study with the obtained coefficients is shown in Figure 3. In brief, SEM indicated that the economic capital in 1997 is associated with both cultural and economic capital in 2007. Cultural capital but not economic capital as measured in 2007 is associated with the rate of participation in center-based ECED in 2009 and

2019. At least in part, cultural capital in 2007 may mediate a historical effect played by the economic or cultural capital, as measured in 1997.

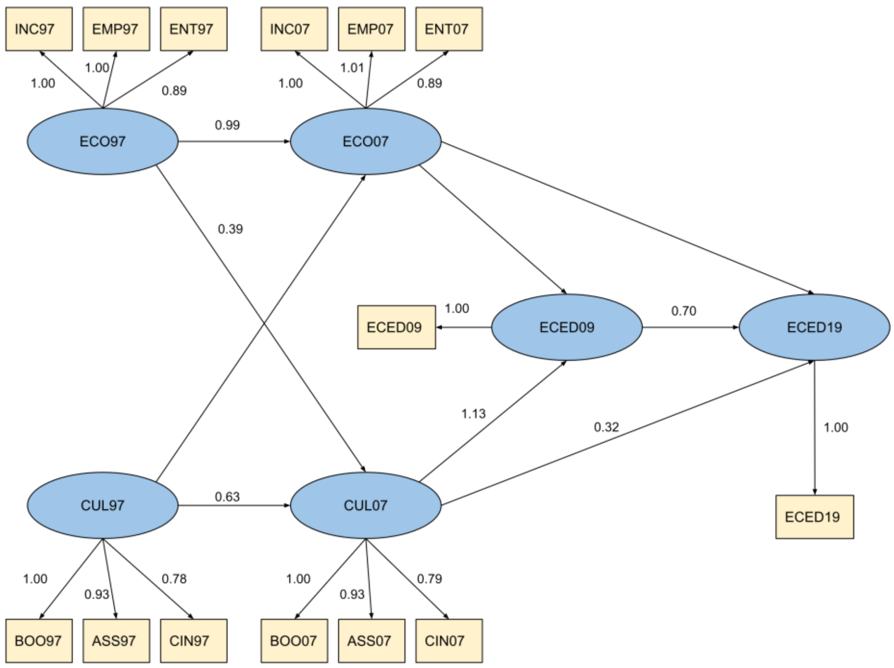


Fig. 3 Diagram for the SEM estimated for all the Italian provinces, and for the multi-sample SEM. We report factor loadings and Beta coefficients. CUL97 and CUL07 indicate cultural capital in 1997 and 2007, respectively; ECO97 and ECO07 indicate economic capital in 1997 and 2007, respectively; ECED09 and ECED19 indicate early childhood educational development in 2009 and 2019, respectively. INC = income pro capite; EMP = employment; ENT: number of enterprises; BOO = number of book shops; ASS = number of associations; CIN = number of cinemas.

5 Conclusions

Based on our data, we have observed a noteworthy association between cultural capital and the rate of participation in center-based ECED, as well as the historical influence of economic capital. We suggest that policy makers should monitor the rate of participation in center-based ECED while considering factors such as cultural capital. Our findings highlight the relevance of local context in evaluating ECED, as demonstrated by the case of Central Italian provinces. Therefore, implementing appropriate policies could improve the cultural capital of local areas and potentially enhance the rate of participation in center-based ECED in the medium term, reducing inequalities. Even though early environments are not irreversible as developmental trajectories are influenced by them, they play a pivotal role in shaping children's future outcomes. The presented analysis could prompt further studies on a microsystem level.

References

1. Carta, F. and De Philippis, M.: The impact of the COVID-19 shock on labour income inequality: evidence from Italy. Bank of Italy Occasional Paper, (606) (2021)
2. Heckman, J. J. Giving kids a fair chance. Boston: Mit Press (2013)
3. Musatti, T. and Picchio, M.: Early education in Italy: Research and practice. International Journal of Early Childhood, 42(2), 141–153 (2010)
4. Ripamonti, E. and Barberis, S.: The effect of cultural capital on high school dropout: An investigation in the Italian provinces. Social Indicators Research, 139(3), 1257-1279 (2018)
5. Ripamonti. E. and Barberis, S.: The association of economic and cultural capital with the NEET rate: Differential geographical and temporal patterns. Journal for Labour Market Research, 55(13), 13 (2021)

High- and Low-Performing students and future career: a gender and social issue

La carriera futura degli studenti fragili ed eccellenti: una questione di genere e sociale

Patrizia Falzetti and Roberto Ricci

Abstract University enrolment has been the focus of much attention in recent years both in public policies and from a scientific perspective. In addition, the EU Next Generation has emphasized the importance of increasing the number of people with a tertiary degree. The paper shows how many of the differences have their roots in educational outcomes. Raising achievement levels seems to be the most promising mean to foster successful participation in tertiary studies. Early results show large differences in the tendency of young people to pursue tertiary studies. Parts of these differences (territorial, socio-economic) were easily predictable, others, such as gender differences, are much less so and certainly require further research.
Abstract *L'iscrizione all'università è negli ultimi anni oggetto di grande attenzione sia delle policies pubbliche sia dal punto di vista scientifico. Inoltre, il Next Generation UE ha enfatizzato l'importanza di aumentare il numero di persone con un titolo di studio terziario. Il paper mostra come molte differenze affondino le loro radice negli esiti del percorso scolastico. L'innalzamento dei livelli di achievement sembra essere lo strumento più promettente per favorire la partecipazione con successo agli studi terziari. I primi risultati mostrano grandi differenze nella tendenza dei giovani a proseguire negli studi terziari. Parti di queste differenze (territoriali, socio-economiche) erano facilmente prevedibili, altre, come quelle di genere, lo sono molto meno e richiedono certamente ulteriori approfondimenti di ricerca.*

Key words: Low performers, high performers, territorial differences, gender gap.

Patrizia Falzetti
INVALSI, Statistic Department; e-mail: patrizia.falzetti@invalsi.it

Roberto Ricci
INVALSI; e-mail: presidente@invalsi.it

1 Introduction

In this paper we intend to analyse the school career up to the choice, or not, of the university by low performer or high performer students. What we would like to understand, also by means of a descriptive analysis of data at our disposal, is whether there are aspects, like gender, social and geographical origin, which might influence the choice to undertake a university career more than others.
Today, it is known that the background of students considerably may influence the choices of youngsters and their families. In this paper, we would also like to highlight the weight of other dimensions that are scarcely considered. Other factors might also exert a considerable influence on career choices. They are mainly related to the level of learning achievements, but also with significant gender differences. Moreover, the aforementioned factors exert interrelated effects, leading to considerable impacts on particular groups of students affected by joint characteristics that severely hamper them.

2 High and low performer students

The main purpose of this paper is to attempt to investigate the relationship of student performance levels, along with other factors, and subsequent university careers. For this reason, it is first to define what is meant by low performer and high performer students [7,1]. However, the problem is not only a definitional one, but also a methodological one. Indeed, it is well known that the categorisation of students according to their level of achievement requires a reliable standardised measure.
From a theoretical point of view, a correct way to measure the entry level of competence of students enrolling at university might be their grade in the state exam at the end of secondary school. However, it is well known that the state exam grade is hardly comparable between different schools, but especially across regions of the country. It therefore seems appropriate to supplement this grade with an external national standardised measure that could considerably reduce the aforementioned problems [4].
The data used in this paper refer to the INVALSI (Italian National Institute for evaluation) test results of the grade 13 population (school year 2018-19), for all the subjects investigated (reading comprehension, mathematics, English reading and listening) and to the enrolment of the same students in the academic year 2019-20 (source: Ministry of University and Research - MUR). Students from Valle D'Aosta and the Autonomous Province of Bolzano were not considered as they are not included in the national student register.
The results of the INVALSI tests (reading comprehension and mathematics) are expressed on a scale of 5 increasing levels [3], where levels 1 and 2 identify inadequate results with respect to the expected standards [5,6]. English results, on the other hand, are given according to the CEFR (Common European Framework of Reference for Languages) [2]. At the end of grade 13, attainment of level B2 is expected.

A low performer is defined as a student who does not reach level 3 in reading comprehension and Mathematics and does not reach B1 in neither English Listening nor English Reading. Instead, a high performer is defined as a student who reaches level 5 in both Italian and Mathematics (on the 5-level scale) and achieves B2 in both English Listening and English Reading.
According to the definition provided, Table 1 shows the distribution of students in Italy at the end of the 2018-19 school year.

Table 1 Low and High performers at the end of school year 2018-19

Level of achievement	***Number***	***Frequency (%)***
Low performers	33,963	7.4
High performers	30,730	6.6
Others	397,482	86.0
Total	462,175	100.0

Table 2 Low and High performers at the end of school year 2018-19 among Italian regions

Region	***Low performers***	***High performers***
Piemonte	3.4%	9.3%
Liguria	4.0%	7.9%
Lombardia	2.2%	11.9%
Veneto	2.1%	10.5%
Provincia Autonoma di Trento	1.1%	13.7%
Friuli-Venezia Giulia	1.8%	12.8%
Emilia-Romagna	3.6%	9.9%
Toscana	5.6%	7.0%
Umbria	5.7%	6.4%
Marche	5.8%	7.4%
Lazio	8.5%	4.8%
Abruzzo	9.2%	4.5%
Molise	10.9%	4.4%
Campania	12.2%	2.3%
Puglia	9.9%	3.8%
Basilicata	13.4%	4.0%
Calabria	16.2%	1.9%
Sicilia	13.0%	2.3%
Sardegna	13.9%	2.8%
ITALIA	7.4%	6.6%

3 Performance and enrolment at university

The level of achievement obviously has a weight on the decision to enrol at university. Table 3 shows the increasing relationship between performance level and enrolment at university.

Table 3 Enrolment rate according to achievement (academic year 2019-20)

Level of achievement	*Enrolled*	*Not enrolled*
Low performers	15.5%	85.5%
High performers	89.5%	10.5%
Others	53.2%	46.8%

At a first glance, there would seem to be a very strong association between achievement level and enrolment rate. Even this first simple observation would seem to suggest that probably the strongest action to raise the enrolment rate would be to increase the achievement level of students.
There are also very different trends in university enrolment rates among the regions of the country. But the most important aspect is that the differences observed mainly refer to low performers. In fact, Table 4 shows differences among Italian regions.

Table 4 Enrolment rate of low and high performers in the academic year 2019-20 among Italian regions

Region	*Low performers*	*High performers*
Piemonte	13.1%	88.7%
Liguria	15.5%	88.3%
Lombardia	14.2%	88.0%
Veneto	7.9%	87.3%
Provincia Autonoma di Trento	9.3%	87.1%
Friuli-Venezia Giulia	4.5%	85.4%
Emilia-Romagna	10.9%	90.3%
Toscana	15.6%	90.0%
Umbria	12.3%	90.8%
Marche	19.6%	91.6%
Lazio	21.2%	91.8%
Abruzzo	18.5%	95.1%
Molise	10.6%	91.2%
Campania	12.3%	94.3%
Puglia	13.8%	93.2%
Basilicata	18.8%	93.0%
Calabria	18.9%	93.0%
Sicilia	16.6%	91.8%
Sardegna	17.9%	90.8%
ITALIA	15.5%	89.5%

Overall, the data in Table 4 show that enrolment rates are generally higher in central and southern Italy. However, while in northern Italy the percentage of low performers enrolling in university is lower than average, in the centre-south of the country the opposite phenomenon is obviously observed. It is also of great interest to observe that the percentage of low performers enrolling in university is particularly low in northern Italian regions where vocational training is highly represented, as well as being a productive environment perhaps better able to employ these young people.
In general, it seems to be possible to say that the share of students enrolling at university is characterised by a smaller share of low performers. This aspect is far from being of secondary importance. In fact, if one considers that the objective of the university is to guarantee high levels of education, the presence of a student

population with higher entry performance certainly benefits the universities that receive these students.
To further investigate the gap in enrolment among students, it is certainly useful to observe whether there are any relevant differences according to gender. It is commonly believed that females enrol more in university than males. In fact, data seem to confirm this belief. However, it is worth examining whether this is true for all ability levels of students or whether there are differences.
Indeed, data in Table 5 seem to show important differences with respect to gender.

Table 5 Enrolment rate by gender according to achievement (academic year 2019-20)

Level of achievement	***Female***	***Male***
Low performers	20.8%	11.8%
High performers	89.7%	89.4%
Others	59.7%	46.4%

While the enrolment rate between males and females does not vary among high performers, for middle and lower skill levels the percentage of females who proceed to tertiary studies is considerably higher. This first result finds little echo in the literature and certainly deserves further investigation. This initial finding casts a different light on all the evaluation made so far of young people's university careers. In particular, for those courses of study in which the majority of students are females who did not attend the so-called strong high schools (classical high school and scientific high school).

4 Final remarks

In this paper, several descriptive analyses are examined through which an attempt is made to highlight career development in tertiary studies linked to certain student characteristics.
Already the first data show different dynamics that make the propensity to enrol at university very different among different groups of students. Some of these trends were easy to imagine, others much less so.
Firstly, the level of performance at the end of secondary school plays a very important role in the choice of whether to continue with university studies. While in part this outcome was entirely predictable, the strength of this association is remarkable. It therefore seems reasonable to argue that probably the most powerful driver for achieving the desired increase in tertiary enrolment is to try to raise learning levels at school. It might be possible that this improvement may play a very important role. This result thus seems to support the idea that guidance should be played more on the didactic level (didactic guidance) rather than on the informational level.
Secondly, territorial differences also seem to play a very important role. In the centre-south regions, there is a larger tendency of the population with lower levels of educational performance to pursue tertiary studies. Obviously, this may have a non-

negligible impact on the performance of universities, especially considering that they attract very different student populations.
In addition, there is a difference in choice of whether to enrol at university between males and females. While no gender gap is observed among high performers, at the lower end of the entrance ability scale the percentage of girls enrolling at university is markedly higher. This last aspect certainly deserves further research as it may have potential non-negligible impacts on all hypotheses of the gender gap in college careers. This aspect becomes even more important for those university curricula that mainly receive female students who did not attend the so-called strong high schools (classical high school and scientific high school).
The first simple analyses proposed in this paper indicate some research directions. It will be very important to observe the outcomes of these students during their university studies, that is, broadening the observation from the dynamics of choosing university (or not enrolling) to the outcome of university studies themselves.

References

1. Barabanti P.: Gli studenti eccellenti nella scuola italiana. Opinioni dei docenti e performance degli studenti, FrancoAngeli, Milano (2018)
2. Council of Europe: Common European Framework of Reference for Languages: Learning, teaching, assessment, (2001) https://rm.coe.int/1680459f97
3. De Simoni: I livelli per la descrizione degli esiti delle prove INVALSI. (2018) https://invalsi-areaprove.cineca.it/docs/2018/Livelli_INVALSI_g8.pdf
4. Falzetti P., Martini A.: L'esame di "maturità" e le prove INVALSI. WP 64 (05/2022) Fondazione Agnelli (2022)
5. INVALSI (a): Quadro di riferimento delle prove INVALSI di Italiano. (2018) https://invalsi-areaprove.cineca.it/docs/file/QdR_ITALIANO.pdf
6. INVALSI (b): Quadro di riferimento delle prove INVALSI di Matematica. (2018) https://invalsi-areaprove.cineca.it/docs/file/QdR_MATEMATICA.pdf
7. Vrapi, R., Alia, A., Brese, F.: Characteristics of High- and Low-Performing Students. In: Japelj Pavešić, B., Koršňáková, P., Meinck, S. (eds) Dinaric Perspectives on TIMSS 2019. IEA Research for Education, vol 13. Springer, Cham. (2022) https://doi.org/10.1007/978-3-030-85802-5_9

Solicited Session SS23 - *Statistical methods for environmental monitoring and sustainability*

Organizer and Chair: Francesca Fortuna

1. *Clustering spatial data through optimal transport* (Balzanella A. and Verde R.)
2. *New interpretative insights for environmental air quality by means of FDA* (Terzi S., Naccarato A. and Fortuna F.)
3. *A Bayesian State-Space Model to Mitigate Unmeasured Confounding* (Zaccardi C., Valentini P. and Ippoliti L.)
4. *Mining social media data for damage assessment in environmental disasters* (del Gobbo E., Cafarelli B., Ippoliti L. and Fontanella L.)

Clustering spatial data through optimal transport

Clustering di dati spaziali attraverso il trasporto ottimale

Antonio Balzanella and Rosanna Verde

Abstract This paper introduces a strategy for clustering point clouds generated by a spatial point process. The input dataset is a set of points in $\Re^d$ describing several events, each one made by a subset of the dataset. Our aim is to discover groups of similar events by means of an appropriate clustering strategy. We propose to cluster the events using a variant of the k-means algorithm based on the Sliced Wasserstein distance for probability measures. Preliminary results show the effectiveness of our proposal.

Abstract *Questo paper introduce una strategia per il clustering di nubi di punti generate da un processo di punto spaziale. Il dataset di input è un insieme di punti in* $\Re^d$ *che descrive diversi eventi, ognuno dei quali è composto da un sottoinsieme del dataset. Il nostro obiettivo è scoprire gruppi di eventi simili mediante una strategia di clustering appropriata. In particolare, proponiamo di raggruppare gli eventi utilizzando una variante dell'algoritmo k-means basata sulla distanza di Wasserstein Sliced per misure di probabilità. I risultati preliminari mostrano l'efficacia della nostra proposta.*

Key words: spatial data, optimal transport, clustering, Sliced-Wasserstein distance

Antonio Balzanella
Università della Campania Luigi Vanvitelli, Viale Lincoln 5 Caserta - Italy, e-mail: antonio.balzanella@unicampania.it

Rosanna Verde
Università della Campania Luigi Vanvitelli, Viale Lincoln 5 Caserta - Italy, e-mail: rosanna.verde@unicampania.it

1 Introduction

Many datasets collect space or space-time data concerning events that occur in specific geographic locations. For example, seismic events, crime data in a large city, geo-referenced disease cases, and mobile phone user positions.

In these examples, each event can be considered as a point pattern, and our dataset is made up of several point patterns. For instance, we can mention independent replications of biological experiments or different seismic events occurring over the years in a geographic area.

In this paper, we provide a strategy that partitions the events into homogeneous clusters.

We consider the cloud of points of an event as a realization of a spatial point process that follows a probability distribution. Therefore, to cluster the events, we propose a variant of the k-means algorithm based on the Wasserstein distance for probability measures and a consistent definition of centroid.

The Wasserstein distance has its roots in Optimal Transport Theory [8]. Generally, the optimal mass transport problem seeks the most efficient way to transform one distribution of mass into another, considering a given cost function for transporting individual masses.

The problem was initially studied by the French mathematician Gaspard Monge[6] in 1781. Later, in 1942, Leonid V. Kantorovich proposed a general formulation of the problem by considering optimal mass transport plans. Unlike Monge's formulation, Kantorovich's approach allows for mass splitting [5].

Viewing the problem of clustering the realizations of a spatial point process as an optimal transportation problem allows us to have clusters of events that are homogeneous in terms of the distance among the empirical distributions of the events. It also enables us to obtain optimal transportation plans that illustrate how a distribution should be transformed into another.

While analyzing point process data benefits from the new perspectives offered by Optimal Transport, computing the Wasserstein distance on multi-dimensional probability measures often becomes a bottleneck, limiting its practical application in large-scale data analysis. However, recent advancements have introduced several numerical methods aimed at accelerating the computation of the Wasserstein distance. Of particular interest is the Sliced Wasserstein distance, proposed in [2].

The underlying concept of the Sliced Wasserstein metric involves obtaining a collection of one-dimensional representations for a probability distribution in higher dimensions by performing projections (or slicing the measure). Subsequently, the distance between two input distributions is calculated as a functional of the Wasserstein distance between their respective one-dimensional representations. This approach allows the distance to be derived by solving multiple one-dimensional optimal transport problems, which conveniently have closed-form solutions.

The use of one-dimensional representations also facilitates the search for k-means centroids, which can be computed in a similar manner as in the one-dimensional case [1].

In the following sections, we provide a detailed explanation of our proposal and present some preliminary results on simulated data.

2 Clustering point process data based on the Sliced Wasserstein distance

Let $x_i = \{o_1, \ldots, o_j, \ldots, o_m\}$ be a spatial point pattern on an observation window $W \subset \Re^d$, $d \geq 2$. The points $o_j \in \Re^d$ form the event x_i occurring inside W.

Let Ψ be the the space of all spatial point patterns on W. Consistently with [4], x_i is a realization of a spatial point process X_i taking values in Ψ according to a probability distribution $\mathbb{P}_{X_i}$.

We consider the set $x = \{x_1, \ldots, x_i, \ldots, x_n\}$ as input of a clustering algorithm which provides a partition $P = \{C_1, \ldots, C_k, \ldots, C_K\}$ of x into K non overlapping clusters by minimizing the following heterogeneity function:

$$\Delta(P,G) = \sum_{k=1}^{K} \sum_{x_i \in C_k} d_W(x_i; g_k) \tag{1}$$

where $g_k \in G$ is the centroid of the cluster C_k.

The previous equation provides a k-means like criterion where x_i and g_k are realizations of spatial point processes. Thus, we need to use an appropriate distance $d_W(\cdot,\cdot)$ and define the corresponding Frechet minimizer g_k.

Let $p \in [1, +\infty)$ and $\mathscr{P}(\Re^d)$ be the set of probability measures on $\Re^d$ with finite moment of order p. The Wasserstein distance of order p between any $\mu, \nu \in \mathscr{P}(\Re^d)$ is defined as:

$$\mathbf{W}_p^p(\mu,\nu) = inf_{\pi \in \Pi(\mu,\nu)} \int_{\Re^d x \Re^d} \|o-q\|^p d\pi(o,q) \tag{2}$$

where $\|\cdot\|$ is the Euclidean Norm, $\Pi(\mu,\nu)$ the set of probability measures on $\Re^d x \Re^d$ whose marginals are μ and ν.

As introduced above, the optimization problem in eq.2 can be computationally expensive, however, for $d = 1$ and $\mu, \nu \in \Re$ it admits a closed-form solution:

$$\mathbf{W}_p^p(\mu,\nu) = \int_0^1 |Q_\mu(\tau) - Q_\nu(\tau)|^p d\tau \tag{3}$$

where $Q_\mu(\tau) = F_\mu^{-1}(\tau)$ and $Q_\nu(\tau) = F_\nu^{-1}(\tau)$ are the quantile function of μ and ν.

A recent proposal for addressing the high computational requirements for data in $\Re^d$, with $d > 1$, is the Sliced Wasserstein distance.

Let $\mathbb{S}^{d-1}$ be the d-dimensional unit sphere and σ the uniform distribution on $\mathbb{S}^{d-1}$. Following [7], for $\theta \in \mathbb{S}^{d-1}$ it is possible to define the map $R : \Re^d \to \Re$ denoting the linear form $o \mapsto \langle \theta, o \rangle$, with $\langle \cdot, \cdot \rangle$ the Euclidean inner-product.

The Sliced Wasserstein distance of order $p \in [1, \infty)$ between two measures $\mu, \nu \in \mathscr{P}(\Re^d)$ is:

$$\mathbf{SW}_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p(R_{\sharp\mu}, R_{\sharp\nu}) d\sigma(\theta) \tag{4}$$

where $R_{\sharp\mu}, R_{\sharp\nu}$ are the univariate distributions (push forward measures) of the projection of μ, ν over θ.

Being $R_{\sharp\mu}, R_{\sharp\nu}$ univariate distributions, $\mathbf{SW}_p^p(\mu,\nu)$ is computed by eq.3 so that the initial problem of computing $\mathbf{W}_p^p$ is addressed using projections of the measures in $\Re^d$ to $\Re$. On practice, Eq. 4 is approximated by:

$$\mathbf{SW}_p^p(\mu,\nu) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{W}_p^p((R_t)_{\sharp\mu}, (R_t)_{\sharp\nu}) \tag{5}$$

where $(R_t)_{\sharp\mu}, (R_t)_{\sharp\nu}$ project μ, ν on $\Re$ according to θ_t i.i.d from σ

In eq. 1, which expresses the clustering optimization criterion, d_W is replaced by the Sliced Wasserstein distance in eq.5. This allows to evaluate the proximity between each event x_i and the cluster centroid g_k through an analytical form based on quantile functions.

Consistently, the centroid g_k can be defined for each θ_t as in the univariate case, as the measure associated to the quantile function $Q_k^{\theta_t}(\tau)$:

$$Q_k^{\theta_t}(\tau) = \frac{1}{|C_k|}\sum_{x_i \in C_k} (Q_{x_i}^{\theta_t}(\tau) \tag{6}$$

where $Q_{x_i}^{\theta_t}(\tau)$ is the empirical quantile function associated to a realization x_i of the spatial point pattern X_i according to the projection on θ_t

3 Preliminary results and conclusions

In this paper, we have introduced a strategy for clustering events described by realizations of spatial point processes. To consider the distribution that generates each event, we proposed using the Wasserstein distance as a measure for aggregating events into homogeneous clusters. However, computing the Wasserstein distance becomes computationally demanding when dealing with data in more than one dimension. To address this issue, we utilized the Sliced Wasserstein distance, which allows us to apply the 1D Wasserstein distance on suitable projections of $\Re^d$ probability measures.

Fig. 1 shows a set of events generated according to two Gaussian random variables in $\Re^3$. On the left side, the initial events are depicted, while on the right side, it is evident that our procedure successfully recognizes the correct partition of the events into two clusters.

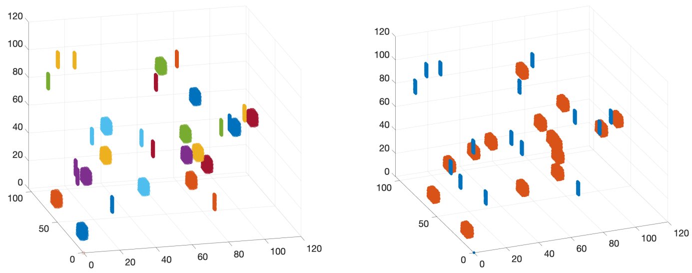


Fig. 1 The left side shows the input events. The right side, provides the clustering according to the Sliced Wasserstein distance

References

1. Balzanella, A., Verde, R.: Histogram-based clustering of multiple data streams. Knowl Inf Syst 62, 203–238 (2020). https://doi.org/10.1007/s10115-019-01350-5
2. Bonneel N., Rabin J. , Peyré G., Pfister H.: Sliced and Radon Wasserstein barycenters of measures. J. Math. Imaging Vision, vol. 51, no. 1, pp. 22–45 (2015)
3. Ester, M., H. P. Kriegel, J. Sander, and X. Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. (1996)
4. Jalilian, A., Mateu, J.: Assessing similarities between spatial point patterns with a Siamese neural network discriminant model. Adv Data Anal Classif 17, 21–42 (2023) https://doi.org/10.1007/s11634-021-00485-0
5. Kantorovich L. V.: On translation of mass, Dokl. AN SSSR, 37:199–201 (1942)
6. Monge G.: Mémoire sur la théorie des déblais et des remblais. Paris, France: De l'Imprimerie Royale (1781)
7. Nadjahi K., Durmus A., Jacob P.E., Badeau R., Simsekli U.: Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections. ArXiv abs/2106.15427 (2021)
8. Villani C.: Optimal Transport. Old and New.Springer Berlin, Heidelberg (2008)

New interpretative insights for environmental air quality by means of FDA

Nuovi spunti interpretativi per la qualità dell'aria attraverso l'approccio FDA

Silvia Terzi, Alessia Naccarato and Francesca Fortuna

Abstract A common practice to assess air quality is to express concentrations of pollutants in terms of a single measure, called air quality index. This strategy emphasizes the need to investigate the relationships between pollutants. To this end, an original area-based measure of association among pollutants of the same station is proposed using the functional data analysis approach. The suggestion is based on the area between two curves: the upper and lower envelopes, that is the curves passing through the maximum and minimum values of pollutants at each time instant. The area-based measure of association provides new insights into the interrelationships between pollutants and also reflects their evolutionary behaviour over time.
Abstract *Una pratica comune per valutare la qualità dell'aria è esprimere le concentrazioni di inquinanti in termini di una singola misura, chiamata indice di qualità dell'aria. Questa strategia enfatizza la necessità di investigare le relazioni tra gli inquinanti. A tal fine, si propone un'originale misura di associazione tra inquinanti della stessa stazione usando l'approccio dell'analisi funzionale dei dati. La proposta si basa sull'area tra due curve: l'envelope superiore e inferiore, ovvero le curve che passano attraverso i valori massimi e minimi degli inquinanti in ogni istante di tempo. La misura di associazione basata sull'area fornisce nuove informazioni sulle interrelazioni tra gli inquinanti e riflette anche il loro comportamento evolutivo nel tempo.*

Key words: multivariate association, functional data analysis, air quality index

Silvia Terzi
Roma Tre University, Rome, e-mail: silvia.terzi@uniroma3.it

Alessia Naccarato
Roma Tre University, Rome, e-mail: alessia.naccarato@uniroma3.it

Francesca Fortuna
Roma Tre University, Rome, e-mail: francesca.fortuna@uniroma3.it

1 Introduction

Air quality represents a crucial issue in assessing sustainable development worldwide as it is closely related to public health. As a consequence, countries continuously assess air quality by means of monitoring stations, which collect data on air pollution concentrations and meteorological information at several locations with different time span resolution. In this context, common air quality strategies have been established to allow comparisons across countries and the "Air quality guidelines for Europe" of the World Health Organization (WHO) [15] represents a fundamental reference to promote standard setting procedures. However, standard guidelines address single pollutants whereas in real life, exposure to mixtures of chemicals occurs with additive, synergistic or antagonistic effects [10]. Although consideration should be given to the interrelationships between the various air pollutants, our knowledge on this aspect is still rudimentary [10]. Since concentrations of individual pollutants can be difficult to interpret, a common practice is to express them in terms of air quality indices (AQIs) [10, 1], which summarize complex situations in a single measure. AQIs are impact-related with respect to people's well-being and actually represent the most diffused approach in air quality assessment [10]. In the last decades, several AQIs have been proposed, and no unique index has been established, showing the lack of a common strategy. Plaia and Ruggieri [10] have provided a detailed review of AQIs, distinguishing them into indices that consider the conjoint effect of pollutants and indices that are based only on the actual more dangerous pollutants.

Taking moves from the widely recognised necessity to investigate the association among time series of pollutants, we suggest and implement, by means of the functional data analysis (FDA) approach [11], an original area-based measure of association among pollutants of the same station. FDA provides several advantages for the analysis of air quality data compared to traditional multivariate approaches; indeed it can handle high-dimensional data, and allows to deal with data measured at different time resolutions as well as with missing data. Moreover, the functional approach can be advantageous in detecting trends and patterns in data that may be hard to discern when dealing with discrete observations. Furthermore, the continuous functional form of the data provides additional information, such as rate of change, acceleration, and dynamic changes over a large-scale domain [4, 5]. In other words, the FDA approach renders the data more informative and provides an easy visualization of the pollutants patterns throughout the entire domain. The use of FDA techniques for the analysis of environmental data has received considerable attention in the past two decades. Indeed, several authors have used this approach to model the dynamic pattern of pollutants [2, 8, 9, 3], to identify outliers [7, 12, 13], to forecast pollutant values [14] or to cluster countries according to their pollution level [6]. In this paper, the representation of pollutants as functions makes it possible to exploit functional tools such as the area under the curve, which reflects the temporal dynamics of pollutants throughout the temporal domain. Specifically, the proposed measure is based on the area under two curves: the upper and the lower envelopes of the pollutants.

The remainder of the paper is the following: Sect. 2 presents an area-based measure of association, Sect. 3 deals with a real case study concerning air quality in the Tuscany region (Central Italy), and Sect. 4 provides some conclusions.

2 An area-based measure of association

Suppose a multidimensional phenomenon is observed on S units with I different dimensions in several time occasions. In our case, air pollution is observed on S monitoring stations, where I pollutants are measured over time. A common way to express a multidimensional phenomenon such as air quality, is to consider a synthetic indicator such as the AQUI, which synthesises the different dimensions in a single measure.
For each station, both temporal observations on I components and the synthetic indicator are represented as functions using the FDA approach. Then, we can derive a distance-based measure of the temporal association among the components, computing the distance between each functional dimension and a reference curve taken as a synthesis. Naturally, the smaller the sum of these distances, the greater the association among the components.
In this paper, we suggest an original area-based measure obtained by computing the area between two reference functions: $AQUI^{(s)}_{max}$ and $AQUI^{(s)}_{min}$, that is the upper and lower envelopes of the observed pollutants. $AQUI^{(s)}_{max}$ is the curve passing through the temporal maxima of the observed values of the pollutants, and $AQUI^{(s)}_{min}$ is the function passing through the temporal minima. Of course, the smaller the area, the greater the association among the pollutants. Then, to derive a relative area-based measure of association, we need to compute the maximum area and consider its ones complement to obtain a relative measure of association. Specifically, the proposed measure for the S-th unit, IA_s, is computed as follows:

$$IA_s = 1 - \left(\frac{\int AQUI^{(s)}_{max}\, dt - \int AQUI^{(s)}_{min}\, dt}{\gamma_s} \right), \tag{1}$$

where $\gamma_s = \max\left(\int AQUI^{(s)}_{max}\, dt - \int AQUI^{(s)}_{min}\, dt \right)$.

3 Application

The data set consists in daily air quality data collected in 2021 from 8 monitoring stations installed in the Tuscany region (Central Italy): 2 stations in the province of Arezzo (Acropoli and S. Giovanni), 2 in the province of Florence (Bassia and Gramsci), 3 in the province of Livorno (Enistagno, Lapira and Piombino) and 1

in the province of Prato (Roma). The pollutants under consideration are: Benzene, Nitrogen oxides, $NOx = NO_2 + NO$, Toluene and fine particulate matter, $PM10$. Thus, for each s-th station, $s = 1, 2, \ldots, 4$, $x_i(t_l)$ indicates the value of the i-th pollutant, $i = 1, \ldots, 4$, for the l-th day of the year, $l = 1, \ldots, 365$. Data are available on the ARPAT web site https://www.arpat.toscana.it/temi-ambientali/aria/qualita-aria/archivio_dati_orari.
Although the concentrations are all expressed in micro-grams per meter cube (μ/m^3), data have been normalized using the min-max scaling as follows:

$$X_i(t_l) = \frac{X_i(t_l) - \min\Big(X_i(t)\Big)}{\max\Big(X_i(t)\Big) - \ min\Big(X_i(t)\Big)}, \tag{2}$$

where $\min\Big(X_i(t)\Big)$ and $\max\Big(X_i(t)\Big)$ represent the minimum and the maximum value of the i-th pollutant in the entire temporal domain for each station, respectively. Equation (2) allows to consider a set of comparable and dimensionless observations, whose values range from 0 to 1. Starting from the normalized data, for each station and for each day, $AQUI^{(s)}_{max}$ and $AQUI^{(s)}_{min}$ are computed as the envelope passing through the maxima and the minima of the 4 pollutants, respectively. Then, normalised raw data are converted into functions, $X_i(t)$ by using a Fourier basis system for constrained functions (see [11] for further details). Also $AQUI^{(s)}_{max}(t_l)$ and $AQUI^{(s)}_{min}(t_l)$ are converted into smooth functions using the same basis system. Fig. 1 shows the functional pollutants $X_i(t)$ together with the $AQUI_{max}(t)$ and $AQUI_{min}(t)$ functions (with dotted lines) for each s-th station. It clearly emerges that $AQUI_{max}(t)$ is determined almost exclusively by $PM10$ (green lines), except in the case of Bassi (Florence), where Toluene (blu line) also plays an important role. Gramsci station (Florence) differs greatly from the others because $AQUI_{max}(t)$ reaches the maximum value over the entire domain and is determined almost exclusively by NOx (red line); moreover, $AQUI_{min}(t)$ shows values that are farther from zero. Once a functional approximation of pollutants has been provided, the area-based association measure can be computed for each station as in Equation (1). Table 1 shows the IA values. The associations range between a minimum of 0,49 and a maximum of 0,7, consistently with what can be seen in Fig. 1.

4 Conclusions

In the definition of an air quality index, the role of both the interaction among pollutants and their evolutionary behaviour are widely recognised. However, the understanding of these aspects is still rudimentary. The internal association measure proposed in this contribution aims to provide a deeper insight in these issues. Taking moves from the functional reconstruction of the pollutants, the suggested measure is based on the area under two curves: the upper and the lower envelopes of the

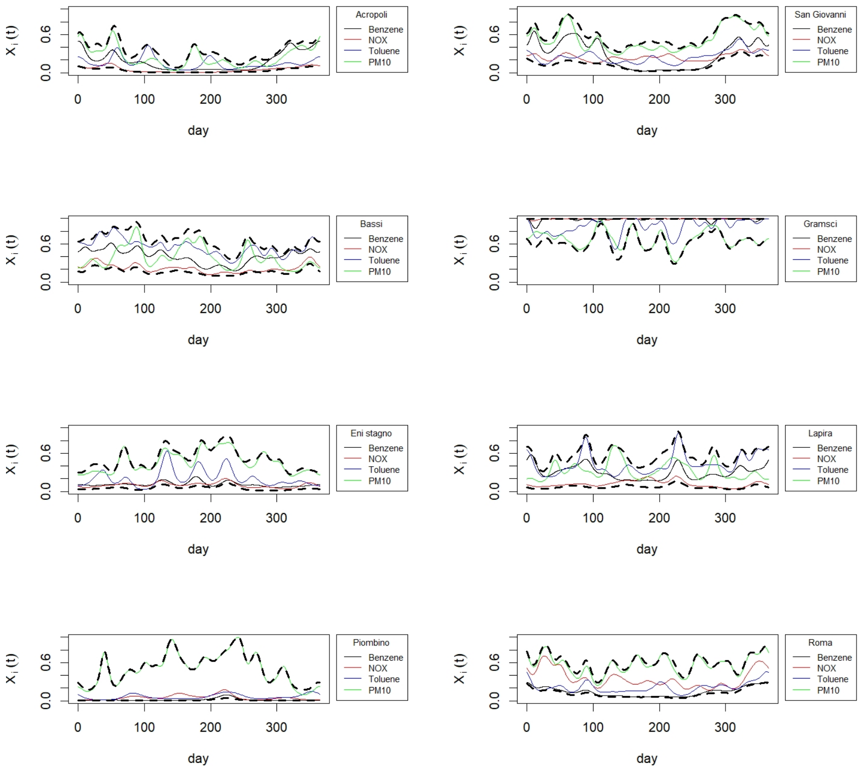


Fig. 1 Functional pollutants, functional $AQUI_{max}$ and $AQUI_{max}$ for the 8 stations of the Tuscany region.

pollutants of each station, providing an internal association measure, able to reflect

Table 1 Area-based association values for the 8 stations of the Tuscany region

Station	A_s
Arezzo Acropoli	0.7
Arezzo San Giovanni	0.54
Florence Bassi	0.52
Florence Gramsci	0.62
Livorno Enistagno	0.54
Livorno Lapira	0.51
Livorno Piombino	0.49
Prato Roma	0.52

the behaviour of the functions in the entire domain.
As for further developments, the suggested procedure can be usefully applied to compare association among stations with respect to a specific pollutant.

References

1. Bruno, F., Cocchi, D.: A unified strategy for building simple air quality indices. Environ Geol 13, 243–261 (2002)
2. Caligiuri, L.M., Costanzo, G.D., Reda, A.: The study of ground ozone concentration levels: a functional analysis approach based on principal components analysis. WIT Trans Ecol Environ 67, 82–59 (2005)
3. Escabias, M., Aguilera, A.M., Valderrama, M.J.: Modeling environmental data by functional principal component logistic regression. Environmetrics 16(1), 95–107 (2005)
4. Fortuna, F., Naccarato, A., Terzi, S.: Country rankings according to well-being evolution: composite indicators from a functional data analysis perspective. Ann Oper Res (2022a) doi: doi.org/10.1007/s10479-022-05072-w
5. Fortuna, F., Naccarato, A., Terzi, S.: Evaluating countries' performances by means of rank trajectories: functional measures of magnitude and evolution. Comput Stat (2022b) doi: doi.org/10.1007/s00180-022-01313-5
6. Ignaccolo, R., Ghigo, S., Giovenali, E.: Analysis of air quality monitoring networks by functional clustering. Environmetrics 19, 672–686 (2008)
7. Martínez, J., Saavedra, A., Garcia-Nieto, P.J. et al: Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). Appl Math Comput 241, 1–10 (2014)
8. Meiring, W.: Oscillations and time trends in stratospheric ozone levels: a functional data analysis approach. J Am Stat Assoc 102(479), 788–802 (2007)
9. Park, A., Guillas, S., Petropavlovskikh, I.: Trends in stratospheric ozone profles using functional mixed models. Atmos Chem Phys, 11473–11501 (2013)
10. Plaia, A., Ruggieri, M.: Air quality indices: a review. Rev Environ Sci Biotechnol 10, 165–179 (2011)
11. Ramsay, J.O., Silverman, B.: Functional Data Analysis, 2nd ed., Springer International Publishing: New York (2005)
12. Sancho, J., Martínez, J., Pastor, J.J. et al.: New methodology to determine air quality in urban areas based on runs rules for functional data. Atmos Environ 83, 185–192 (2014)
13. Shaadan, N., Jemain, A.A., Latif, M.T., Deni, S.M.: Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. Atmos Pollut Res 6, 365–375 (2015)
14. Valderrama, M.J., Ocana, F.A., Aguilera, A.M., Ocana-Peinado, F.M.: Forecasting pollen concentration by a two-step functional model. Biometrics 66, 578–585 (2010)
15. WHO: Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005. Summary of risk assessment. Geneva, World Health Organization (2006) https://apps.who.int/iris/handle/10665/69477

A Bayesian State-Space Model to Mitigate Unmeasured Confounding

Un modello bayesiano per mitigare gli effetti dei fattori confondenti non misurati

Carlo Zaccardi, Pasquale Valentini and Luigi Ippoliti

Abstract When assessing the consequences of air pollution exposure on public health, the magnitude of the relationship between a pollutant and an outcome is unlikely to be constant in space or time. This research seeks to add to the existing literature by suggesting a Bayesian state-space regression model that can account for unmeasured confounders. The proposed model is validated thanks to a simulation study, wherein different scenarios are explored but the coefficient representing the effect of the exposure is recovered even in the most difficult case.

Abstract *Quando si vogliono fare valutazioni sulle conseguenze negative che l'esposizione all'inquinamento atmosferico ha sulla salute pubblica, è molto improbabile che la magnitudine della relazione tra un inquinante atmosferico e un "outcome" sia costante nello spazio o nel tempo. Con questo articolo, si vuole contribuire alla letteratura scientifica esistente dato che viene proposto un modello di regressione bayesiano nello spazio degli stati che sia in grado di catturare l'effetto di fattori confondenti non misurati. Il modello proposto viene validato grazie ad uno studio di simulazione in cui sono presi in considerazione diversi scenari, ma il coefficiente che rappresenta l'effetto dell'esposizione viene recuperato anche nel caso più complesso.*

Key words: DGLM, Bayesian, air pollution, health, unmeasured confounding.

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

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

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

1 Introduction

The association between a disease outcome and an exposure is of primary interest when doing research to examine the health impacts of air pollution on population. These variables typically not only reveal long-term trends and seasonal variations, but also spatial correlation, so that spatio-temporal modeling is essential to comprehend the exposure–outcome relationship in greater detail and investigate fresher, better answers. Since the major pollutants in the atmosphere exhibit both temporal and spatial patterns, we should be aware that their effect on human health may have different magnitudes in each geographical unit and time instant: in a regression framework, coefficients that vary in the spatio-temporal domain, for instance, are discussed by [1, 14]. However, in most applications the association of main interest is likely to be confounded by, for example, environmental factors (such as temperature or humidity), influenza epidemics and human habits: whereas for the former ones data is available at many temporal and spatial resolutions, information is sparse and imprecise for the latter ones [4]. Confounding problems are difficult because they make it impossible for a model to recover the effect of the exposure if the unmeasured confounders are not appropriately taken into account in the model specification.

Varying-coefficients models (VCMs, [7]) are popular in many different study fields to increase flexibility of statistical models [6]. Basically, VCMs can be envisioned as a large family comprising, among the others, the linear multiple regression model, the generalized additive model (GAM, [15]), and the time-varying coefficient model (TVCM), that is a VCM with regression coefficients that vary in time. VCMs have been used in health studies as well: some authors take advantage of GAMs in their specifications to incorporate smooth functions of measured confounders (when these have non-linear effects on the linear predictor), and of calendar time to account for unmeasured factors [4, 8, 13, 16]. Other authors model regression coefficients with a temporal dynamic proposing a variety of TVCMs [9, 11].

A subgroup of TVCMs gathers the dynamic generalized linear models (DGLMs), wherein the temporal dynamic is specified in a state-space formulation and the Markovian assumption is made on the time-varying coefficients [12]. DGLMs could simultaneously emphasize seasonal and non-systematic variations due to their capacity to capture gradual changes [2]; in other words, this class of models could account for unmeasured confounding and potentially restore the desired relationship. Additionally, we prefer the use of DGLMs over GAMs because, while both are capable of handling potential non-linear effects, the coefficients in the former may be interpreted easily, whereas it is more challenging to identify the association of primary interest in the latter. To our knowledge, time-varying regression coefficients and unmeasured confounding have not been taken into account together in spatio-temporal health research, hence our intention is to address this knowledge gap in the existing literature.

The neighboring scheme among the areal units is another modeling issue that we would rather overcome with our model discussed in the next Section: in fact, it is typically not taken into account in multi-site time series studies (e.g., [8, 9, 13, 16]). As a result, any pair of measurements taken on the outcome at two areal units are assumed to be conditionally independent, even if these areas are adjacent. The conditional independence assumption should be relaxed because the residuals may show some spatial correlation. This often entails including a random effect in the model with a conditionally autoregressive (CAR) prior specification.

2 The Model

For $t = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, N$, let Y_{it} represent the adverse health outcome observed on time instant t in areal unit i, and let X_{it} represent concentration or emission of a pollutant at time t in area i. Assume that there are p measured spatio-temporal confounders, defined here as $M_{1it}, \ldots, M_{pit}$, whereas U_{it} represents an unmeasured confounder. We additionally assume that the confounding factors vary at both temporal and spatial scales coarser than those of the exposure, which is a crucial assumption under which the confounding bias may be mitigated. If the outcome, Y_{it}, represents count data (e.g., mortality counts, morbidity counts, or number of Emergency Department visits for a certain disease), then it should be modeled using a Poisson distribution with conditionally independent realizations, namely:

$$Y_{it}|\tilde{Y}_{it} \stackrel{ind}{\sim} Poi(E_{it} \exp\{\tilde{Y}_{it}\}), \tag{1}$$

$$\tilde{Y}_{it} = \eta_{it} + \varepsilon_{it}, \tag{2}$$

$$\eta_{it} = \beta_0 + \beta_{1t} X_{it} + f(M_{1it}, \ldots, M_{pit}) + g(U_{it}) + v_{it}, \tag{3}$$

where E_{it} denotes the expected count in area i at time instant t; η_{it} is the true (latent) spatio-temporal process that is represented as a function of exposure and confounders; ε_{it} is a spatially and temporally uncorrelated Gaussian random variable, assumed to have zero mean and variance σ^2_ε; in addition, $f(\cdot)$ and $g(\cdot)$ in Equation (3) are unknown smooth functions, and v_{it} is an error term such that the N-dimensional vector $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$ is modeled using a zero-mean CAR specification with non-singular precision matrix $\mathbf{Q} = \sigma^{-2}(\mathbf{D} - \rho\mathbf{W})$, where $\mathbf{W}$ is the spatial contiguity matrix with element on the ith row and jth column either equal to 1 (if areal units i and j are contiguous) or equal to 0 (if units i and j are not adjacent or if $i = j$), $\mathbf{D}$ is a diagonal matrix with the ith diagonal element equal to the number of neighbors of unit i, $\sigma^2 > 0$ represents the conditional variance of the process, and ρ is a smoothing parameter ensuring that $\mathbf{Q}$ is not singular [1].

Whereas the most popular method to account for unmeasured confounding is to include bases derived from a natural cubic spline of calendar time [5, 8, 10], we suggest instead to specify a temporal dynamic for area-specific intercepts and the exposure's coefficient as follows. Under the assumption that $g(\cdot)$ is a smooth function, Equation (3) can be rewritten as:

$$\eta_{it} = \beta_{0it} + \beta_{1t} X_{it} + f(M_{1it}, \ldots, M_{pit}) + v_{it},$$

wherein the time-varying intercepts β_{0it} adjust for unmeasured confounding, so the parameter of main interest, β_{1t}, can be recovered. To model the dynamics of the regression parameters, β_{0it} and β_{1t}, a DGLM is specified. The probabilistic inference for the parameters is aided by a Markov chain Monte Carlo (MCMC) sampler, which is based on adaption to the proposed model of conventional DGLM algorithms. This is done after specifying the prior distributions for all hyperparameters.

3 Simulation Study

In this section, a simulation study is set up to evaluate the performance of the suggested DGLM in terms of its capacity to recover the main regression coefficient of interest, β_{1t}. The objective is to evaluate the magnitude of the effect on mortality of $PM_{2.5}$ concentrations, i.e. airborne particles with a diameter of less than 2.5 μm. We collected $T = 450$ daily satellite observations of $PM_{2.5}$ concentrations (in μg/m^3), temperature, and relative humidity (two variables that are well-known to be confounders) from the Copernicus Atmosphere Monitoring Service's Atmosphere Data Store [3], in order to simulate data that are similar to a real scenario. The discrete spatial domain is partitioned in $N = 117$ health districts within Piemonte and Lombardia regions. Further details on the data collection will be given in an extended version of this paper.

For each district, we simulate daily death counts in accordance with Equations (1)–(3), where $f(\cdot)$ and $g(\cdot)$ are taken to be natural cubic splines. U_{it}, in particular, is generated using lagged versions of a natural cubic spline of time, where the lag value is different in each health district: it could represent influenza or other epidemic outbreaks. It is assumed that the effect of interest, β_{1t}, either is constant in time or has a periodicity with maximum and minimum values experienced during the winter and summer, respectively. Figure 1 represents the temporal average (left) and time series (right) of the simulated standardized mortality ratio (SMR), calculated as $T^{-1}\sum_{t=1}^{T} Y_{it}/E_{it}$, for each district $i = 1,\dots,N$, when β_{1t} is a periodic function.

The simulation findings are shown in Figure 2 as a percentage increase in (simulated) daily mortality corresponding to a 10-unit increase in exposure level. The ”true” simulated effects are shown as red dashed lines, whereas the posterior means are shown as black solid lines, along with their corresponding 95% credible intervals (CIs), which never contain the zero. The left plot displays the result of the simplest case, in which the effect is considered to be constant throughout time: the dashed line is included in the CI, indicating that the relationship of interest is fully recovered. The right plot shows the scenario when the effect of exposure varies over time. Similar to the preceding instance, the relationship is almost fully recovered here as well, and the pattern described by the CI over time suggests that the main effect cannot be considered time-invariant, as expected.

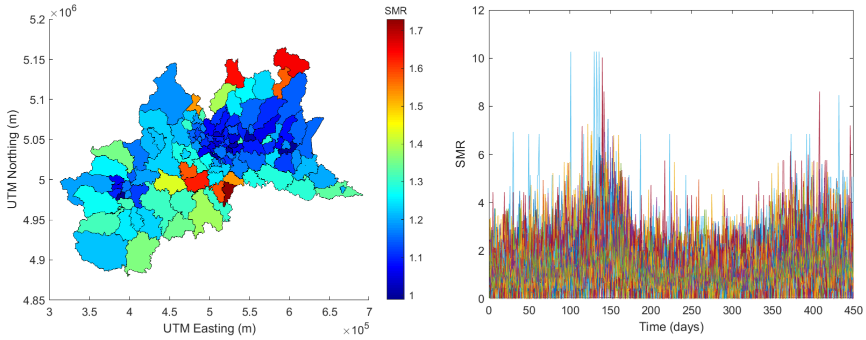


Fig. 1 On the left, the simulated SMR is averaged over the entire study period, for each health district in Piemonte and Lombardia regions. On the right, the corresponding time series are shown.

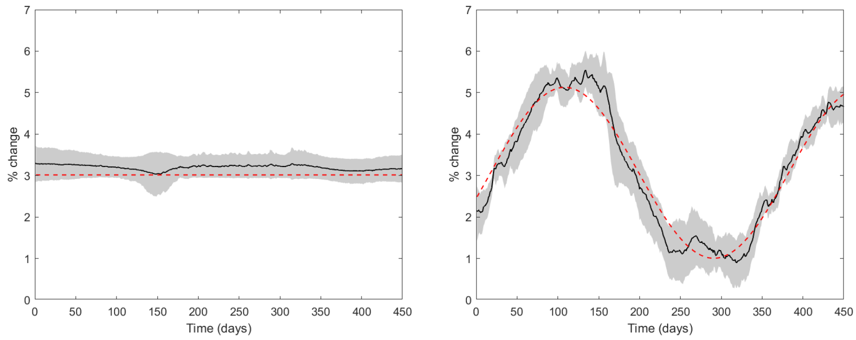


Fig. 2 Relations from the simulation study between exposure and health outcome. Results are presented as percent changes in risk for exposure increases of 10 units. Assuming that the exposure-outcome association is either constant over time (left) or time-varying (right), posterior means (black solid lines) and corresponding 95% credible intervals (shaded grey regions) are displayed against time instants. Red dotted lines represent the "true" simulated relationships.

4 Conclusion

A DGLM is proposed in the present article for assessing the impact of air pollution exposure on a health outcome. Our proposal results in a flexible model since it can straightforwardly capture both spatial and temporal heterogeneity. A simulation study has demonstrated that this model is capable of accurately recovering the main relationship. Future improvements might include extending the model to take into consideration any non-linear relationships between exposure and linear predictor: this topic will be covered in greater detail in an expanded edition of the paper, along with a more thorough simulation analysis and a real-world application.

References

1. Banerjee, S., Carlin, B. P., Gelfand, A. E.: Hierarchical modeling and analysis for spatial data. CRC press (2014)
2. Bitto, A., Frühwirth-Schnatter, S.: Achieving shrinkage in a time-varying parameter model framework. J. of Econometrics 210(1), 75–97 (2019)
3. Copernicus Atmosphere Monitoring Service: CAMS Global Atmospheric Composition Forecasts. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on April 12, 2023) `https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-atmospheric-composition-forecasts?tab=overview`
4. Dominici, F., Peng, R. D.: Statistical methods for environmental epidemiology with R: a case study in air pollution and health. Springer (2008)
5. Dominici, F., Peng, R. D., Zeger, S. L., White, R. H., Samet, J. M.: Particulate air pollution and mortality in the United States: did the risks change from 1987 to 2000? Am. J. of Epidemiology 166(8), 880–888 (2007)
6. Fan, J., Zhang, W.: Statistical methods with varying coefficient models. Statistics and its Interface 1(1), 179–195 (2008)
7. Hastie, T., Tibshirani, R.: Varying-coefficient models. J. of the Royal Stat. Soc.: Ser. B (Methodological) 55(4), 757–779 (1993)
8. Janes, H., Dominici, F., Zeger, S. L.: Trends in air pollution and mortality: an approach to the assessment of unmeasured confounding. Epidemiology 18(4), 416–423 (2007)
9. Klompmaker, J. O., Laden, F., James, P., Sabath, M. B., Wu, X., Schwartz, J., Dominici, F., Zanobetti, A., Hart, J. E.: Effects of long-term average temperature on cardiovascular disease hospitalizations in an American elderly population. Environ. Res. 216, 114684 (2023)
10. Peng, R. D., Dominici, F., Louis, T. A.: Model choice in time series studies of air pollution and mortality. J. of the Royal Stat. Soc.: Ser. A (statistics in society) 169(2), 179–203 (2006)
11. Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L., Samet, J. M.: Seasonal analyses of air pollution and mortality in 100 US cities. Am. J. of Epidemiology 161(6), 585–594 (2005)
12. Prado, R., Ferreira, M. A., West, M.: Time series: modeling, computation, and inference. CRC Press (2021)
13. Qiu, X., Danesh-Yazdi, M., Wei, Y., Di, Q., Just, A., Zanobetti, A., Weisskopf, M., Dominici, F., Schwartz, J.: Associations of short-term exposure to air pollution and increased ambient temperature with psychiatric hospital admissions in older adults in the USA: a case–crossover study. The Lancet Planet. Health 6(4), e331–e341 (2022)
14. Sahu, S.: Bayesian modeling of spatio-temporal data with R. CRC Press (2022)
15. Wood, S. N.: Generalized additive models: an introduction with R. CRC press (2017)
16. Zhou X., Josey K., Kamareddine L., Caine M. C., Liu T., Mickley L. J., Cooper M., Dominici F.: Excess of COVID-19 cases and deaths due to fine particulate matter exposure during the 2020 wildfires in the United States. Sci. Adv. 7(33), eabi8789 (2021)

Mining social media data for damage assessment in environmental disasters

Analisi dei social media per la valutazione dei danni durante i disastri ambientali

Emiliano del Gobbo, Barbara Cafarelli, Luigi Ippoliti and Lara Fontanella

Abstract The increasing prevalence of social media usage has led to the emergence of mining social media data as a valuable resource for disaster response. Mining their textual data presents opportunities and challenges. Techniques leveraging natural language processing and machine learning extract relevant information while filtering out noise and misinformation. Real-world examples, such as Hurricane Harvey in 2017, demonstrate the value of social media in coordinating relief efforts. Challenges include unstructured and ambiguous data, diverse user credibility, and overwhelming data volume. The aim of this research is to develop a methodology that integrates textual classification of social media data, spatial analysis, and visual analytics to provide rapid responses during natural disasters.

Abstract *L'aumento dell'uso dei social media ha portato allo sviluppo del datamining dei social media come una preziosa risorsa per la gestione delle emergenze. L'analisi dei loro dati testuali presenta opportunità e sfide. Tecniche che sfruttano l'elaborazione del linguaggio naturale e l'apprendimento automatico estraggono informazioni rilevanti filtrando il rumore e la disinformazione. Esempi reali, come l'uragano Harvey nel 2017, dimostrano il valore dei social media nel coordinare gli sforzi di soccorso. Le sfide includono dati non strutturati e ambigui, la diversa credibilità degli utenti e un volume considerevole. Questa ricerca mira allo sviluppo*

Emiliano del Gobbo
Department of Economics, Management and Territory, University of Foggia, Via da Zara 11, Foggia, 71121, Italy, e-mail: emiliano.delgobbo@unifg.it

Barbara Cafarelli
Department of Economics, Management and Territory, University of Foggia, Via da Zara 11, Foggia, 71121, Italy, e-mail: barbara.cafarelli@unifg.it

Luigi Ippoliti
Department of Economics, University of Chieti-Pescara, Via dei Vestini 42, Pescara, 65127, Italy, e-mail: luigi.ippoliti@unich.it

Lara Fontanella
Department of Legal and Social Science, University of Chieti-Pescara, Via dei Vestini 42, Pescara, 65127, Italy, e-mail: lara.fontanella@unich.it

di metodologie che integrano la classificazione testuale dei dati dei social media e modelli spaziali per migliorare la gestione delle emergenze.

Key words: disaster response, social media, text mining, data mining

1 Introduction

The rise of social media platforms has transformed the way people communicate and interact with each other, leading to a massive increase in the amount of textual data generated every day. During times of crisis and natural disasters, social media can provide a valuable source of real-time information that can help disaster responders make informed decisions and allocate resources more effectively. However, the sheer volume and variety of data generated by social media users can make it difficult to extract relevant and actionable information.

Mining social media textual data for disaster response has emerged as a promising research area that aims to address these challenges. By leveraging advances in natural language processing, machine learning, and information retrieval, researchers and practitioners are developing techniques and tools to extract key information from social media textual data in real-time. These techniques can be used to identify critical information, such as the location of individuals in need of assistance, the severity and extent of the disaster, and the response efforts underway. Additionally, these techniques can be used to identify false information and misinformation, and to filter out irrelevant or redundant data.

Mining social media textual data for disaster response has already shown its value in several real-world scenarios [7]. For example, during the 2015 Nepal earthquake, social media played a crucial role in providing support and assistance [1, 5]. According to various reports and studies, platforms like Twitter and Facebook were used extensively by affected individuals, local communities, and relief organizations to communicate and coordinate relief efforts. Similarly, during Hurricane Harvey in 2017, social media users provided real-time updates on the location of stranded individuals, the severity of the flooding, and the availability of shelter and supplies.

However, mining social media textual data for disaster response poses several challenges. Firstly, the data is often unstructured, noisy, and ambiguous, making it challenging to extract relevant and actionable information. Secondly, the data is generated by a diverse range of users with varying levels of expertise, credibility, and bias, leading to issues of reliability and accuracy. Thirdly, the volume of social media textual data generated during a disaster can be overwhelming, making it challenging to categorize crisis-related messages during the sudden-onset of natural or man-made disasters. Finally, extracting the location information and estimating the event location is also a difficult task to maintain satisfactory situation awareness [6].

In this work, we focus on the problem of damage assessment during disasters, which is a critical task for emergency responders and government agencies to determine the severity and scope of the disaster and to allocate resources effectively.

In particular, we employ state-of-the-art machine learning techniques to perform an experimentation of damage assessment using tweets related to the hurricane Ida [3].

2 The hurrican Ida dataset

Twitter has been extensively used as an active communication channel, especially during mass convergence events such as natural disasters like earthquakes, floods and typhoons [7, 1, 5]. In this work, the dataset of interest consists of tweets posted during the Hurricane Ida [3], which was a deadly and destructive Category 4 Atlantic hurricane that became the second-most damaging and intense hurricane to make landfall in the U.S. state of Louisiana on record, behind Hurricane Katrina in 2005. In particular, hurricane Ida formed August 26, 2021 and dissipated September 4th, 2021.

The dataset includes over 1,800,000 tweets (all in English) selected by searching for the keywords *#ida*, *'hurricane ida'* and *hurricaneIda*. The distribution over time of the tweets is shown in Fig. 1 which highlights the days when the hurricane peeked its severity. During the onset of a crisis, a variety of information is posted in real-time by affected people; by people who are in need of help or by people who are willing to offer volunteering services. Hence, these messages contain significant actionable and tactical information. To give a flavour of the content of the tweets, Fig. 2 shows a word cloud with the most frequent words posted during the event.

At the onset of the event, assessment of the level of damage is one of the key situational awareness requirements of humanitarian response organizations to un-

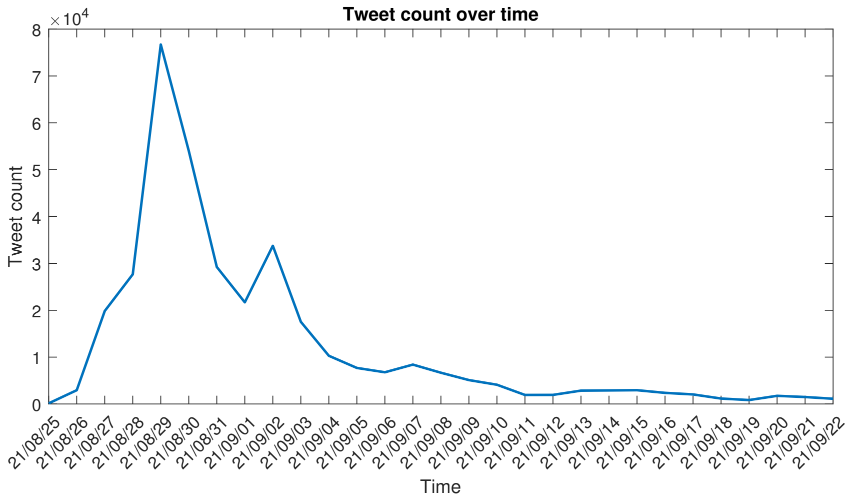


Fig. 1 Distribution of tweets over time from Twitter dataset Hurricane Ida [3]

derstand the severity of destruction and to plan relief efforts accordingly. To this purpose, Natural Language Processing (NLP) techniques, such as automatic summarization, information classification, named-entity recognition, information extraction, have become increasingly important in processing and analyzing social media messages. NLP techniques can help to quickly and efficiently process and analyze large volumes of social media messages during disasters, providing valuable insights for damage assessment.

3 Methodology and data analysis

To classify the damages and required interventions in the Hurricane Ida dataset, we leverage the information and labels available in IDRISI-RE, which contains a collection of tweets related to major disasters such as hurricanes, earthquakes, and floods. In particular, the involvement of state-of-art NLP model, such as Bidirectional Encoder Representations from Transformers (BERT)[2, 4], will be used to classify tweets by their type of support request. BERT is specifically designed to generate deep bidirectional representations through pre-training on unlabeled text, considering both left and right context in all layers [2]. The pre-training phase of BERT involves training the transformer network concurrently on two tasks: (i) predicting the masked words in a sentence (Masked Model language), and (ii) predicting the subsequent sentence (Next Sentence Prediction). A comprehensive explanation of

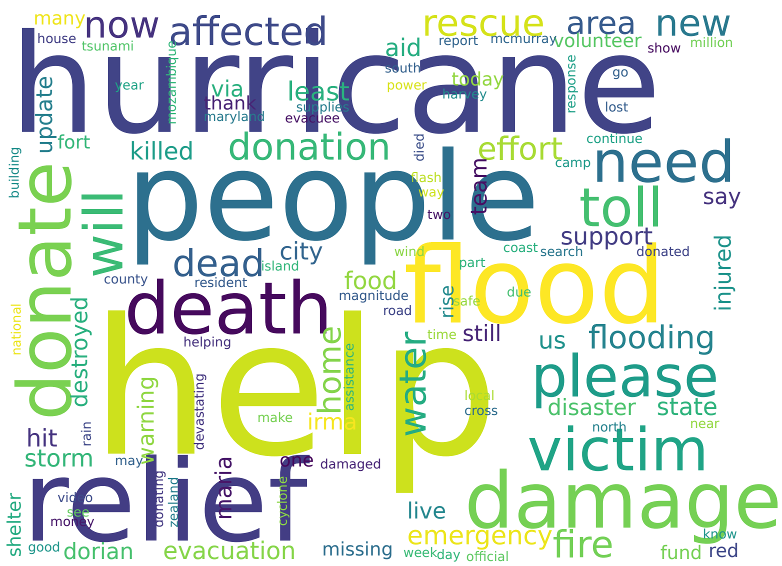


Fig. 2 Word cloud of the top 200 words in emergency tweets of hurrican datasets

the BERT implementation can be found in Devlin et al. (2018) [2]. Once the pre-trained BERT model is obtained, it can be fine-tuned by incorporating additional output layers, enabling the creation of models for various tasks, such as question answering and language inference, without significant alterations to the task-specific architecture (Devlin et al., 2018). In this case, fine-tuning processing will be performed using the IDRISI-RE dataset using the tweets related to hurricane events, to build a model to be used for predicting the class of tweets on new hurricane events. This methodology will be experimented on the hurricane Ida dataset and results discussed in an extended version of this work.

References

1. Bossu R., Laurin M., Mazet-Roux G., Roussel F. and Steed R.: The importance of smartphones as public earthquake-information tools and tools for the rapid engagement with eyewitnesses: A case study of the 2015 nepal earthquake sequence. Seismological Research Letters, 86:1587–1592 (2015)
2. Devlin J., Chang M.W., Lee K. and Toutanova K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
3. Mark E.P.: Hurricane ida twitter dataset (2021) https://digital.library.unt.edu/ark:/67531/metadc1913080/, accessed May 16, 2023), University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu
4. Reimers N. and Gurevych I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
5. Subba R.: Online convergence behavior, social media communications and crisis response: An empirical study of the 2015 Nepal earthquake police Twitter project (2016)
6. Suwaileh R., Elsayed T. and Imran M.: IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. Information Processing & Management, 60(3):103340 (2023)
7. Young C., Kuligowski E. and Pradhan A.: A review of social media use during disaster response and recovery phases (2020)

Solicited Session SS24 - *Satisfaction and behavior in tourism*

Organizer and Chair: Antonio Lucadamo

1. *The evaluation of the hotel stay through a new development of correspondence analysis coping with ordinal variables* (D'Ambra A. and Amenta P.)
2. *Assessing the role of knowledge and authenticity in the formation of attendee loyalty at cultural festivals* (Rivetti F., Lucadamo A. and Rossi C.)
3. *Residents' Opinions and Perceptions of Tourism Development in the Historic City of Matera* (Sarnacchiaro P., Di Gennaro R. and Di Taranto E.)
4. *Exploring tourism at religious sites: The case of Assisi* (Rivetti F., Dini M. and Splendiani S.)

The evaluation of the hotel stay through a new development of correspondence analysis coping with ordinal variables

La valutazione del soggiorno alberghiero tramite una estensione dell'analisi delle corrispondenze con variabili ordinali

Antonello D'Ambra and Pietro Amenta

Abstract Correspondence Analysis is a valuable data science visualization technique to find and display the relationships between two variables. When we cope with ordinal categorical variables, we cannot apply this analysis because Pearson's chi-square contingency coefficient does not take this characteristic into account. Taguchi [6, 7] introduced a statistic that considers the ordinal nature of a variable using the cumulative frequency of the table's cells across this variable. This index is at the base of several extensions of Correspondence Analysis (CA) that have been proposed in the literature, also in the presence of more than two variables. In this paper, we propose an enhancement of the inferential tools to highlight the easy applicability and graphical reading of the results of these approaches.
Abstract *L'analisi delle corrispondenze è un'utile tecnica grafica per visualizzare le relazioni tra due variabili. Questa analisi non può essere applicata in modo efficiente quando siamo in presenza di variabili categoriali ordinali. Infatti questa tipologia non viene considerata direttamente dal coefficiente chi-quadrato del Pearson. Taguchi [6, 7] ha introdotto un indice che considera la natura ordinale di una variabile utilizzando la frequenza cumulativa delle celle della tabella. Questo indice è alla base di diverse estensioni dell'Analisi delle Corrispondenze che sono state proposte in letteratura, anche in presenza di più di due variabili. In questo lavoro proproniamo uno sviluppo degli strumenti inferenziali per evidenziare la facile applicabilità e lettura grafica dei risultati forniti da questi approcci.*

Key words: Contingency table, Taguchi's index, Orthogonal decomposition, Inferential tools.

Antonello D'Ambra
Department of Economics, University of Campania "L.Vanvitelli", Italy,
e-mail: antonello.dambra@unicampania.it

Pietro Amenta
Department of Law, Economics, Management and Quantitative Methods.
University of Sannio, Italy,
e-mail: amenta@unisannio.it

1 Notation

Let A, B and Y be three categorical variables with $i = 1,\dots,I$, $k = 1,\dots,K$ and $j = 1,\dots,J$ categories, respectively. Suppose that at least one variable has an ordinal nature with increasing scores (say Y). Suppose $(A_1,B_1,Y_1),\dots,(A_n,B_n,Y_n)$ is a random sample of the random vector (A,B,Y). n is the known total number of observations. Let N_{ikj} be a random variable, which counts the number of observations that fall into the cross-category $i \times k \times j$. We interactively code the categorical variables A and B into a new variable AB that consists of all combinations of the I and K categories [2]. This new variable consists of IK categories, such that $N_{ik\bullet}$ and $N_{\bullet j}$ represent the counts for the joint categories (i,k) and j, respectively. Suppose also that we have the $(IK \times J)$ contingency table $\mathbf{N} = (n_{ikj})$ under the product-multinomial sampling model. Let Y be the response variable which depends on the two predictor variables A and B. Let $\mathbf{n}_{ik}$ be the ik-th row of $\mathbf{N}$. We denote by p_{ikj} the probability of having an observation fall in the ik-th row and j-th column of the table $\mathbf{P} = (p_{ikj})$ of order $IK \times J$. Therefore, let $\mathbf{D}_{IK}$ and $\mathbf{D}_J$ represent the diagonal matrices of row and column marginal probabilities p_{ik+} and $p_{\bullet j}$, respectively, where $p_{ik\bullet} = \sum_{j=1}^{J} p_{ikj}$ and $p_{\bullet\bullet j} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ikj}$ denote the probabilities that A, B and Y are in categories i, k and j, respectively. We indicate with F_s the cumulative distribution of Y evaluated in s; that is, $F_Y(s) = Pr(Y \le s) = \sum_{j=1}^{s} p_{\bullet j} = p_{\bullet s}$. Finally, let $E_{iks} = \sum_{j=1}^{s} n_{ikj}$ and $E_{\bullet\bullet s} = \sum_{j=1}^{s} n_{\bullet\bullet j}$ be the cumulative count and the cumulative column total up to the s-th column category, respectively, with $s = 1,\dots,J-1$.

2 Cumulative Correspondence Analysis and Taguchi's Statistic for a matricisation of three ways contingency tables

D'Ambra and Amenta [1] developed a new CA extension that considers the decomposition of Taguchi's statistic facing to a matricisation of the three-way contingency table where one of the variables consists of ordered responses. The matricisation process [3] transforms a three-way array (of order $I \times K \times J$) into a supermatrix of order $IK \times J$, with mode I entities nested within mode K entities. They proposed a generalisation of Taguchi's index (called Multiple Taguchi) as a measure of the association between the row and ordered column variables:

$$T_M = \sum_{s=1}^{J-1} w_s \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ik\bullet}\left(\frac{E_{iks}}{n_{ik\bullet}} - d_s\right)^2 \tag{1}$$

where $d_s = E_{\bullet\bullet s}/n$ and $w_s = [d_s(1-d_s)]^{-1}$. It is possible to show that T_M statistic can be expressed as a sum of Pearson's chi-squared tests applied to the cumulative frequencies in the columns: $T_M = \sum_{s=1}^{J-1} X_s^2$, where X_s^2 is the Pearson chi-squared for the s-contingency sub-table, of size $(IK \times 2)$, which is obtained by aggregating the first s column categories and the remaining categories $(s+1)$ to J, respectively. Moreover, the authors pointed out that the T_M-statistic can be also computed as:

$$T_M = n \times ||\mathbf{B}||^2_{\mathbf{D}_{IK}} = n \times ||\mathbf{D}^{-1}_{IK}\mathbf{P}\mathbf{A}^T\mathbf{W}^{\frac{1}{2}}||^2_{\mathbf{D}_{IK}} = n \times tr\left(\mathbf{D}^{-\frac{1}{2}}_{IK}\mathbf{P}\mathbf{A}^T\mathbf{W}\mathbf{A}\mathbf{P}^T\mathbf{D}^{-\frac{1}{2}}_{IK}\right)$$

where $\mathbf{A} = \mathbf{L} - [\mathbf{D}(\mathbf{1}_{J-1}\mathbf{1}^T_J)]$, $\mathbf{L}$ is lower uni-triangular matrix of dimension $(J-1) \times J$, $\mathbf{D} = diag(d_s)$, $\mathbf{1}_{J-1}$ is an all-ones vector of dimension $(J-1)$ and $\mathbf{W}$ is a diagonal square matrix of dimension $[(J-1)\times(J-1)]$ with general term w_s . Moreover, this index can be expressed in terms of total inertia: $T_M/n = ||\mathbf{B}||^2_{D_{IK}} = trace(\mathbf{B}^T\mathbf{D}_{IK}\mathbf{B}) = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J} p_{ik\bullet}b^2_{ikj} = \sum_{m=1}^{M}\lambda^2_m$ where λ_m is the m-th generalised singular value of the Generalised Singular Value Decomposition of the matrix $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ with $\mathbf{U}^T\mathbf{D}_I\mathbf{U} = \mathbf{I}_M$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_M$, respectively.

Finally, the CA coordinates for the graphical representation of the row and column cumulated categories are then given by $\mathbf{F} = \mathbf{U}\mathbf{\Lambda}$ and $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}$, respectively.

Using the Satterthwaite's method [5], Nair [4] approximated the distribution of the Taguchi's index T. The distribution of T_M has been approximated following the same approach. Indeed, the limiting distribution of T_M is a linear combination of iid chi-squared random variables:

$$T_M \xrightarrow[H_0]{D} \sum_{s=1}^{J-1} \gamma_s \chi^2_{s,(IK-1)}$$

which can be approximated by $r_M(IK-1)\chi^2_{v_M}$ with

$$v_M = \frac{1}{r_M}\sum_{s=1}^{J-1}\gamma_s = \frac{(IK-1)(J-1)}{\omega}$$

degrees of freedom (df) where:

- $\chi^2_{s,(IK-1)}$ is the chi-squared distribution for the s-th components with $(IK-1)$ df;
- γ_s is the s-th eigenvalue of matrix $\mathbf{A}^T\mathbf{W}\mathbf{A}\mathbf{D}_J$;
- $r_M = (IK-1)^{-1}(\sum_{s=1}^{J-1}\gamma^2_s / \sum_{s=1}^{J-1}\gamma_s)$ and
- $\omega = (\sum_{s=1}^{J-1}\gamma^2_s)/(\sum_{s=1}^{J-1}\gamma_s)$.

Finally, the statistic $\tilde{T}_M$ is given by $\tilde{T}_M = T_M/r_M(IK-1)$ which can be approximated by a chi-squared random variable $\chi^2_{v_M}$ with v_M df.

Confidence circles identify which categories contribute to the hypothesis of independence. D'Ambra and Amenta [1] highlighted that the joint distribution between the first two dimensions for the ik-th row principal coordinate is given by $f^2_{ik1} + f^2_{ik2} \sim \frac{\omega}{N_{ik\bullet}}\chi^2_{\frac{2}{\omega}}$ leading to a circle of radii length $r_{ik} = \sqrt{\omega/N_{ik\bullet} \times \chi^2_{2/\omega}}$ with center equal to (f_{ik1}, f_{ik2}). They considered the confidence circles only for each joint category of the row predictor variables, highlighting which categories significantly influence the response categories. If the origin of the axes $(0,0)$ lies within a confidence circle of a predictor category, then this does not significantly influence the response categories.

The development of a suitable decomposition of the Multiple Taguchi's index allowed also to separate the main effects and interaction term with several advantages, as follows: to consider only statistically significant components, to examine the effect of each predictor variable on collapsed criterion variable, to investigate the impact of each category of significant predictor variable, and to detect if there are combined effects of criterion variables. See [1] for deeper theoretical aspects.

We highlight that confidence circles can be also applied for evaluating each effect of the T_M decomposition.

We point out that it is also possible to obtain several indices proposed in literature by using suitable values of w_s as well as to develop inferential tools for the aggregated column categories.

Finally, in an extended version of this paper, we will consider the stays evaluations of several hotels in Naples, highlighted by the Tripadvisor website, to highlight the easy applicability and graphic reading of the results of our approach.

References

1. D'Ambra, A., Amenta, P.: An extension of correspondence analysis based on the multiple Taguchi's index to evaluate the relationships between three categorical variables graphically: an application to the Italian football championship. Ann. Oper. Res. (2022) doi: 10.1007/s10479-022-04803-3
2. Greenacre, M.: Correspondence analysis in practice. Chapman & Hall/CRC (2017)
3. Kiers, H.: Towards a standardized notation and terminology in multiway analysis, J. Chemom. 14, 105–122 (2000)
4. Nair, V.: Chi-squared type tests for ordered alternatives in contingency tables. J. Am. Stat. Assoc. 82, 283–291 (1987)
5. Satterthwaite, F.: An approximate distribution of estimates of variance components, Biom. J., 2: 110–114 (1946)
6. Taguchi, G.: Statistical analysis, Tokyo: Maruzen (1966)
7. Taguchi, G.: A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test, Saishin Igaku. 29: 806–813 (1974)

Assessing the role of knowledge and authenticity in the formation of attendee loyalty at cultural festivals

Valutazione del ruolo della conoscenza e dell'autenticità nella formazione della loyalty dei partecipanti ai festival culturali

Francesca Rivetti, Antonio Lucadamo and Carla Rossi

Abstract The study of knowledge in the formation of loyalty at cultural festivals has only recently attracted the attention of scholars. Much remains to be understood. This study develops a knowledge-based model of attendee loyalty in which the role of authenticity is assessed. It considers knowledge acquired during the festival and its role in loyalty formation, also in relation to authenticity. A Structural Equation Model is applied to a sample of visitors to a festival annually held in southern Italy. The results show that acquired knowledge has a positive impact on loyalty and that this relationship is mediated by authenticity.
Abstract *Lo studio della conoscenza nella formazione della loyalty nell'ambito dei festival culturali ha attirato solo di recente l'attenzione degli studiosi. Al riguardo, molto resta da capire. Questo studio sviluppa un modello di attendee loyalty basato sulla conoscenza, in cui viene valutato il ruolo dell'autenticità. Viene considerata la conoscenza acquisita durante il festival e il suo ruolo nella formazione della loyalty, anche in relazione all'autenticità. Un modello di equazioni strutturali viene applicato a un campione di visitatori di un festival che si tiene annualmente nel Sud Italia. I*

Francesca Rivetti
University of Basilicata; e-mail: francesca.rivetti@unibas.it

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

Carla Rossi
University of Basilicata; e-mail: carla.rossi@unibas.it

risultati mostrano che la conoscenza acquisita ha un impatto positivo sulla loyalty e che questa relazione è mediata dall'autenticità.

Key words: acquired knowledge, authenticity, loyalty, cultural festival, attendees

1 Introduction

In recent years, tourism scholars have begun to examine the role of knowledge in relation to loyalty. Acquired knowledge has been considered as an antecedent of attendee behaviour, with specific reference to sporting events [16] and cultural festivals [11]. In the latter case, the role of knowledge acquisition in the formation of behavioural intentions is considered particularly important, because learning is "one of the main outcomes" of this type of events [11]. Therefore, acquired knowledge is expected to impact directly on loyalty and the following hypothesis is formulated:
H1: Acquired knowledge positively influences behavioural loyalty.
In addition to loyalty, scholars have identified other consequences of knowledge. One construct that has been extensively researched in tourism studies, but has been rarely associated with acquired knowledge, is authenticity. In fact, although scholars have linked learning to authenticity [12], the relationship between the two constructs has not yet been sufficiently explored in tourism and events in particular. We believe that in the context of festivals, satisfying the knowledge needs of participants would increase their perception of authenticity. Therefore, the following hypothesis is formulated:
H2: Acquired knowledge positively influences authenticity.
In addition, many studies have investigated the existence of a direct relationship between authenticity and loyalty. Kolar and Zabkar [5] noted the "importance and centrality of authentic experiences for understanding the loyalty of cultural tourists". Novello and Fernandez [6] hypothesised a positive effect of authenticity on loyalty, but found no confirmation in their empirical study of a religious event. Ramkisson and Uysal [9] found a positive relationship between perceived authenticity and cultural behavioural intentions. Shen [15] highlighted that festival authenticity does not predict re-visit intention, contrary to what Robinson and Clifford [13] found. Instead, Casteran and Roederer [2] suggested that authenticity affects behavioural loyalty. Akhoondnejad [1] showed that festival authenticity has no effect on attendee loyalty, while it has an impact on satisfaction, which in turn affects loyalty. This conflicting evidence highlights the need to further investigate whether festival loyalty can be considered a direct consequence of authenticity. In line with some of the literature mentioned above, we believe that perceived authenticity could be a predictor of loyalty in cultural festivals and formulate the following hypothesis:
H3: Perceived authenticity positively influences behavioural loyalty.
Beyond the direct relationships mentioned so far, we believe that authenticity is more than a simple consequence of acquired knowledge and an antecedent of behavioural loyalty. In particular, it can help explain the relationship between acquired knowledge and loyalty, by acting as a mediator in the model. This means that perceived

authenticity can contribute to explain why acquired knowledge leads to increased behavioural loyalty.
H4: Perceived authenticity mediates the relationship between acquired knowledge and behavioural loyalty.

2 Methodology

The study was developed with reference to a convenience sample of participants in a cultural festival, called "Sponz Fest", organised annually in Calitri, a town located in southern Italy.
The items used in the questionnaire were derived from the literature. With regard to perceived authenticity, we referred to the studies of Kolar and Zabkar [5], Ram et al. [8], Akhoondnejad [1], Girish and Chen [4]. Acquired knowledge was developed on the basis of the study by Rivetti and Lucadamo [11], based on the items used by Yeh and Lin [17], Coetzee et al. [3], and Oh et al. [7]. Finally, loyalty was based on the items proposed by Girish and Chen [4] and Akhoondnejad [1], and used by Rivetti and Lucadamo [11].
In order to test the hypothesised relationships, a covariance-based structural equation model (SEM) was developed by using the Lavaan package [14] in the "R" software [10].

3 Results and discussion

Visitors to the Sponz Fest 2022 were the population of this study. The research was developed through convenience sampling, and more than 300 visitors participated in the survey.
In the first step of the analysis, the goal was to test the first hypothesis (H1), which states that acquired knowledge positively influences loyalty. Before looking at the results of the SEM, it is important to consider the reliability measures (Table 1).

Table 1 Reliability measures for the initial model

Construct	***CR***	***AVE***	***Alpha***
Acquired Knowledge	0.900	0.752	0.893
Loyalty	0.958	0.884	0.955

The two constructs meet the reliability requirements. The values of CR and Cronbach's alpha are, in fact, always higher than 0.7 – the threshold generally accepted – and the AVE value is higher than 0.5; therefore, convergent validity is acceptable.

Table 2 Structural model results for the initial model

Hypothesis	*Dependent Variable*	*Independent Variable*	*Estimates*	*Stand. estimates*	*St. error*	*P-value*
H1	Loyalty	Acquired Knowledge	0.541	0.734	0.030	0.000

In Table 2 we can see that the first hypothesis is validated, the acquired knowledge has a significant positive impact on loyalty.
The second step is to assess whether there is a mediation by authenticity. Therefore, in the second analysis it is necessary to verify that acquired knowledge has a positive effect on authenticity (H2), which in turn should have a positive influence on loyalty (H3). Table 3 shows that the reliability measures have not significant changes for the two constructs already considered, and they also assume reliable values for authenticity.

Table 3 Reliability measures for the final model

Construct	*CR*	*AVE*	*Alpha*
Acquired Knowledge	0.887	0.726	0.893
Authenticity	0.947	0.817	0.946
Loyalty	0.958	0.884	0.955

In Table 4 we can see that the relationship between acquired knowledge and loyalty is still significant, but with a slightly higher p-value than before and, more importantly, with lower estimated coefficients (standardised and unstandardised) than in the initial model.

Table 4 Structural model results for the final model

Hypothesis	*Dependent Variable*	*Independent Variable*	*Estimates*	*Stand. estimates*	*St. error*	*P-value*
H1	Loyalty	Acquired Knowledge	0.161	0.223	0.093	0.017
H2	Authenticity	Acquired Knowledge	0.692	0.849	0.021	0.000
H3	Loyalty	Authenticity	0.482	0.545	0.089	0.000

The influence of the dependent variable on the mediator is also significant and positive, as is the relationship between the mediating construct and loyalty. This means that there is a partial mediating effect. Table 5 allows for a better understanding of all the links, showing the important role of authenticity in the relationship between acquired knowledge and loyalty.

Table 5 Mediating effects

Effect	*Estimates*	*St. estimates*	*St. error*	*P-value*
Direct Effect	0.161	0.223	0.093	0.017
Indirect Effect	0.334	0.463	0.077	0.000

Total Effect	0.495	0.686	0.034	0.000

The study contributes to the understanding of the antecedents of behavioural loyalty in the context of cultural festivals. In particular, it focuses on the impact of knowledge acquired during festivals and the role of perceived authenticity. In addition to the theoretical implications, important managerial implications are found. In this respect, it is emphasised that, in order to increase attendee loyalty, organisers of this type of festival should focus as much as possible on cultural content and promote knowledge transfer through the various information channels used by the festival. As learning should not only take place in a passive form (involving listening, observing, and memorising), but above all in an active form that requires participation, critical thinking and problem solving, event organisers should make every effort to provide visitors with a meaningful context and different kinds of multi-sensory stimulation during their visit to the festival in order to enhance participants' involvement and their learning experience during the cultural event.

They should also create events within the festival that are not produced solely for tourist consumption, but are authentically rooted in local history, culture, and traditions, and are therefore more relevant to local people than to tourists. In other words, they should be aware of the commoditising effects of tourism, which, if poorly managed, tends to reduce local culture to a series of inauthentic events.

References

1. Akhoondnejad, A.: Tourist loyalty to a local cultural event: The case of Turkmen handicrafts festival. Tourism Management, 52, 468-477 (2016)
2. Casteran, H., Roederer, C.: Does authenticity really affect behavior? The case of the Strasbourg Christmas Market, Tourism Management, 36, 153-163 (2013)
3. Coetzee, W.J., Lee, C., Faisal, A.: Predicting intentions to revisit and recommend a sporting event using the event experience scale (EES). Event Management, 23(3), 303-314 (2019)
4. Girish, V.G., Chen, C.F.: Authenticity, experience, and loyalty in the festival context: Evidence from the San Fermin festival, Spain. Current Issues in Tourism, 20(15), 1551-1556 (2017)
5. Kolar, T., Zabkar, V.: A consumer-based model of authenticity: An oxymoron or the foundation of cultural heritage marketing? Tourism Management, 31(5), 652-664 (2010)
6. Novello, S., Fernandez, P.M.: The influence of event authenticity and quality attributes on behavioral intentions. Journal of Hospitality & Tourism Research, 40(6), 685-714 (2016)
7. Oh, H., Fiore, A.M., Jeoung, M.: Measuring experience economy concepts: Tourism application. Journal of Travel Research, 46(2), 119-132 (2007)
8. Ram, Y., Björk, P., Weidenfeld, A.: Authenticity and place attachment of major visitor attractions. Tourism Management, 52, 110-122 (2016)
9. Ramkissoon, H., Uysal, M. S.: The effects of perceived authenticity, information search behaviour, motivation and destination imagery on cultural behavioural intentions of tourists. Current Issues in Tourism, 14(6), 537-562 (2011)
10. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021)
11. Rivetti, F., Lucadamo A.: Cultural festival attendees: a path from motivation to loyalty. Current issues in tourism, 1-17 (2022)
12. Robertson M., Junek, O., Lockstone-Binney, L.: Is this for real? Authentic learning for the challenging events environment. Journal of Teaching in Travel & Tourism, 12(3), 225-241(2012)
13. Robinson, R.N., Clifford, C.: Authenticity and festival foodservice experiences. Annals of Tourism Research, 39(2), 571-600 (2012)

14. Rosseel, Y.: Lavaan: An R package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36 (2012)
15. Shen, S.: Intention to revisit traditional folk events: A case study of Qinhuai Lantern Festival, China. International Journal of Tourism Research, 16(5), 513-520 (2014)
16. Yazici, T., Kocak, S., Altunsöz, I.H.: Examining the effect of experiential marketing on behavioral intentions in a festival with a specific sport event. European Sport Management Quarterly, 17(2), 171-192 (2017)
17. Yeh, H.R., Lin, L.Z.: Exploring tourists' nostalgic experiences during culture festivals: the case of the Sung Chiang Battle Array. Current Issues in Tourism, 20(4), 391-424 (2017)

Residents’ Opinions and Perceptions of Tourism Development in the Historic City of Matera

Opinioni e percezioni dei residenti sullo sviluppo del turismo nella città storica di Matera

Pasquale Sarnacchiaro, Roberta Di Gennaro and Enrico Di Taranto

Abstract The paper aims to understand the residents’ perceptions of tourism impact in destination development process. In particular, the development of tourism in Matera should affect the daily life of residents and affect the perception of residents of the impacts of tourism in their communities. The growing number of urban tourists increases the use of natural resources, origins socio-cultural impact, and causes stress on infrastructure, mobility and other facilities. For these reasons, the objectives of our study are to explore the perceptions of the residents of Matera about the impacts of tourism. This paper reports a study, based on Structural Equation Model, about residents’ perception on tourism impact. The data are collected with a self-administered questionnaire on a sample of 250 residents.

Abstract *La ricerca mira ad analizzare le percezioni dei residenti sull’impatto del turismo. In particolare, lo sviluppo del turismo a Matera ha toccato la vita quotidiana dei residenti e il numero crescente di turisti urbani aumenta l'uso delle risorse naturali, l'impatto socio-culturale delle origini e provoca stress su infrastrutture,*

Pasquale Sarnacchiaro

Università of Naples Federico II – Department Economia, Management e Istituzioni; e-mail: sarnacch@unina.it

Roberta Di Gennaro

Università of Naples Parthenope – Department of Engeneering; e-mail: roberta.digennaro@uniparthenope.it

Enrico Di Taranto

Università of Naples Federico II – Department Economia, Management e Istituzioni; e-mail: ditarant@unina.it

mobilità e altre strutture. A tal fine l'obiettivo del nostro studio è di misurare le percezioni dei residenti di Matera riguardo l'impatto del turismo. Partendo dalla somministrazione di un questionario ad un campione di 250 residenti, lo studio, attraverso la costruzione di un modello ad equazioni strutturali, mira ad identificare i fattori che maggiormente influenzano l'impatto del turismo sulla vita della popolazione residente.

Key words: Tourism growth, Overtourism, Structural Equation Model, Partial Least Squares, Tourism policy

1 Introduction

Tourism development stimulates communities growth, inspires business and investment, improves government services for the public sector, improves countries' images amongst other countries and supports international tolerance and understanding among people around the world. Many countries consider tourism a tool for economic development and an engine of economic diversification because of various benefits that tourism offers to a community. Interest in residents' perceptions of the impact of tourism is began in various parts of world, primarily as case studies in the USA [2, 12] and Japan [17, 22]. Most of these studies categorize tourism impacts into social, cultural, environmental and economic impacts [1, 20]. The terms social and cultural impacts of tourism are used in research interchangeably, measured using a single variable called socio-cultural impacts [8, 13].
The fundamental hypothesis is that residents' perceptions of the impacts of tourism influence their support for tourism development [2, 15, 18]. Residents are primary stakeholders during sustainable development of tourism [16]; they are part of the tourism industry in their communities because subjects to be observed and interacting with tourists. Their perceptions of tourism's impacts influence hospitality towards tourists. Understanding residents' perceptions of the impacts of tourism offers important information to government officials, tourism developers and policymakers when developing tourism strategies and services [5].

2 Literature review

Most studies investigating residents' perceptions of tourism's impact focus on perceptions of local or national residents [6, 10, 22]. Doxey [7] associates tourism development with the feeling of locals evolving from euphoria and apathy during the beginning of the life cycle, to annoyance and antagonism in the latter phases. Starting from the criticisms and extensions of such models, the key hypothesis of some scholars is that tourism development, measured by increasing tourists' number is associated with locals' deteriorating sentiment and corresponding reactions to tourism. On this premise, it follows logically that 'over-tourism' and its presence in the media is the expected result of some destinations reaching the last two stages of their Tourist Area Life-Cycle'(TALC). Tourist numbers exceed the carrying capacity of the destination and its infrastructure, while the locals feel overrun and displaced. The term 'over-tourism' has been defined by Richardson [19] as any destination that suffers from the tension of tourism. The term is also to be linked to what is more generally known as the carrying capacity, that is to say the maximum limit to tourism development [4]. In academia, over-tourism has become commonplace overnight, too. Whereas it was largely non-existent prior to 2017 [14]. The marketability and popularity of the term over-tourism seems to be at least partially responsible for its entry in academia, rather than its explanatory value, as exemplified by a recent paper using the term overtourism in its title, but does not mention it at all in the main text at all [9]. We want to assess the residents' perception of tourism growth to understand if in Matera, elected European Capital of Culture in 2019, we can already talk about over-tourism.

3 Research Method

The empirical phase of the research was conducted through a survey on residents of the Matera destination.
Starting from Ap & Crompton [3], scale consists of a belief component asking respondents to assess the level of change, associated with 35 items, and an evaluative component asking residents to rate their level of like or dislike for each item. We have verified social-cultural, economic, and environmental domains. Data were collected using a self-administered questionnaire, conducted during May and July 2019, in the city of Matera. The survey consisted of two sections, the first capturing socio-demographics features and the second one the perceptions of tourism's impact. The survey has three independent latent constructs that examined respondents' perceptions of tourism's impacts, and one dependent. All four latent variables included several elements, derived from extant studies of residents' perceptions of tourism's social, cultural, environmental and economic impacts [6, 10, 12, 15]. Due to the variety of attributes available in the literature, selection was based on common items applicable to all destinations, such as improvements to infrastructures and services. Similarly, attributes inappropriate to the context of Matera were excluded, such as nightlife. Items were measured using a five-point Likert-type scale anchored by strongly disagree and strongly agree, with 3 as a neutral point, to capture residents' perceptions of the four domains of tourism impacts. In total, 250 usable questionnaires were received and analyzed. To investigate the phenomena, we propose a Structural Equation Model in which the variable "Tourism impact" depends by some factors as: "Economic Impact", "Socio-cultural Impact", "Enviromental Impact" that influence the overall experience of residents, crucial for public administrators in their management choices.
In the social sciences, SEM has become a useful research technique for analysing the relationships between latent variables. There are two approaches to estimating the relationships in a SEM: the covariance-based SEM [11] and the variance-based method [21], known as Partial Least Squares Path Modeling (PLS-PM). In this paper,

we chose the PLS-PM, performed by Smart-PLS (Version 4), because it has less stringent assumptions for the distribution of variables and error terms [21].
PLS-PM statistical properties in particular provide robust estimations when the data presents normal and very extreme non-normal distributions (skewness and/or kurtosis) and can work with both reflective and formative measurement models. PLS-PM is formally defined by two sets of linear equations called inner (or structural) and outer (or measurement) models, respectively. The structural model specifies the relationships between latent variables (LVs), while the measurement model specifies the relationships between a LV and its manifest variables (MVs). PLS-PM includes two different types of measurement models, defined as reflective and formative measurement models. A PLS-PM is analyzed and interpreted in two stages: (1) evaluate the measurement model; (2) assessing the structural model.

4 Findings

From a theoretical point of view, the paper is placed within the flow of literature on the impacts generated by tourism and tries to understand how the tourism destination development affects the liveability of residents. From a managerial point of view, these findings suggest useful guidelines to policy makers and destination managers to implement a tourism development process, be managed so as not to generate negative impacts on the resident population and thus be as sustainable as possible from an economic, environmental and social point of view. A SEM was elaborated to examine residents' perception, because in social sciences, it is a research technique useful for analyzing the relationships between latent variables.

References

1. Almeida-García, F., Peláez-Fernández, M.Á., Balbuena-Vázquez, A., Cortés-Macias, R.: "Residents' perceptions of tourism development in Benalmádena (Spain)", Tourism Management, 54, 259-74 (2016)
2. Andereck, K.L., Valentine, K.M., Vogt, C.A., Knopf, R.C.: A cross-cultural analysis of tourism and quality of life perceptions. Journal of Sustainable Tourism, 15(5), 483-502 (2007)
3. Ap, J., Crompton, J.L.: Developing and testing a tourism impact scale. Journal of travel research, 37(2), 120-130 (1998)
4. Borg, J.V.d., Costa, P., Gotti, G.: Tourism in European heritage cities. Annals of Tourism Research 23, 306-321. http://dx.doi.org/10.1016/0160-7383(95)00065-8 (1996).
5. Brida, J.G., Pulina, M.: A literature review on the tourism-led-growth hypothesis (2010)
6. Brunt, P., Courtney, P.: Host perceptions of sociocultural impacts. Annals of tourism Research, 26(3), 493-515 (1999)
7. Doxey, G.V: A causation theory of visitor-resident irritants' methodology and research inferences. Proceedings of the Sixth Annual Conference of the Travel Research Association, 195--198, San Diego CA: Travel and Tourism Research Association (1975)
8. Fredline, L., Jago, L., Deery, M: The Development of a Generic Scale to Measure the Social Impacts of Events. Event Management 8, 23-37 (2003)

9. Muler Gonzalez, V., Coromina, L., Gali, N.: Overtourism: residents' perceptions of tourism impact as an indicator of resident social carrying capacity-case study of a Spanish heritage town. Tourism Review, 73(3), 277-296 (2018)
10. Gursoy, D., Chi, C.G., Dyer, P.: Local's attitudes toward mass andalternative tourism: The case of Sunshine Coast, Australian Journal of Travel Research, 49(3), 381-394 (2010)
11. Jöreskog, K.G.: A general method for estimating a linear structural equation system. ETS Research Report Series (1970)
12. Kim, K., Uysal, M., Sirgy, M.J.: How does tourism in a community impact the quality of life of community residents? Tourism management, 36, 527-540 (2013)
13. King, B., Pizam, A., Milman, A.: Social impacts of tourism: Host perceptions. Annals of Tourism Research, 20, 650-665 (1993)
14. Koens, K., Postma, A., Papp, B.: Is Overtourism overused? Understanding the impact of tourism in a city context. Sustainability, 10(12) (2018)
15. Látková, P., Vogt, C.A.: Residents' attitudes toward existing and future tourism development in rural communities. Journal of Travel Research, 51(1), 50-67 (2012)
16. Lundberg, E.: The level of tourism development and resident attitudes: A comparative case study of coastal destinations. Scandinavian Journal of Hospitality and Tourism, 15(3), 266-294 (2015)
17. Miyakuni, K.: Residents' attitudes toward tourism, focusing on ecocecentric attitudes and perceptions of economic costs: The case of iriomote island, Japan. Michigan State University. Park, Recreation and Tourism Resources (2012)
18. Rasoolimanesh, S.M., Ringle, C.M., Jaafar, M., Ramayah, T.: Urban vs. rural destinations: Residents' perceptions, community participation and support for tourism development. Tourism Management, 60, 147-158 (2017)
19. Richardson, D.: Suffering the strain of tourism. Retrieved from: (TTG@wtm).Dod, J. (2017)
20. Rivera, M., Gregory, A., Cobos, L.: Mobile application for the timeshare industry: the influence of technology experience, usefulness, and attitude on behavioral intentions. Journal of Hospitality and Tourism Technology, 6(3), 242-257 (2015)
21. Wold, H.: Path models with latent variables: The NIPALS approach. In Quantitative sociology, 307-357 (1975)
22. Zamani-Farahani, H., Musa, G.: Residents' attitudes and perception towards tourism development: A case study of Masooleh, Iran. Tourism Management, 29(6), 1233-1236 (2008)

Exploring tourism at religious sites: The case of Assisi

Esplorare il turismo nei siti religiosi: Il caso di Assisi

Francesca Rivetti, Mauro Dini and Simone Splendiani

Abstract The research is aimed at exploring tourism at religious sites. In particular, it aims to investigate the impact of different types of visitor motivations on behavioral loyalty, distinguishing between excursionists and tourists. Considering a sample of about 590 individuals who visited Assisi, a structural equation model is constructed. The model is tested separately for the two groups of visitors. Results show that motivations impact differently on loyalty and important differences are found for the two groups.

Abstract *La ricerca è finalizzata ad esplorare il turismo nei siti religiosi. In particolare, si propone l'obiettivo di investigare l'impatto di diversi tipi di motivazioni sulla loyalty comportamentale dei visitors, distinguendo tra turisti ed escursionisti. Prende a riferimento un campione di circa 590 soggetti che hanno visitato Assisi, viene realizzato un modello di equazioni strutturali. Il modello è testato separatamente per due gruppi di visitors. I risultati mostrano che le motivazioni hanno un diverso impatto sulla loyalty e sono rilevate importanti differenze per i due gruppi.*

Key words: religious tourism, motivation, loyalty, visitors

Francesca Rivetti
University of Basilicata; e-mail: francesca.rivetti@unibas.it

Mauro Dini
University of Urbino; e-mail: mauro.dini@uniurb.it

Simone Splendiani
University of Perugia; e-mail: simone.splendiani@unipg.it

1 Introduction

Tourism to religious sites has attracted scholarly attention in relatively recent times. Especially in the last decade there has been a relevant growth in the number of contributions on the topic. Despite this, to date the phenomenon is still poorly explored and, under some respects, it remains unclear.

One of the aspects that scholars are focusing on concerns the motivations that can lead tourists to visit this type of destinations. In this regard, Kim et al. [9] highlighted that religious motivation certainly underlie visits to religious sites, but it could be accompanied by other motives. Scholars have focused primarily on the framing motifs related to religion and, more generally, spirituality, trying to understand how they can be framed and articulated. For example, Tsai [24] identified motives concerning religious zeal, spiritual comfort and the search for God's protection. Liro et al. [13], distinguishing between domestic and non-domestic tourists, considered three types of motivations, that is sightseeing, prayer and pilgrimage. Besides these motivations, scholars have been concerned to consider the set of additional reasons that may drive tourists to visit religious destinations. Among these, they found the use of leisure time [15], the desire to travel [24], socialization [1, 15, 21], escape from routine [1, 26], nature enjoyment [4, 5, 8], self-growth [5], learning and cultural exploration [5, 19, 25], education [19], recreation and shopping [14, 7], sports activities [4], the fulfilment a promise [4], the willingness to attend sacred events [6], and service quality [7].

Recently, scholars have begun to question the consequences of the different types of motivations characterizing religious tourists. In particular, motivations have been framed within models aimed at assessing the satisfaction and/or behavioral intentions of visitors. With reference to satisfaction, Chun et al. [5, p. 432] considered the impact of various motivations (concerning the willingness to be connected with the nature, self-growth, learning, and relaxation), distinguishing between domestic and non-domestic tourists. Instead, Verma and Sarangi [25] evaluated the connection with satisfaction of the acquisition of spiritual and religious knowledge, escapism, the search for new experiences, and the willingness to have fun with family. Hassan et al. [7] developed a model in which religious, social/cultural and shopping motivations are not only related with satisfaction, but also with loyalty; they found a particularly relevant role of religious motivations in determining both satisfaction and loyalty. Instead, Piramanayagam et al. [17] focused on the direct relationships between motivations and behavioral intentions, finding that service quality and religious belief positively impact intentions, while historical attractions and cultures do not exert a significant influence.

While progress has been made with recent research, much remains to be understood regarding the consequences of the different types of motivations that drive tourists to visit religious destinations. In this regard, some observations should be made. First of all, it should be emphasized the variety of motivations considered in the studies aimed at investigating their consequences, an aspect that leaves wide open spaces for research. Furthermore, it must be considered that religious visitors can be framed within a multiplicity of categories, not limited to domestic and non-domestic tourists. In this regard, it would be possible to distinguish between excursionists and tourists,

who could be moved by different motivations, and whose motivations could have a different impact on the formation of loyalty.
Stating on these premises, the aim of this paper is to investigate the relationships between various types of motivations and behavioral loyalty with reference to a religious destination, considering two categories of visitors: excursionists and tourists. With reference to motivations, the focus is on cultural, spiritual/religious and escape motives.

2 Methodology

The study is based on a convenience sample of about 590 visitors to Assisi, one of the best known and most visited religious destinations in the world. Data collection took place in 2022, through the administration of a structured questionnaire. The items to measure motivation were derived from various studies. More specifically, to measure cultural motivation, we used three items from Li and Cai [12], Yoon and Uysal [27], Rivetti and Lucadamo [20], Kolar and Zabkar [11], and Schofield and Thompson [23]. Spiritual/religious motivation was identified through three items from Albayrak et al. [2] and Tsai [24]. To measure escape motivation, three items proposed by Li and Cai [12] were employed. Loyalty was constructed considering four items from Patwardhan et al. [16], Al-Msallam [3], and Kim & Thapa [10].
To test the relationships between the types of motivations and loyalty, we developed a co-variance-based structural equation models by using the "R" software [18] with the Lavaan package [22]. This model was tested separately with reference to excursionists and tourists in the sample.

3 Results and discussion

The reference population consisted of visitors to Assisi. We adopted a convenience sampling method and collected about 590 questionnaires. With the objective to examine the relationships between visitor motivations and behavioral loyalty, detecting the differences between excursionists and tourists, we tested the model separately with reference to the two groups of visitors. We found that the three types of motivations contribute differently to the formation of behavioral loyalty at religious destinations. Moreover, the impact of motivations is different with reference to the specific types of visitors.
The study contributes to the investigation of the antecedents of religious destination loyalty, focusing on the role of motivation. Managerial implications are relevant, especially considering the opportunity to define specific actions aimed at attracting visitors driven by different motivations, which impact differently on their behavioral intentions.

References

1. Abbate, C. S., and Di Nuovo, S.: Motivation and personality traits for choosing religious tourism. A research on the case of Medjugorje. Current Issues in Tourism, 16(5), pp 501-506 (2013)
2. Albayrak, T., Herstein, R., Caber, M., Drori, N., Bideci, M., and Berger, R.: Exploring religious tourist experiences in Jerusalem: The intersection of Abrahamic religions. Tourism Management, 69, pp 285-296 (2018)
3. Al-Msallam, S.: The impact of tourists' emotions on satisfaction and destination loyalty–an integrative moderated mediation model: tourists' experience in Switzerland. Journal of Hospitality and Tourism Insights, 3(5), pp 509-528 (2020)
4. Amaro, S., Antunes, A., and Henriques, C.: A closer look at Santiago de Compostela's pilgrims through the lens of motivations. Tourism Management, 64, pp 271-280 (2018)
5. Chun, B., Roh, E. Y., Spralls, S. A., and Kim, Y.: Predictors of templestay satisfaction: A comparison between Korean and international participants. Leisure Sciences, 40(5), pp 423-441 (2018)
6. Dar, H.: Hindu religious motivations in Kashmir valley. International Journal of Religious Tourism and Pilgrimage, 8(3), pp 1-14 (2020)
7. Hassan, T., Carvache-Franco, M., Carvache-Franco, W., and Carvache-Franco, O.: Motivations as predictors of religious tourism: the Muslim pilgrimage to the city of Mecca. Journal of Cultural Heritage Management and Sustainable Development, forthcoming (2022)
8. Kainthola, S., Chowdhary, N., Kaurav, R. P. S., and Tiwari, P.: Motivations of urban millennials for spiritual travel in India. Tourism Recreation Research, pp 1-16, forthcoming (2021)
9. Kim, B., Kim, S., and King, B.: Religious tourism studies: evolution, progress, and future prospects. Tourism Recreation Research, 45(2), pp 185-203 (2020)
10. Kim, M., and Thapa, B.: Perceived value and flow experience: Application in a nature-based tourism context. Journal of Destination Marketing and Management, 8, pp 373-384 (2018)
11. Kolar, T., and Zabkar, V.: A consumer-based model of authenticity: An oxymoron or the foundation of cultural heritage marketing? Tourism Management, 31(5), pp 652-664 (2010)
12. Li, M., and Cai, L. A.: The effects of personal values on travel motivation and behavioral intention. Journal of Travel Research, 51(4), pp 473-487 (2012)
13. Liro, J., Sołjan, I., and Bilska-Wodecka, E.: Visitors' diversified motivations and behavior–the case of the pilgrimage center in Krakow (Poland). Journal of Tourism and Cultural Change, 16(4), pp 416-435 (2018)
14. Liro, J.: Visitors' motivations and behaviours at pilgrimage centres: Push and pull perspectives. Journal of Heritage Tourism, 16(1), pp 79-99 (2021)
15. Liutikas, D.: In search of miracles: pilgrimage to the miraculous places. Tourism Review, 70(3), pp 197-213 (2015)
16. Patwardhan, V., Ribeiro, M. A., Payini, V., Woosnam, K. M., Mallya, J., and Gopalakrishnan, P.: Visitors' place attachment and destination loyalty: Examining the roles of emotional solidarity and perceived safety. Journal of Travel Research, 59(1), pp 3-21 (2020)
17. Piramanayagam, S., Kumar, N., Mallya, J., and Anand, R.: Tourist's motivation and behavioural intention to visit a religious Buddhist site: a case study of Bodhgaya. International Journal of Religious Tourism and Pilgrimage, 8(8), 5 (2020)
18. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021)
19. Robina Ramírez, R., and Fernández Portillo, A.: What role does touristś educational motivation play in promoting religious tourism among travellers?. Annals of Leisure Research, 23(3), pp 407-428 (2020)
20. Rivetti, F., Lucadamo A.: Cultural festival attendees: a path from motivation to loyalty. Current issues in tourism, 1-17 (2022)
21. Rodrigues, S., and McIntosh, A.: Motivations, experiences and perceived impacts of visitation at a Catholic monastery in New Zealand. Journal of Heritage Tourism, 9(4), pp 271-284 (2014)

22. Rosseel, Y.: Lavaan: An R package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36 (2012)
23. Schofield, P., and Thompson, K.: Visitor motivation, satisfaction and behavioural intention: the 2005 Naadam Festival, Ulaanbaatar. International journal of tourism research, 9(5), pp 329-344 (2007)
24. Tsai, H. Y. M.: Exploring the motivation-based typology of religious tourists: A study of Welcome Royal Lord Festival. Journal of Destination Marketing and Management, 21, 100623 (2021)
25. Verma, M., and Sarangi, P.: Modeling attributes of religious tourism: A study of Kumbh Mela, India. In Journal of Convention and Event Tourism, 20(4), pp 296-324 (2019)
26. Wang, W., Chen, J. S., and Huang, K.: Religious tourist motivation in Buddhist Mountain: The case from China. Asia Pacific Journal of Tourism Research, 21(1), 57-72 (2016)
27. Yoon, Y., and Uysal, M.: An examination of the effects of motivation and satisfaction on destination loyalty: a structural model. Tourism Management, 26(1), 45-56 (2005)

Session of free contributes SFC1 - *Education and labour*

1. *Local concordance among the items of questionnaires on student's opinion (OPIS)* (Terzi S. and Petrarca F.)
2. *High School Proficiency of Future University Students: An Analysis based on INVALSI Data* (Santelli F., Di Credico G. and Di Caterina C.)
3. *Employment vulnerability of immigrants in the labour market – Does origin matter?* (Bittaye M.)
4. *The effect of pricing policies on students' use of university canteens* (Masserini L., Bini M. and Lorenzoni V.)
5. *Gig workers' identikit* (Zavarrone E. and Forciniti A.)

Local concordance among the items of questionnaires on student's opinion (OPIS)

Valutazione della concordanza locale tra le risposte al questionario sull'opinione degli studenti (OPIS)

Silvia Terzi and Francesca Petrarca

Abstract By means of a local concordance curve – a function designed to detect positive association amongst the single components of multivariate phenomena, in particular among the best performing units and vice-versa among the worst performing units - we investigate the association among the items of the questionnaire on Italian university student's opinion (OPIS). In particular we will examine whether - among the classes with at least one element of dissatisfaction - there is a concentration of organizational and teaching deficiencies or whether the lacking skills are equally distributed.

Abstract *Per mezzo della curva di concordanza locale - una funzione concepita per rilevare la presenza di associazione positiva tra le singole componenti di un fenomeno multivariato, in particolare tra le unità con le migliori prestazioni e viceversa tra le unità con le peggiori prestazioni - viene indagata l'associazione tra gli item del questionario sull'opinione degli studenti universitari italiani (OPIS). In particolare, esamineremo se - tra gli insegnamenti che presentano almeno un elemento di insoddisfazione - vi sia una concentrazione di carenze organizzative e didattiche o se le competenze che vengono ritenute carenti siano equamente distribuite.*

Key words: Local concordance, quality assessment and evaluation.

Silvia Terzi
University of Rome Tre, Via Silvio D'Amico 77, e-mail: silvia.terzi@uniroma3.it

Francesca Petrarca
University of Rome Tre, via del Castro Pretorio 20, e-mail: francesca.petrarca@uniroma3.it

1 Introduction

In the student's opinion survey (OPIS) concerning the degree of appreciation for the subjects in which they are about to take the examination, students are asked some questions, concerning three distinct macro-environments, relating to: (1) organization of the class, (2) teaching skills and openness of the lecturer, (3) overall satisfaction. For each question, the respondent must indicate her or his degree of satisfaction with item being asked, i.e. whether she or he is totally dissatisfied, more critical than appreciative, quite appreciative, or totally satisfied.

Then, at the end of the year, having collected all questionnaires, for each class and for each question, the four distinct percentages of totally dissatisfied, partially dissatisfied, partially satisfied or totally satisfied students are computed. These percentages can be aggregated into a composite indicator measuring overall satisfaction, overall dissatisfaction, or critical (or positive) aspects of one or more study courses.

Aggregating different dimensions, it would be extremely useful to embed in the performance indicator some information concerning the association between the single dimensions. Aim of this paper is to assess this issue, in particular to evaluate whether among the classes with at least one element of dissatisfaction - there is a concentration of organizational and teaching deficiencies or whether the lacking skills are equally distributed. In other words we want to investigate whether among the least appreciated classes, there is strong concordance among the unsatisfactory items. We could of course resort to Kendall's W, computing it on the basis of the % of totally unsatisfied with each item. However, it is interesting to explore whether – within the data set pertaining to the percentages of totally unsatisfied with each item - as the percentage of dissatisfaction decreases, the concordance varies.

In case of maximum local concordance, the most unappreciated classes will be the most unappreciated in all items; conversely, in the case of low or no concordance, the elements of highest dissatisfaction are distributed among classes.

We will make use of the local concordance curve [1] to explore these aspects.

2 Local concordance

Let X be the data matrix of n units and d variables: $X = (x_{ih})$, $i = 1,...,n$, $h = 1, ...,d$. In our case, the units are the classes that have recorded at least 1% of dissatisfaction in at least one item and the variables are the percentages of totally dissatisfied with the each of 12 selected items of the questionnaire; so that $n = number\ of\ classes$ and $d = 12$. Following Terzi and Moroni([1]), we rank all the observations within each of the d dimensions. Then we partition our ranked observations in contiguous subsets (slices) of fixed size s.[1] For each ranking we have a first, a second,..., a last slice; we call window the union of the d corresponding slices. To assess local concordance,

[1] For the sake of simplicity, we assume $s \leq n/d$.

we count how many units are ranked in each window. The smallest the number of units, the greatest the local agreement. Vice-versa maximum disagreement is when the units belonging to the $r-th$ slice of the $h-th$ distribution do not belong to the $r-th$ slice of any of the other $d-1$ distributions, so that the total number of units ranked in the window is maximum. Thus, if we count the number of units belonging to the $r-th$ window a minimum (s) is reached in case of maximum concordance and a maximum (sd) in case of maximum disagreement.

Following their same notation, let Cr be the number of units ranked in the $r-th$ window on at least one of the d rankings. A relative measure of concordance is defined as:

$$Kr = \frac{(sd - Cr)}{(sd–s)} \tag{1}$$

The local concordance curve is derived and plotted by computing the local concordance coefficient Kr for all the distinct slices of size s of the multivariate distribution.

The choice of s needs some clarification. In fact it affects not only the maximum and minimum values of Cr, the head count criterion, but also what is being defined as agreement. For $s \leq n/d$, $max\,Cr = sd$; for $s > n/d$, $max\,Cr = n$.[2] As for the notion of agreement, setting $s = 5$ means that the first 5 ranks are equivalent: if one unit i is assigned—respectively— ranks 1, 2, 3, 4, 5 on $d = 5$ distributions, we are stating that these rankings agree, coincide; when $s = 10$ we are stating equivalence among the first 10 ranks. Thus, local agreement is agreement among classes of ranks of size s. As a consequence, it is likely that for different values of s we will have different values not only of Cr but also of Kr, not directly comparable.

3 Application and results

We analysed the questionnaires pertaining to the Bachelor's degree classes of 3 Departments of Roma Tre University: Economics, Business Administration and Political Science.

Our 12 selected items are[3]:

- D1: background knowledge
- D2: adequacy of study load
- D3: adequacy of course material
- D4: clarification of examination modalities
- D5: adherence to class timetable
- D6: capability of stimulating interest

[2] So that in this case a relative measure of concordance, say $K'r = \frac{(n-Cr)}{(n-s)}$.

[3] The student's opinion questionnaire consists of 16 items out of which we selected 12. We chose to leave out four items: D8 concerning supplementary teaching activities, D12, D13 concerning - respectively - suitability of classrooms and structure, D16 concerning the capacity of the teaching material to adequately replace some classroom activities.

- D7: expository clarity
- D9: consistency with website
- D10: lecturer's approachability
- D11: regularity of the lecturer's attendance in class
- D14: interest for the topics covered
- D15: overall satisfaction

The variables are the % of totally unsatisfied with each of the 12 items. The $n = 98$ units are the number of classes with at least 1% of dissatisfaction in at least one of the items. First of all we computed the correlation matrix:

CORRELATION MATIRX												
D1	1,00											
D2	0,40	1,00										
D3	0,28	0,49	1,00									
D4	0,16	0,36	0,64	1,00								
D5	0,22	0,44	0,51	0,67	1,00							
D6	0,30	0,38	0,64	0,48	0,56	1,00						
D7	0,23	0,20	0,28	0,25	0,24	0,49	1,00					
D9	0,22	0,38	0,48	0,57	0,73	0,51	0,18	1,00				
D10	0,17	0,39	0,49	0,66	0,74	0,53	0,19	0,73	1,00			
D11	0,27	0,29	0,29	0,49	0,56	0,35	0,13	0,66	0,78	1,00		
D14	0,43	0,35	0,36	0,22	0,22	0,35	0,29	0,24	0,13	0,13	1,00	
D15	0,43	0,47	0,70	0,56	0,50	0,78	0,39	0,49	0,49	0,34	0,62	1,00

Fig. 1 Correlation Matrix

The pairwise correlation coefficients vary from a minimum of 0.13 and a maximum of 0.78. The maximum correlation is among items D10 and D11 (i.e. lecturer's approachability and regularity of attendance in class) and – maybe surprisingly - among items D7 and D15 (expository clarity and overall satisfaction); while the minimum is reached between D7 and D11 (expository clarity and regularity of attendance in class); and between D14 (interest for the topics covered) and, respectively D10 (lecturer's approachability) and D11(regularity of the lecturer's attendance). We computed Kendall's concordance coefficient W [2], obtaining $W = 0.45$. As for the concordance, setting $s = 5$, $s = 7$ and then $s = 10$, we can plot the three curves as in Figure 2. A first observation is that although the shape of the curves, are similar, the curve for $s = 7$ and $s = 10$ are shifted downwards. As we already stressed, the choice of s affects maximum and minimum values of Cr, the head count criterion, but it is also related to what is being defined as agreement. Consequently for different values of s, the values of Kr are not directly comparable, and for increasing values of s we do not know which curve will lie above the others, or if some curves intersect. For this particular data set, for a less restrictive definition of agreement, local concordance decreases.

Taking into account the correlation matrix, the value of W and the average level of the curves we can notice that the correlations and the concordance among the items is not very high. What is particularly interesting with respect to the focus of this paper is that the local concordance in the first and in the last windows is lower than the local concordance in the central windows. This means that where the dissatisfaction for one or more items is higher (the first windows), or lower (the last), it is less concentrated than where it is non at its maximum or minimum.

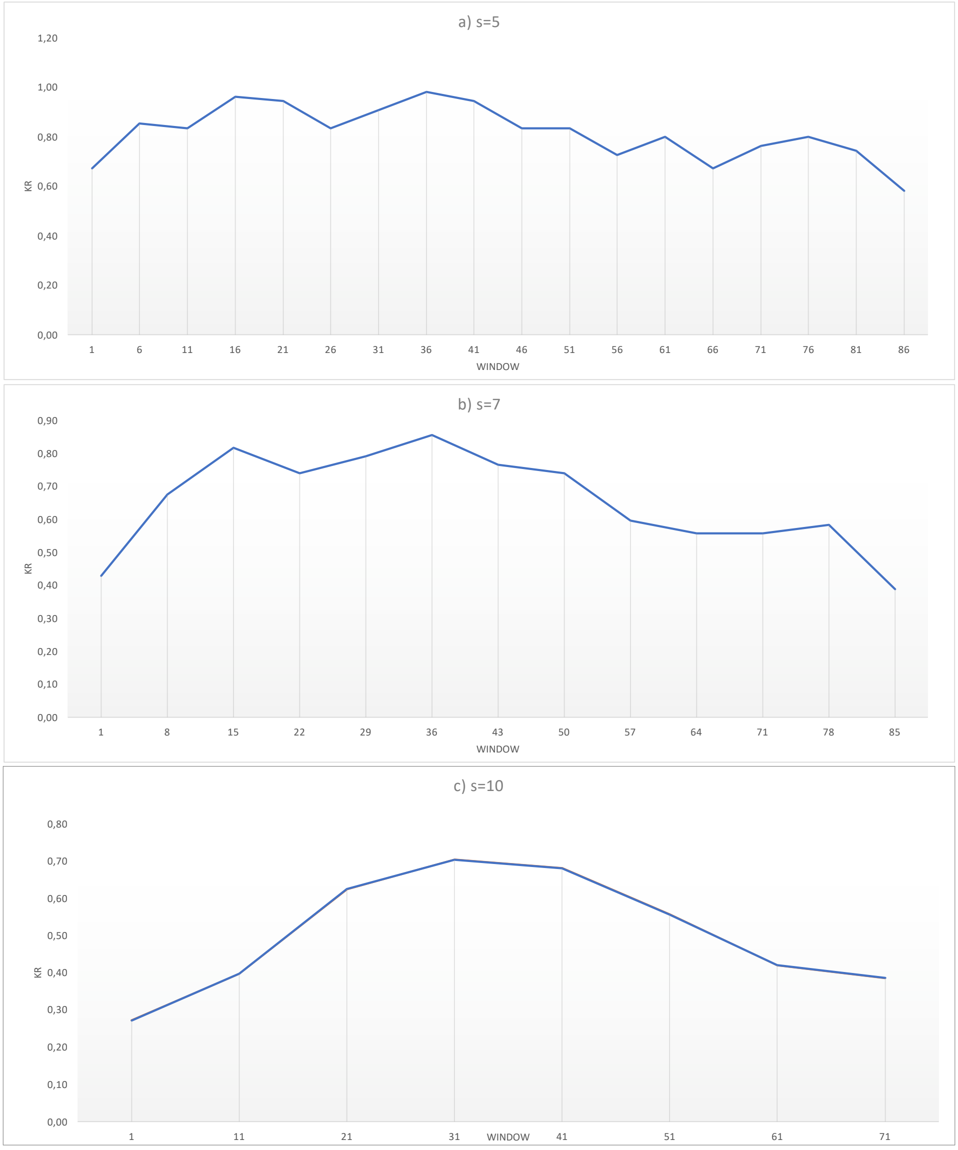


Fig. 2 Local concordance curve with $s = 5$ (a)), $s = 7$ (b) and $s = 10$ (c))

4 Final Remarks

Aim of the present work is to assess the degree of association among the items of a student's opinion survey. In particular, we analysed the rankings of the classes with respect to each of the items of the questionnaire, to investigate whether the most dissatisfactory classes are ranked in the first positions in all items of the survey (maximum local concordance in the first windows) or on the contrary, the worst ratings are equally distributed among the classes that show some elements of dissatisfaction. For our data set we found that both overall and local concordances are not very high and that where the dissatisfaction for one or more items is higher, (the first windows) or lower (the last), it is less concentrated than where it is non at its maximum or minimum. In order to proceed in this direction and deepen the analysis, we intend not only to broaden the data set, but also to perform an analogous analysis on the concordance among the most appreciated items.

References

1. Terzi, S. and Moroni, L.: Local Concordance and Some Applications. Social Indicator Research. 161, 7-8 (2022)
2. Kendall M.G., Babington Smith, B.: The problem of m rankings The Annals of Mathematical Statistics 10, 275-287 (1939)

High School Proficiency of Future University Students: an Analysis based on INVALSI Data

Profitto Scolastico dei Futuri Studenti Universitari: un'Analisi basata sui Dati INVALSI

Francesco Santelli, Gioia Di Credico and Claudia Di Caterina

Abstract Large-scale assessment in the education field is key in every Country. In Italy, the institute that is in charge of evaluating pupils' proficiency is the INVALSI, via a set of standardized tests, that go in parallel with traditional school evaluation. Data collected in a such way at the individual level pose a statistical challenge, given the nested structure of students-classroom-school and the repeated measure longitudinal observations that are obtained for each student. We propose in this context the streamlined version of the mean field variational Bayes (MFVB) algorithm for linear mixed models with crossed random effects, in order to obtain plausible predictors of pupils' performances. The results and interpretation of model coefficients are in line with the literature on educational data.

Abstract *La valutazione su larga scala nel campo dell'istruzione è fondamentale in ogni Paese. In Italia, l'istituto che si occupa di valutare le competenze degli alunni è l'INVALSI, attraverso una serie di test standardizzati, che vanno in parallelo senza sostituirsi alla tradizionale valutazione scolastica. I dati così raccolti a livello individuale pongono una sfida alla costruzione di modelli statistici adeguati. Proponiamo in questo contesto una versione semplificata del "mean field variational Bayes" (MFVB) per modelli lineari misti con effetti random incrociati, al fine di ottenere predittori plausibili delle prestazioni degli alunni. I risultati e l'interpretazione dei coefficienti del modello sono in linea con la letteratura sui dati delle performance scolastiche.*

Key words: crossed design, education, mean field variational Bayes, random effects.

Francesco Santelli
Università degli studi di Trieste, Trieste, e-mail: `fsantelli@units.it`

Gioia Di Credico
Università degli studi di Trieste, Trieste, e-mail: `gioia.dicredico@deams.units.it`

Claudia Di Caterina
Università degli studi di Verona, Verona, e-mail: `claudia.dicaterina@univr.it`

1 Introduction

Evaluation of the Italian educational system's performance is carried out by a set of standardized procedures. For what concerns the school system, the evaluation is performed from a quantitative internal benchmarking perspective [2]. The definition of such analysis refers to the fact that benchmarking has also been applied to contexts that are supposed to be apart from the logic of profit. For that reason, it has been adopted in private companies and public sectors [5]. The self-assessment concept is the core of internal quantitative benchmarking in the public sector. The idea is that the public sector can manage and produce by itself a considerable amount of data regarding its sub-entities to evaluate their performances. This benchmarking procedure often takes the name of *large scale scholastic assessment*, given the number of units involved and the fact that a set of standardized tools is used simultaneously over a considerable number of pupils.

The Italian national institute for the evaluation of the school system *INVALSI* is in charge of such assessment: it gathers data from various sources and provides comprehensive analyses. Several publications and technical reports are available on the INVALSI website. Data produced by INVALSI have been deeply explored by scholars, especially in investigations on the gender gap in mathematics [3, 4], or on the impact of the pandemic on the large-scale assessment governance [8]. Starting from INVALSI individual micro-data, several attempts have been made to produce statistical tools to properly analyze the proficiency values of students at different times and in different domains. To this aim, we propose a Bayesian approach based on a streamlined mean field variational Bayes (MFVB) algorithm (see Section 3). We fit a crossed random effects model to account for two data variability levels: students and tests [1].

2 Dataset

The data drawn from the Italian "Anagrafe Nazionale della Formazione Superiore" have been processed according to the research project "From high school to the job market: analysis of the university careers and the university North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero dell'Università e della Ricerca" (MUR), and INVALSI. This dataset, known as MOBYSU, has been collected to link university students' career to their scholastic proficiency. Several domain skills are tested on each pupil, such as Math, English language, Italian language, and Science.

Here, we focus on data from the cohort of pupils who graduated from high school in 2018/19 in Italy and then enrolled at an Italian university in the academic year 2019/20. Overall, such students are more than 240000. We select only those who attended for the first time the Italian school system and have no missing relevant information. A random sample of 14337 students, corresponding to 10% of the selected subset of interest, is then used for our analyses. The response variable of

Table 1 Model predictors, their description and reference categories.

Variable	Description	Reference
Gender	Male, Female	Female
Age	Reception (1-year ahead), Regular, Failed	Regular
Nation	Foreigner, Italian	Italian
Student escs (EscsStud)	student socio-economic level	
School escs (EscsSch)	school socio-economic level	
School type (SchTy)	13 categories	Classical Lyceum
Work Mother (Work.M)	5 categories	Unemployed
Work Father (Work.F)	5 categories	Unemployed
Year	School year	
NUTS1 classification (NUTS1)	5 areas	Center
School	Private, Public	Public

our model refers to students' marks: two recorded at the end of the first school term (Italian and Math, oral) and two through the standardized INVALSI tests (Italian and Math, written), during their 10th and 13th high school grades. Predictors involved in the analysis are listed in Tab.1. In detail, females constitute 57% of the observations, and only 3% of the students have non-Italian nationalities. Most students had a regular scholastic career (84%), while 11% of them started school one year before, and only 5% were held back one year. A vast majority of pupils (75%) attended lyceum type of schools (40% scientific-technological; 12% linguistic; 11% classical; 12% other). Vocational institutions were instead chosen by 21% of the students. Around 4% attended private schools, while the remaining were public. Following the Nomenclature of territorial units for statistics (NUTS), 40% of the students are from Northern Italy (29% North-West; 20% North-East), 20% from the Center and the others from South and Islands.

The Economic, Social and Cultural Status index (ESCS) gives information on students' family background with a mean value of 0.35 and a standard deviation of 0.94 at the individual level and 0.23 and 0.40 at the school level. Regarding parents' job positions, mothers are mainly unoccupied/retired (30%) or teachers/employees (26%). Fathers are less likely to be unoccupied/retired (5%), but they are involved in upper-level roles such as managers, university professors and freelance positions (29%). We also ensured that the selected sample reflects the composition of the original population in terms of subgroups defined by key covariates, like geographic location and type of school.

3 Methods and analysis

Each combination student/test is observed 2 times, namely in the 10th and 13th grades of high school. For each ith student, we assume these two scores $\boldsymbol{y}_{ii'}$ on test i' follow a linear mixed model with two crossed random effects:

$$\begin{aligned} \boldsymbol{y}_{ii'}|\boldsymbol{\beta},\boldsymbol{u}_i,\boldsymbol{u}'_{i'},\sigma^2 &\overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ii'}\boldsymbol{\beta}+\boldsymbol{Z}_{ii'}\boldsymbol{u}_i+\boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'},\sigma^2\boldsymbol{I}), \quad i=1,\dots,14\,337, \qquad (1)\\ \boldsymbol{u}_i|\boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}), &\qquad \boldsymbol{u}'_{i'}|\boldsymbol{\Sigma}' \overset{\text{ind.}}{\sim} N(0,\boldsymbol{\Sigma}'), \quad i'=1,\dots,4, \end{aligned}$$

where $\boldsymbol{X}_{ii'}$ is the 2×33 design matrix, $\boldsymbol{\beta}$ is the 33-dimensional vector of fixed-effect coefficients, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$ are 2×2 matrices corresponding to the random effects and u_i and $u'_{i'}$ are the random effects for students and tests, respectively. $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are the random effects 2×2 covariance matrices and σ^2 is the error variance. Model 1 includes a random intercept and slope for each student and test, that is

$$y_{ii'j}|\boldsymbol{\beta},u_{0i},u_{1i},u'_{0i'},u'_{1i'},\sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0+u_{0i}+u_{0i'}+(\beta_1+u_{1i}+u_{1i'})x_{1,ii'j}+\sum_{k=1}^{33}\beta_k x_{k,ii'j},\sigma^2)$$

for $j=1,2$, where $x_{1,ii'j}=1,2$ is the year indicator encoding the high school grades. We adopt a Bayesian approach with non-informative priors on all the model parameters. In particular, we use $\boldsymbol{\beta}\sim N_{33}(\mathbf{0},10^{10}\boldsymbol{I})$ and for variances we follow [6]:

$$\begin{aligned} \sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(1,1/a_{\sigma^2}), &\quad a_{\sigma^2}\sim \text{Inverse-}\chi^2(1,10^{-5}),\\ \boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} &\sim \text{Inverse-G-Wishart}(G_{\text{full}},4,\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}),\\ \boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'} &\sim \text{Inverse-G-Wishart}(G_{\text{full}},4,\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1}), \qquad (2)\\ \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), &\quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}=\{2\,\text{diag}(10^5,10^5)\}^{-1},\\ \boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}}), &\quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}'}}=\left\{2\,\text{diag}(10^5,10^5)\right\}^{-1}, \end{aligned}$$

We use product restriction III ([7], Sect. 3) on the mean field approximation of the joint conditional density of all parameters in (1) with the above priors:

$$q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}')=q(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')\,q(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}').$$

This restriction allows for a full joint posterior covariance matrix of $(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')$, leading to higher inferential accuracy but challenging computing that can be streamlined because the number of tests is small. The q-density parameters are obtained using a coordinate ascent algorithm running for 100 iterations.

INVALSI marks are standardized with respect to the national mean (200) and standard deviation (40), while first term scores are centered to the national high school diploma mean of 2018/19 scholastic year (73/10). This step permits obtaining a comparable scale of the two types of test scores that also accommodates our prior distribution specification 2.

4 Main findings

Fixed marginal effects suggest that female students and those from Northern Italy regions perform the best (especially from North-East), while South and Islands show the lowest proficiency. The socio-economic status dimension has a significant positive effect, as expected, especially at the individual level, but interestingly also at the school level. This means that a context effect is affecting individual performances. Lyceums pupils show the best scores on average, with Scientific lyceum overperforming all the others. Parents' jobs effect depicts a pattern in which students with a father who is involved in teaching and a mother who is not working, maybe having more time to dedicate to the children, are the ones with the best results. On the contrary, when the mother is involved in demanding activities (entrepreneur, manager), pupils' scores tend to decrease. No clear effects are found for students one year ahead, but students that already failed at least one year are obtaining results way below average. Students who failed are not recovering, maybe lacking adequate support in order to do that. Pupils with no Italian nationality also perform below the mean, and in this case, for sure the language barrier plays a key role. For what concerns random effects, INVALSI items have a positive effect and first-term negative effects on the intercept, thus students on average are judged with less generosity by their own teachers. Slope variability is pretty low compared to intercept variability, so the effect is more on the starting level than on the overtime scores.

Table 2 Random effects tests' estimates and standard deviations for tests and students (approximate posterior mean). Square root of diagonal entries of $\hat{\boldsymbol{\Sigma}}$ ($\hat{\boldsymbol{\Sigma}}'$) are denoted by $\hat{\sigma}$ ($\hat{\sigma}'$). The residual error standard deviation estimate is 0.848.

	Intercept	Slope
INVALSI - Ita ($\hat{u}'_1$)	0.514	-0.055
INVALSI - Mate ($\hat{u}'_2$)	0.568	-0.064
First term - Ita ($\hat{u}'_3$)	-0.402	0.089
First term - Mate ($\hat{u}'_4$)	-0.679	0.030
$\hat{\sigma}'$	0.649	0.075
$\hat{\sigma}$	0.586	0.047

5 Conclusions

The work analyzed Italian students' proficiency data using the streamlined MFVB algorithm, exploiting the potentialities of such a method. The obtained results align with the literature on deepening proficiency data and, more specifically, Italian INVALSI data. Interestingly, given that it is known that results improve with the increase in socio-demographic status, some unexpected results emerged from the parents' jobs. It emerges that a different dimension deviates from the mere socio-demographic status, related to the awareness of the importance of having good marks (teachers have a positive effect) and more time to dedicate to children support (when a mother is unemployed, results improve). Some specific groups of students, such as recent immigrants with not yet Italian nationality or pupils that already failed

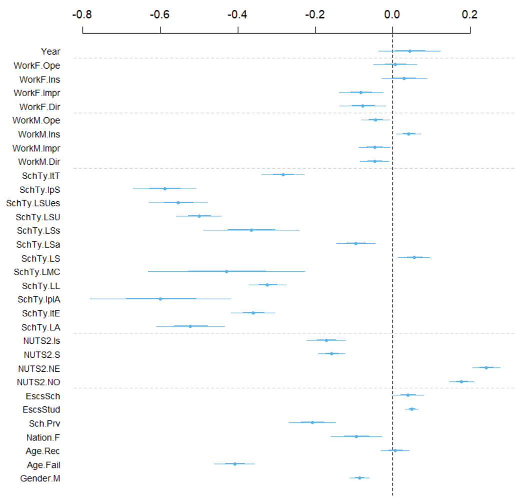


Fig. 1 Fixed effects: approximate posterior means (dots) and 95% credible intervals for the MFVB. The intercept estimate is equal to 0.243.

at least one year, seem more vulnerable. The geographical South-North divide is clear from the findings, and the gender gap emerges, favoring females.

References

1. Baayen, R. H., Davidson, D. J., Bates, D. M.: Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59, 390-412 (2008)
2. Binder, M., Clegg, B., Egel-Hess, W.: Achieving internal process benchmarking: guidance from BASF. Benchmarking: An Int. J. 13, 662-687 (2006)
3. Cascella, C., Giberti, C., Bolondi, G.: An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. Stud. in Educ. Eval. 64 (2020) doi:10.1016/j.stueduc.2019.100819
4. Giofré, D., Cornoldi, C., Martini, A., Toffalini, E.: A population level analysis of the gender gap in mathematics: Results on over 13 million children using the INVALSI dataset. Intell. 81 (2020) doi:10.1016/j.intell.2020.101467
5. Kouzmin, A., Löffler, E., Klages, H., Korac-Kakabadse, N.: Benchmarking and performance measurement in public sectors: Towards learning for agency effectiveness. Int. J. of Public Sect. Management 12, 121-144 (1999)
6. Huang, A., Wand. M. P.: Simple marginally noninformative prior distributions for covariance matrices. Bayesian Anal. 8 439-452 (2013)
7. Menictas, M., Di Credico, G., Wand, M. P.: Streamlined variational inference for linear mixed models with crossed random effects. J. Comput. Graph. Stat. 32, 99-115 (2022)
8. Milner, A. L., Mattei, P., Ydesen, C.: Governing education in times of crisis: State interventions and school accountabilities during the COVID-19 pandemic. Eur. Educ. Res. J. 20, 520-539 (2021)

Employment vulnerability of immigrants in the labour market – Does origin matter?

Vulnerabilità occupazionale degli immigrati nel mercato del lavoro – Conta l'origine?

Muhammed Bittaye

Abstract The purpose of this paper is to investigate whether the origin of migrant workers affects the level of employment vulnerability they face. We used the German Socio-Economic Panel (G-SOEP) for the period 2015 – 2018 to investigate the effect of origin on the level of employment vulnerability of employed immigrants in Germany and investigate which immigrant groups may experience higher levels of employment vulnerability because of their origin. The findings of the study show that the origin of the migrant workers affects the level of employment vulnerability they face in Germany. However, employed immigrants from Europe – Non-EU and Western Asia are the regions that experience significantly higher levels of employment vulnerability in Germany.

Abstract *Lo scopo di questo documento è indagare se l'origine dei lavoratori migranti influisce sul livello di vulnerabilità occupazionale che devono affrontare. Abbiamo utilizzato il German Socio-Economic Panel (G-SOEP) per il periodo 2015-2018 per indagare l'effetto dell'origine sul livello di vulnerabilità occupazionale degli immigrati occupati in Germania e indagare quali gruppi di immigrati potrebbero sperimentare livelli più elevati di vulnerabilità occupazionale a causa della loro origine. I risultati dello studio mostrano che l'origine dei lavoratori migranti influisce sul livello di vulnerabilità occupazionale che affrontano in Germania. Tuttavia, gli immigrati occupati dall'Europa - Non UE e dall'Asia occidentale sono quelli che sperimentano livelli significativamente più elevati di vulnerabilità occupazionale in Germania.*

Key words: Immigrant, Employed, Multidimensional employment vulnerability, Mixed effects model, Germany

---

Muhammed Bittaye

Università degli Studi dell'Insubria; e-mail: mbittaye@uninsubria.it

1 Introduction

Employment vulnerability here refers to the risk of being exposed to adverse working conditions which may threaten the well-being and living conditions of workers. Thus, employment vulnerability is not limited to some category of employees, work-related dimensions, or job characteristics but extends to all individuals in all sorts of jobs [1, 2]. As a result, employment vulnerability includes not only employees and those in formal jobs but also includes the self-employed, the unemployed, and those employed in informal and illegal jobs [1].

For migrant workers, the level of employment vulnerability can greatly depend on the host country and the relationship between the host and the sending country. Jayaweera and Anderson [3] argued that language, social networks, and discrimination are part of the main challenges migrant workers are faced with. However, the extent of the challenges migrant workers face may significantly differ from one host country to another.

To the best of our knowledge, [4] conducted the only study of the origin of workers on employment vulnerability. However, the categorization of the regions of origin of the workers may have been few and too broad. Thus, in this study, we use more categories of the region of origin of the workers and also focused only on immigrant workers in Germany. This paper contributes to the literature on labour migration. The paper investigates the possible effect of the origin of migrants on the level of employment vulnerability they may face in Germany. To do this, we analyse the effect of origin on employment vulnerability using two panel data models, a linear mixed effects model and a logistic mixed effects model. We then investigate which immigrant groups are likely to experience higher levels of employment vulnerability because of their origin.

2 Literature review

2.1 Migrant workers and employment vulnerability

In their study of measuring employment vulnerability, [2] focused on 15 countries in the European Union and estimate the employment vulnerability by country. Their findings show that there are differences in the level of employment vulnerability between countries. Countries like the Nordic countries have strict employment protection laws and as such may be expected to experience low levels of employment vulnerability, while countries in the south may experience higher employment vulnerability as employment policies are weaker in these countries.

To investigate whether there is a difference in employment vulnerability between migrants and EU nationals in the European Union, [4] used propensity score matching to compare the levels of employment vulnerability between migrants and EU citizens. Using three categories of origin, developed European countries, EU new members, and non-European countries, their findings show no difference in employment vulnerability between the groups. However, the findings show that there exists a difference in the level of employment vulnerability by skill level. For low-skilled workers, migrants have a lower level of employment vulnerability than native EU workers. While, for high-skilled workers, the study found that migrant workers have a higher level of employment vulnerability than the natives.

2.2 The trend of international migration in Germany

Germany has long been a country that attracts a high inflow of migrants. According to the United Nations Department of Economic and Social Affairs (UN DESA) data on international migrant stocks, in 2015 a total of 10.2 million migrants lived in Germany [5]. However, nine of the sending countries with the highest migrant stock in 2015 made up 61.9% of the total migrants in Germany.

3 Data and estimation strategy

3.1 Data source

For this study, we use the German Socio-Economic Panel (G-SOEP). The study focuses only on immigrants in Germany and excludes all German nationals. Additionally, the study focuses only on immigrants that are employed, and aged 18 years and over. Unemployed migrants, those not in the labor force, self-employed workers, interns and apprentices are excluded from the study. The study, therefore, considers the period from 2015 to 2018, covering a total of 2,955 individuals across 6,499 observations. With 1,981 employed immigrants aged 18 years and over in 2015, 1,622 in 2016, 1,480 in 2017, and 1,416 in 2018.

3.2 Data engineering

As some of the variables of interest have some missing values, we used multiple imputation to address the problem of missing values for our study. As this paper is an extension to the study by [6], we maintained the same imputation technique for the imputation of our missing values. Using a total of 5 imputations (i.e., $M=5$), and using the multiple imputation by chained equations (MICE), which allows to specify different imputations methods for each of the missing variables, we imputed the missing values for the variables education, firm size, length of stay, job-training mismatch, oral ability, reading ability and writing ability.

3.3 Dependent and independent variables

Using two models, a linear mixed effects model, and a logistic mixed effects model, we analysed the effect of origin on the employment vulnerability of employed immigrants in Germany. In the mixed effects model, we used the multidimensional employment score (c_i) as the dependent variable and in the logistic mixed effects model, we used the multidimensional employment vulnerability identification function (v_i^k) as the dependent variable.

The employment vulnerability score (c_i) measures an employee's total weighted univariate vulnerability across all indicators. The employment vulnerability score was constructed using 10 indicators. And for each of the indicators, an employed immigrant is considered vulnerable in that indicator if the employee's achievement, x_{ij}, falls below the indicator-vulnerability cutoff, z_j. A score of 1 is assigned if the employee falls below the indicator-vulnerability cutoff, z_j, and assigned a score 0 zero otherwise [6].

While the variable multidimensional employment vulnerability identification function (v_i^k), a binary variable, equals 1 if the employee is multidimensionally vulnerable in the labour market (given a predefined threshold k) and 0 otherwise. To determine which employed immigrants are multidimensionally vulnerable in the labour market, the multidimensional employment vulnerability score (c_i) was used, and a predefined threshold called the multidimensional employment vulnerability cutoff, k, was applied to determine who is multidimensionally vulnerable in the labour market. With a multidimensional employment vulnerability cutoff of $k=5$, an employee is considered multidimensionally vulnerable in the labour market if his multidimensional employment vulnerability score (c_i) is at least the value of k [6].

3.4 Estimation strategy

The model to be estimated under the mixed effects model is:

$$y_{it} = \boldsymbol{x'}_{it}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it},\ \ i = 1, \dots, n,\ \ t = 1, \dots, T_i$$

with $u_i \sim N(0, \sigma_u)$ and $u_{it} \sim N(0, \sigma_\varepsilon)$

While for the logistic mixed effects model, we have:

$$\Pr(y_{it} = 1 | \alpha_i, \boldsymbol{x}'_{it}\boldsymbol{\beta}) = \frac{e^{(\alpha_i + x'_{it}\boldsymbol{\beta})}}{1 + e^{(\alpha_i + x'_{it}\boldsymbol{\beta})}}$$

where the latent variable y_{it}^* is given by:

$$y_{it}^* = \boldsymbol{x'}_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it},\ \ i = 1, \dots, n,\ \ t = 1, \dots, T_i$$

$$y_{it}^* = \boldsymbol{x'}_{it}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it},\ \ i = 1, \dots, n,\ \ t = 1, \dots, T_i.$$

$$y_{it} = 1(y_{it}^* > 0) = \begin{cases} 1\ if\ y_{it}^* > 0 \\ 0\ if\ y_{it}^* \leq 0 \end{cases}$$

4 Results and discussion

Table 1 Mixed effects regression of employment vulnerability score

Variables	Employment vulnerability score	
	Coef.	**Std. error**
Region of origin (Ref: EU & EFTA)		
Europe - Non-EU	0.7219**	0.2898
Western Asia	0.7339**	0.3568
Rest of Asia	0.5883	0.7051
Africa	0.2705	0.9351
America & Oceania	0.6018	0.6736
Constant	1.7095***	0.1467
Observations	6,499	
Number of individuals	2,955	
Controls	Yes	
F statistic	14.2	
p-value (F)	0.0000	

Note: *** p<0.01, ** p<0.05, * p<0.1

Table 2 Logistic mixed effects regression of the employment vulnerability identification function

Variables	(1) **Vulnerable employee at *k=1***	(2) **Vulnerable employee at *k=3***	(3) **Vulnerable employee at *k=5***	(4) **Vulnerable employee at *k=7***
Region of origin (Ref: EU & EFTA)				
Europe - Non-EU	-0.0866	0.6367***	0.5050***	0.6780**
	(0.3289)	(0.1931)	(0.1849)	(0.3144)
Western Asia	0.6227	0.4778**	0.6519***	0.0190
	(0.4261)	(0.2317)	(0.1943)	(0.4387)
Rest of Asia	-0.0870	0.7286**	0.6791**	1.2297***
	(0.6046)	(0.3306)	(0.3198)	(0.4400)
Africa	0.3663	1.6270***	1.1493***	1.2768**
	(0.9378)	(0.5837)	(0.4204)	(0.5732)
America & Oceania	1.4328	0.5316	0.0388	-0.9155
	(1.1949)	(0.4443)	(0.4732)	(1.3876)
Constant	2.7441***	-2.0831***	-4.9243***	-8.0359***
	(0.6071)	(0.3140)	(0.3359)	(0.8703)
Observations	6,499	6,499	6,499	6,499
Controls	Yes	Yes	Yes	Yes
F statistic	3.3	13.5	12.6	3.2
p-value (F)	0.0000	0.0000	0.0000	0.0000

Note: *** p<0.01, ** p<0.05, * p<0.1; standard errors in parentheses

5 Conclusion

The findings of the study show that even though the average employment vulnerability score of the regions is less than 5, and that the incidence of multidimensionally vulnerable employees at the cutoff *k=5* for each of the regions is not higher than 42%, the origin of the employed immigrants has an effect on their level of employment vulnerability in Germany. The findings of the study revealed that origin does affect the level of employment vulnerability faced by migrant workers in Germany.

Thus, although Germany is one of the countries with the highest labour rights compliance on target 8.8 of the Sustainable Development Goals (SDGs), the level of employment vulnerability is influenced by the origin of the employed immigrants. However, only employed immigrants from Europe – Non-EU and Western Asia are significantly more likely to experience higher levels of employment vulnerability than employed immigrants from EU & EFTA. Employed immigrants from these two regions (i.e., Europe – Non-EU and Western Asia), therefore, appear to be affected more by employment vulnerability in Germany than migrants from other regions.

References

1. Bazillier R., Boboc C.: Labour migration as a way to escape from employment vulnerability? Evidence from the European Union. Applied Economics Letters, 23(16), 1149–52 (2016)
2. Bittaye M.: Measurement of Employment Vulnerability of Immigrant Workers in Germany (2022)
3. Burgess J., Connell J.: Vulnerable work and strategies for inclusion: an introduction. International Journal of Manpower, 36(6), 794–806 (2015)
4. Greenan N., Seghir M.: Measuring vulnerability to adverse working conditions, evidence from European countries. hal-02172377 (2017)
5. Jayaweera H. and Anderson B.: Migrant workers and vulnerable employment: A review of existing data, 1–49 (2008)
6. United Nations.: International Migrant Stock. Department of Economic and Social Affairs, Population Division, (2020). Available from: https://www.un.org/development/desa/pd/content/international-migrant-stock

The effect of pricing policies on students' use of university canteens

L'effetto delle politiche di prezzo sull'uso delle mense universitarie da parte degli studenti

Lucio Masserini, Matilde Bini and Valentina Lorenzoni

Abstract University canteens play an important role in academic life; they not only allow students to benefit from subsidised food services and meals at lower prices than those commonly available at other local eateries but also affect other aspects, such as students' health, social relationships and academic achievement. Using a quasi-experimental design and a difference-in-differences approach, this study aims to evaluate the impact of an income-based pricing policy on students' frequency of using university canteens and their meal choices. Using data from an Italian university, this study shows that users who experienced a meal price increase significantly reduced their use of university canteens.

Abstract *Utilizzando un disegno quasi sperimentale e un approccio basato sulla differenza nelle differenze, questo studio si propone di valutare l'impatto di una politica dei prezzi basata sul reddito sulla frequenza di utilizzo delle mense universitarie da parte degli studenti e sulle loro scelte di pasto. Utilizzando i dati di un'università italiana, questo studio dimostra che gli utenti che hanno subito un aumento del prezzo dei pasti hanno ridotto significativamente l'uso delle mense universitarie.*

Key words: difference-in-differences, food choice, price change, pricing policies, university canteen meals, university finance

---

Lucio Masserini

Department of Economics and Management, University of Pisa, Pisa, Italy; e-mail: lucio.masserini@unipi.it

Matilde Bini

Department of Human Sciences, European University of Rome, Rome, Italy; e-mail: matilde.bini@unier.it

Valentina Lorenzoni

Institute of Management, Sant'Anna School of Advanced Studies, Pisa, Italy; e-mail: valentina.lorenzoni@santannapisa.it

1 Introduction

Universities and related facilities play a strategic role in the promotion of students' well-being by ensuring their academic achievement and enhancing their social life, overall health status and human dignity [3].

Although the literature in the field of university facilities is sparse [1], some studies have revealed that easy access to high-quality food is an important aspect of students' well-being with relevant short- and mid-term consequences. Research has shown that adequate access to food has an impact on students' performance and retention, potentially contributing to their general health and having considerable implications for long-term health preservation and the reduction of the collective health care burden [4,5]. University canteens are generally able to offer low-cost meals and play an important role in the promotion of students' university careers and general well-being from a public health perspective, since they allow students to benefit from subsidised food services and meals at lower prices than those commonly found among their competitors. Using a difference-in-differences (DID) approach, this study aims to evaluate the effect of introducing an income-based pricing system on students' use of university canteens and their meal choices at these canteens. DID is a statistical technique used for policy evaluation in quasi-experimental designs with panel data. Two aspects were analysed as outcome variables to evaluate the impact of the new pricing system on students' eating habits: 1) the frequency of university canteen use and 2) the proportion of light and large meals consumed. The study was carried out in a university in Central Italy that has three university canteens administered by the Financial Aid and Scholarship Office.

2 Method

Given that this study used panel data and involved a natural experiment, where the treatment consisted of the introduction of a new income-based pricing system, a DID approach [2,6] was employed. The DID estimator is a popular tool in quasi-experimental designs for evaluating the impact of a treatment or intervention using a repeated cross-sectional or panel design. In our setting, the same students belonging to the treated and control groups were observed for two periods, before and after treatment, in such a way that the students were made into their own controls. The basic idea behind the DID technique is that in the absence of treatment, the change in the treated outcome would have been the same as the change in the non-treated outcome. Thus, although the outcome levels may differ between the treated and control groups even in the pre-treatment period, the impact of the treatment could be measured by the DID estimator as the difference in average outcomes in the treatment group before and after treatment minus the difference in average outcomes in the control group before and after treatment [6]. The DID estimator can be easily implemented using a regression approach, which can obtain the estimates and corresponding standard errors in one step:

$$y_{it} = \beta_0 + \beta_1 t_{it} + \beta_2 T_{it} + \beta_3 (t_{it} \times T_{it}) + X_{it}\delta + \varepsilon_{it}$$

where y_{it} represents the relevant outcome variable (the frequency of canteen use and the proportions of FM, LM1 and LM2); t_{it} is a binary variable for the period of observation, where $t_{it} = 0$ stands for the period before April 1, 2018, and $t_{it} = 1$ stands for the period from April 1, 2018, onwards; T_{it} is a binary treatment variable, where $T_{it} = 0$ indicates students in the control group (those with unchanged meal prices) and $T_{it} = 1$ stands for students in the treatment group (those whose meal prices increased or decreased); X_{it} is a vector of covariates, entered in the model as control variables; and ε_{it} is the error term. β_3 is the coefficient of interest and represents the DID estimator. It results from the interaction term obtained by multiplying the treatment indicator and the period of observation; it takes a value of 1 for students whose meal prices changed after the treatment. For the frequency of canteen use, the regression equation was estimated with the ordinary least squares technique; for the proportions of meal types, a beta regression approach was carried out (for more details, see [7]). In both cases, robust clustered standard errors were used to control for heteroskedasticity and clustered data.

3 Results

Two outcome variables were used to assess the impact of the new pricing system on the students' eating habits: 1) the frequency of university canteen use and 2) the proportion of meal types (large and light meals) consumed.

Taking the number of accesses as a dependent variable, the effect of price variation was evaluated using two DID models: one for students whose meal prices increased and the other for those whose meal prices were reduced. Taking the proportion of meal types consumed (FM, LM1 and LM2) as a dependent variable and the two intervention groups, additional DID models were also estimated. In both cases, analyses were carried out separately for all students, as well as for frequent and non-frequent users. The general DID model used in our analysis allows for the inclusion of both fixed and time-varying covariates. For each model, we present only the DID estimates summarised as β_3, which represents the main parameter of interest and measures the magnitude and direction of the effect of the price variation.

Table 1 shows the results of the DID linear regression models that were estimated using the frequency of university canteen use as a dependent variable. With regard to the students whose meal prices increased (Table 1), the DID estimate (β_3) indicates a significant decrease in the total number of accesses; this effect was also observed among both frequent and non-frequent users. Given the logarithmic scale of the dependent variable, results can be interpreted more effectively in terms of percentage change. On average, the price increase produced a 29.9% (95% CI: 23.2%–36.0%) reduction in the number of accesses. The magnitude of the effect produced by the price variation is higher among frequent users, whose canteen use was reduced by 36.3% (95% CI: 25.2%–45.7%) as compared to the 20.2% (95% CI: 7.0%–31.5%) decrease among non-frequent users. No effect was detected among the students whose meal prices were reduced.

Table 1 DID estimates on the number of accesses according to the type of price change

	Overall	Non-frequent users	Frequent users
Price increase	-0.356 (0.046)***	-0.225 (0.078)**	-0.450 (0.081)***
Price reduction	0.043 (0.041)	0.071 (0.075)	0.096 (0.065)

The dependent variable is the log-transformed frequency of canteen use. Models also include the individual-level covariates (age, gender, course year and faculty). Robust standard errors are in parentheses. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Other results we obtained show the effect on FM consumption. DID estimates show significant effects only for students whose meal prices increased. Among these students, the probability of always choosing and never choosing FM increased by 40.9% and 87.8%, respectively. This effect was not observed when the sub-groups of frequent and non-frequent users were evaluated. Among frequent users, the price increase was associated with a 12.9% reduction, on average, in choosing FM.

Among the students whose meal prices increased, results show a rise in the proportion of those who chose LM1, both overall (+15.8%) and in the two-subgroups of frequent (+21.9%) and non-frequent users (+29.3%). On the other hand, the probabilities of never and always choosing LM1 were not affected by the price variation. When the effect of price reduction on LM1 selection was analysed, the only significant effect observed was a threefold increase in the probability of always choosing LM1 among non-frequent users.

With regard to the proportion of LM2 consumed, the analysis of students whose meal prices increased shows that the intervention produced a higher probability of not choosing LM2 both overall (+42.2%) and among non-frequent users (+42.8%). Conversely, the probability of always choosing LM2 decreased significantly in the same groups. Overall and among non-frequent users, price reduction had a significant negative effect on the probability of always choosing LM2. No other significant effect was associated with price reduction.

References

1. Blichfeldt, B.D., Gram, M.: Lost in transition? Student food consumption. Higher Education, 65, pp. 277-289 (2013)
2. Card, D., Krueger, A.: Minimum wages and employment: a case of the fast food industry in New Jersey and Pennsylvania. American Economic Review, 84, pp. 772-784 (1994)
3. Gallegos, D., Ramsey, R., Ong, K.W.: Food insecurity: is it an issue among tertiary students? Higher Education, 67, pp. 497-510 (2014)
4. Hughes, R., Serebryanikova, I., Donaldson, K., Leveritt, M.: Student food insecurity: the skeleton in the university closet. Nutrition & Dietetics, 68(1), pp. 27-32 (2011)
5. Kirkpatrick, S.I., Tarasuk, V.: Food insecurity is associated with nutrient inadequacies among Canadian adults and adolescents. Journal of Nutrition, 138, pp. 604-612 (2008)
6. Meyer, B.: Natural and quasi-experiments in economics. Journal of Business & Economic Statistics, 13, pp. 151-161 (1995)
7. Puhani, P.A.: The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. Economics Letters, 115(1), pp. 85–87 (2012)

Gig workers' identikit
L'identikit dei gig workers

Emma Zavarrone and Alessia Forciniti

Abstract The paper analyzes the young gig workers in Italy with the dual aim of: 1) defining their identity by sociodemographic features and professions, and 2) detecting perceived advantages and disadvantages. A simple random sample of 804 respondents was interviewed. To investigate the associations among the features, we used the Multiple Correspondence Analysis (MCA). The results highlighted two main categories of workers: a) freelancers, male, in the Central and North, with high education, middle-to-high income, and specialized professions; b) occasional, female, living in the South and islands without professional specificity. A greater perception of advantages and disadvantages among occasional workers with middle-to-high education was observed compared to freelancers with middle-high education.
Abstract *Il paper analizza i giovani gig worker in Italia con il duplice obiettivo di: 1) definire la loro identità attraverso caratteristiche socio-demografiche e professioni, 2) rilevare vantaggi e svantaggi percepiti. Un campione casuale semplice di 804 rispondenti è stato intervistato. Per studiare le associazioni tra caratteristiche, abbiamo usato l'Analisi delle Corrispondenze Multiple (MCA). I risultati hanno evidenziato due principali categorie di lavoratori: a) freelancer, maschio, nel Centro e Nord, con alta formazione, reddito medio-alto, e professioni specializzate; b) occasionale, femmina, che vive a Sud o Isole, senza specificità professionale. Maggiore percezione di vantaggi e svantaggi è osservata tra gli occasionali con istruzione medio-alta rispetto ai freelancer con istruzione medio-alta.*

Key words: gig economy, gig workers, multiple correspondence analysis

---

Emma Zavarrone
Iulm University, Via Carlo Bo, e-mail: emma.zavarrone@iulm.it

Alessia Forciniti
Iulm University, Via Carlo Bo, e-mail: alessia.forciniti@iulm.it

1 Introduction

Due to the fast growth of information and communication technology started in the 2000s, the industries and labour market underwent a digital revolution, which implicated a digitalization of systems and the increase of on-demand platforms.
This shift led to a new labour force - distinct from already existing offline one - that was defined as autonomous employment or 'gig economy'. The etymology of gig economy is uncertain and referred to an economic model based on independent contractors, temporary employees, freelancers, on-demand workers [6] who work for companies or clients on a project-by-project basis, and are recruited through online platforms. Many European companies are opting for this workers called *gig workers*, because they are more cost-effective than traditional employees and offer more flexibility. Often the gig workers are those who have a limited-term contract or who do more than one job at the same time. During the last eleven years, the second job is a trend that has grown rapidly in countries of European Union (EU), reaching its maximum of 8,165 units in 2019, and recording a slight decrease during the Covid-19 [2]. While, from 2012 to 2021 the contract of limited duration represented a trend for 62.96% of European countries, with an average annual between 11.9 and 13.8 in people aged 15-64 years.
In Italy, the employment rate 2021 in the age group 15-64 for those with a limited-term contract is 13.2, while for those who have a second job it is 9.2 [2]. Despite the potential benefits of gig work, such as time and geographical flexibility, better work-life balance, opportunities for entrepreneurship, there are also many challenges, first of all the legal protections since gig workers often risk losing their income if a client terminates their contract or if they are unable to find new clients. Another issue is the lack of social protection, which is debated in many countries and includes the lack of: a) paid sick time; b) employer-provided health insurance; c) bargaining rights; d) unemployment insurance, among others. In fact, in most EU member states, the laws governing payments and rights to social protection schemes rely primarily on typical workers rather than on workers with non-standard agreements, indicating a lack of adaptation of modern social protection systems. Therefore, self-employed workers or gig workers represent a category of workers with a very uncertain defining and regulatory framework.

Our work aims to investigate the profile of Italian gig workers, with a focus on young people, and perceptions of the advantages and disadvantages characterizing this economic model. Specifically, we examined two research questions (RQs).
RQ_1: Who are the young people in Italy who use the gig economy in the labor market? Can we identify a relationship between the type of gig worker and socio-economic characteristics?
RQ_2: The perception of the advantages and disadvantages of condition of gig workers changes in relation to the education and profession?

The feature of our paper is not of methodological innovation to answer RQs, but the application of a quantitative approach to the a new and constantly evolving cultural and regulatory context of analysis, in order to provide answers that may contribute to better frame the phenomenon of gig workers in Italy.

To follow, Section 2 presents data and methods; Section 3 describes the results; and Section 4 indicates discussion and conclusions.

2 Data e Methods

2.1 Data

Data collection consists of a simple random sample of 804 respondents, aged 18 - 30 years, who live in Italy and work as self-employed workers or gig workers.
Data were collected within a project co-founded by the European Union under the title *Gig Up*, aimed at offering a program to address NEET unemployment in the gradual transition to the labor market by using the gig economy as an activator of employment skills. The survey was based on questionnaire administration in Italian language consisting of 20 multiple answer questions divided into three sections: 1) respondents' demo-economic characteristics, including the type of gig worker (freelancer, occasional or reluctant) and profession; 2) apps used for job search; 3) advantages, disadvantages of the condition of gig worker.

The sample is characterized by a higher percentage of female respondents, 70%, compared to 29% of male and 1% of other. The majority (58%) were between 25 and 30 years, and the most frequent education is the secondary school (42.5%), followed by bachelor degree (24.6%) and master degree (21.9%). In addition, the highest education is prevalent in male respondents.

During the data entry, the answers concerning profession, apps, advantages and disadvantages were encoded such as single variables, whose the modes were represented by a binary answer, equals 1 in the case of an affirmative answer, and 0 in the event of a negative answer. Data has been organized into a response pattern matrix *"cases x variables"*, $\mathbf{M}_{804x97}$, consisting of 804 rows, of respondents, and 97 columns, corresponding to the variables examined for each section of the questionnaire. By means of the cleaning data process, we removed the cases where the answers were missing (n=4). Thus, the matrix used for data analysis consists of 800 rows and 97 columns, $\mathbf{M}_{800x97}$.

2.2 Methods

To answer RQ_1 and outline the young gig workers in Italy, from matrix $\mathbf{M}$, after the data cleaning, we selected only the variables involving answers on profession and demographic characteristics (type of gig worker, gender, age, geographic area, education, income) obtaining a new table $\mathbf{A}_{800x15}$, with 800 individuals (rows) and 15 variables (columns).

On the matrix **A** we performed the Multiple Correspondence Analysis (MCA) [4; 5]. MCA is a well-known multivariate technique for data analysis as extension of the simple correspondence analysis (CA) [1], and it is used to summarize and visualize in a geometric space the data table with more than two categorical variables. Specifically, we performed an eigenvalue-eigenvector decomposition based on the Burt matrix, as cross-product of matrix which concatenates all two - way cross-tabulations between pairs of the variables examined. Furthermore, in order to resolve the structure of Burt matrix that overestimates the total inertia due to sub-matrices on the main diagonal that are cross tabulations of each variable with itself, we adopted the adjustment of inertias approach [3] that rescaled the coordinates of the solution to best fit the pairwise cross-tabulations off the main diagonal.

To respond to RQ_2, instead, we selected from the original **M** matrix only the variables related to the type of profession, education, strengths and weaknesses of this economic model, building a new matrix $\mathbf{B}_{800x18}$, with 800 respondents (rows) and 18 variables (columns), and we replicated the methodological approach of MCA in order to explore the associations that can better describe and identify this aspect of the gig workers in Italy.

3 Results

The MCA map frames the Italian young gig worker by answering RQ_1. Through the first dimension, explains 62.1% of variance, and with the second one, 7.1% (Fig.1).

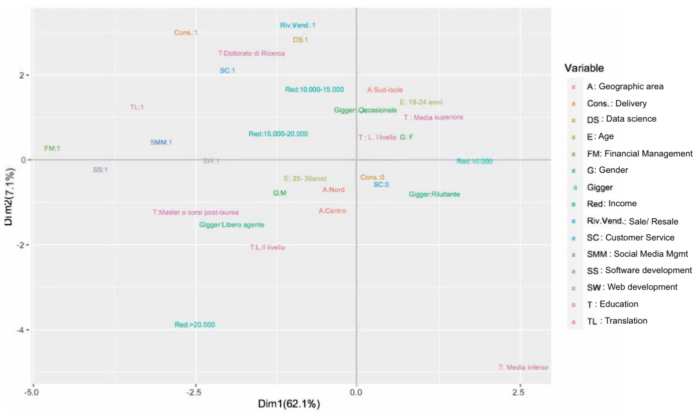


Fig. 1 MCA Map for profiling Italian gig workers

The gig worker's identity may be summarized by the four quadrants: the first is related to those who are occasional, with middle education or bachelors, mostly women living in the South or the Islands, aged 18–24 years, who have a low income

(10,000 euros) and without any professional specificity. The second one is characterized by the highest education (Ph.D), an income between 10,000 and 20,000 euros, and diversified professions: financial management (FM), data science (DS), sales and resale (Riv. Vend.) and translation (TL). A variability of income is observed due to the diversity of professions belonging to this quadrant; different jobs have different degrees of specialization. The third quadrant is characterized by freelancer, male and aged between 25 and 30 years, from Central and Northern, with an income of more than 20,000 euros and medium-high education (post-graduate). The professions are associated with technology, such as software development (SS) and web development (SW). The fourth one is related to workers who are reluctant and are not characterized by any professional profile. This map identifies a greater specialization for male of Central - North compared to females of South and Islands.

The MCA map of Fig. 2 answers RQ_2 and shows strengths and weaknesses of the condition of gig worker based on education and profession.

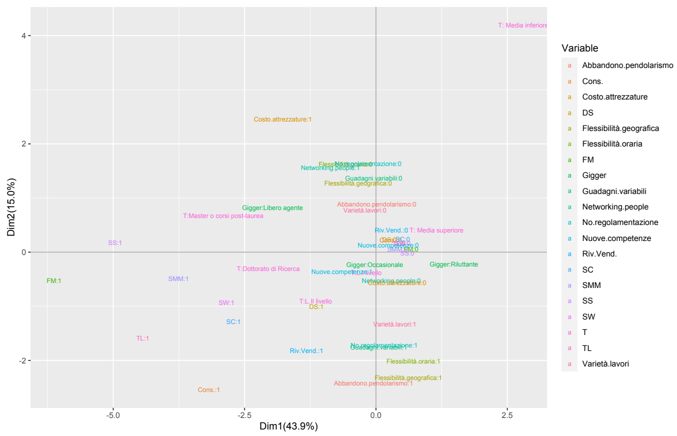


Fig. 2 MCA Map of strengths and weaknesses

The factorial projection with the first dimension explains the 43.9% of variance and with the second 15% appears divided into one upper section and one lower section. In the upper part, we observe the post-graduate characterizes by free agents, mainly associated with software development (SS), that indicates networking people as the main advantage and equipment costs as the main disadvantage. The lower section is characterized by education from bachelor to Ph.D, with a greater number of diversified professions: financial management (FM), translation (TL), web development (SW), customer service (SC), and deliveries (Cons). We see the acquisition of new skills and the abandonment of commuting as strengths, and the lack of regulations and uncertain variability of earnings as weaknesses. Finally, in the fourth quadrant, there are the reluctants which identifiy time and geographic flexibility as prevailing advantages.

4 Discussion and conclusions

The work investigated the young gig worker in Italy with the double objective of: 1) framing the identity of gig workers and identifying the socio-demographic characteristics associated with the different professions; and 2) detecting what advantages and disadvantages are perceived depending on the degree of education and the job. The survey was configured by administering a questionnaire to a simple random sample of 804 respondents between 18 and 30 years of age. We used the MCA to map the variables in a shared geometric space and study their relationships. The characteristic is therefore not the innovation of the method but its application to an evolving economic model. The results highlighted the main types of gig workers: 1) freelancers: men, between 35 and 30 years who live in the Central and Northern, have a middle-to-high education, have an income of more than 20,000 euros, and perform work based on technology ; 2) have a high level of education (Ph.D), a high income (between 10,000 and 20,000 euros), and have diverse professions, from the most specialized (data science, financial management, etc.) to those less specialized (sales, deliveries); 3) occasional workers: women with higher middle education or a bachelor's degree living in the South and Islands with no specialization. The advantage perceived by those who are free agents, have a postgraduate degree, and carry out technological activity is related to networking, while the disadvantage is the cost of equipment. This denotes that the freelancer with medium-high education is oriented to perform gig work as a primary source of income, paying little attention to the pros and cons. The occasional worker with middle-to-high education (bachelor, master, and Ph.D) instead of seeing a diverse range of jobs (delivery, financial management, etc.) defines strengths as new skills and weaknesses as a lack of regulation and variability of earnings, highlighting a greater awareness of the uncertain framework that characterizes the condition of gig worker.

References

1. Benzécri, J. P.: L'analyse des données. Tome 2: L'analyse des correspondances. Dunod, Paris (1973)
2. Eurostat: Self-employment statistics (2022) Available via DIALOG. https://ec.europa.eu/eurostat/statistics-explained/index.php? Cited 1 December 2022
3. Greenacre M. J.: Correspondence Analysis in Practice. Academic Press, London (1993)
4. Greenacre, M.J., Pardo, R.: Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. Sociological Methods and Research 35, 193-218 (2006)
5. Nenadic, O. and Greenacre, M.: Computation of Multiple Correspondence Analysis with Code in R. In: Greenacre, M. and Blasius, J. (eds.) Multiple Correspondence Analysis and Related Methods, pp. 523-551. Chapmann & Hall / CRC, Boca Raton (2007)
6. Vallas, S., Schor, J.B.: What Do Platforms Do? Understanding the Gig Economy. Annual Review of Sociology 46 (2006)

Session of free contributes SFC2 - *Tourism and territory*

1. *Multi-source approach for statistics in tourism sector* (Bianchino A., Fusco D., Giordano P., Liguori M.A. and Summa D.)
2. *Statistical analysis of tourism sustainability in Campania: post Covid-19 review* (Giacalone M., Basile V. and Bellucci M.)
3. *Investigating recent changes in dietary behavior* (D'Uggento A., d'Ovidio F.D., Toma E. and Onorati M.G.)
4. *Depopulation in the Abruzzo municipalities* (Carulli A.L., Di Spalatro D. and Valentini A.)
5. *The Productions System of Inland Areas* (Madia Carucci A.M. and Regano A.)

Multi-source approach for statistics in tourism sector

Approccio multi-fonte alle statistiche sul turismo

Antonella Bianchino, Daniela Fusco, Paola Giordano, Maria Antonietta Liguori and Donato Summa

Abstract Tourism sector, in particular rural tourism, combines customer needs for food quality products, agri-environmental and economical sustainability. This study proposes a theoretical model for building a synthetic index starting from 22 basic indicators declined in three Pillars: 1) Infrastructural density and touristic fluxes; 2) Economical impact of touristic sector and 3) Agricultural sector support. It was used a multi-source approach (statistical and administrative sources, big data), to meet the need to develop a system of homogeneous, comparable and updated statistics. The territorial dimension focuses on municipalities belonging to the National Strategy for Inner Areas (SNAI), where the agricultural sector plays a central role.
Abstract *Il settore del turismo, in particolare il turismo rurale, combina le esigenze dei clienti di prodotti alimentari di qualità, sostenibilità agroambientale ed economica. Questo studio propone un modello teorico per la costruzione di un indice sintetico partendo da 22 indicatori base declinati in tre Domini: 1) Densità infrastrutturale e flussi turistici; 2) Impatto economico del settore turistico; 3)*

Antonella Bianchino
Istat; e-mail: bianchin@istat.it

Daniela Fusco
Istat; e-mail: dafusco@istat.it

Paola Giordano
Istat; e-mail: pgiordano@istat.it

Maria A. Liguori
Istat; e-mail: liguori@istat.it

Donato Summa
Istat; e-mail: donato.summa@istat.it

Sostegno agricolo. L' approccio multi-fonte (fonti statistiche e amministrative, big data), risponde all'esigenza di sviluppare un sistema di statistiche omogenee, confrontabili e aggiornate. Il focus è realizzato sui comuni appartenenti alla Strategia Nazionale per le Aree Interne (SNAI), dove il settore agricolo svolge un ruolo centrale.

Key words: Big data, tourism, Inner Areas, multi-source.

1 Introduction

As registered by the World Tourism Organization (WTO), tourism connected to food and countryside is a rapidly growing segment. A high percentage of tourists choose the tourist destination based on quality food and the need to gain positive experiences by tasting local products in small locations, renowned for their quality and their strong connection to the territory [7].

Typical local products, in Italy, represent a main component of the endogenous development of territorial systems, due to the meaningful economic, social and tourist impacts they can produce. Agri-food production, although consisting of specialties, is in many cases a necessary, but not a sufficient development condition: so, it must be combined with other services, such as the entire cultural heritage of which a certain local system is provided, for the production of the so-called integrated specialties [3].

Italy's National Strategy for Inner Areas (SNAI) is an innovative policy for development and territorial cohesion to counteract marginalization and demographic decline within IAs throughout the Country [1].

According to SNAI, in 2020, the 7,903 Italian municipalities are classified in 7 categories, from Urban Pole (182) to Ultra-peripheral ones (382). IAs include Peripheral and Ultra-peripheral areas (total 1,906), fragile territories with a far "distance" from essential services. In the IAs, the agricultural, pastoral and forestry sectors play a central role as opportunities for economic growth and for the value of care and environmental prevention [5].

The aim of the work is to represent the complexity of rural tourism in IAs, compared with Urban Poles and "Belt" municipalities, analysing main components and driving forces, using a multi-source approach. The study identifies useful indicators for the evaluation of these phenomena by exploiting the opportunity given by using both Big data and traditional sources. The result could provide a tool for monitoring tourism sector by stakeholders and policy makers.

2 Methodology

Many National Statistical Institutes (NSIs), especially in Europe, are moving from single source statistics to multi-source statistics. This is due to higher quality demands on the statistics produced: more detailed data, more timely data and a general request for a quicker response from NSIs to societal events. Additionally, many NSIs face budget cuts that make large-scale surveys too costly to set up and maintain. NSIs have traditionally produced single-source statistics, where ultimately only data from a single data source is used.

Other data sources are also often used in this process, but only as ancillary data, for example to calibrate or improve the estimates, or as supplementary data to validate the statistics produced. In most cases, the single data sources are surveys, although nowadays administrative data are increasingly used as unique data sources and Big Data are starting to be used. By combining survey data with already available administrative data and Big Data, NSIs can reduce data collection and processing costs and reduce the burden on respondents [4]. As a result, in this study, multi-source approach combines 8 different data sources. Data are used to build 22 basic indicators.

2.1 The proposed model: identification of basic indicators

Following the OECD [6] indication for the composite indicator construction, the definition should give the reader a clear sense of what is being measured by the indicator. It should refer to the theoretical framework, linking various sub-groups and the underlying indicators.

For the assessment of driving forces that affect tourism in IAs, this study proposes a theoretical model for building a synthetic index starting from 22 basic indicators declined in three Pillars: 1) Infrastructural density and touristic fluxes; 2) Economical impact of touristic sector and 3) Agricultural sector support. The estimation of a complex phenomenon as tourism, with the use of a synthetic index, summarizes the concept at the highest levels, leaving little space to the analysis of the individual facets, but represents a photograph of the phenomenon, useful for the evaluation of touristic and agricultural *ex post* policies.

Indicators are described in table 1.

For the construction of the indicators, the following sources have been used:

- Capacity of collective accommodation establishment survey, ISTAT. (https://www.istat.it/it/archivio/216328). Indicators 1-5.
- The permanent Census of the population and housing, ISTAT (https://www.istat.it/it/censimenti/popolazione-e-abitazioni). Indicators 1-3.
- Survey of museums and similar institutions, ISTAT and Ministry of Culture (MiC), (https://www.istat.it/it/archivio/6656). Indicator 6.
- Statistical Register of Active Enterprises (ASIA), ISTAT (http://www.istat.it/it/archivio/216767). Indicators 7, 9 and 10.
- FRAME SBS, ISTAT (https://www.istat.it/it/archivio/249448). Indicator 8.

- Big Data on Agritourisms [2] from http://www.agriturismoitalia.gov.it. Indicators 11-15.
- Statistical Atlas of the Municipalities, ISTAT and SISTAN (https://asc.istat.it/ASC/). Indicators 4, 11 and 12.
- General Agricultural Census 2020, ISTAT. (https://www.istat.it/it/censimenti/agricoltura/7-censimento-generale). Indicators 16-22.

Regarding to Agritourisms, usually, the main source of information for these farms is represented by the administrative data of the authorised Agritourisms, collected both by Istat and by Ministry of Agriculture in different surveys. To enhance the data quality (i.e. timeliness, accuracy, punctuality) and to increase the amount of information related, in this work was proposed the use of web scraping techniques. Web scraping is a process of extracting data from websites using software programs; it has become an increasingly popular technique in recent years due to the growth of the internet and the large amount of data available online.

There are two main types of web scraping: specific and generic web scraping. Specific web scraping involves scraping websites where the structure and content are known in advance. This means that the software program can be programmed to replicate the behaviour of a human user visiting the website and extract only the relevant information from the website. Generic web scraping, on the other hand, involves scraping websites where no prior knowledge of the content is available. This means that the software program must scrape the entire website and then use machine learning or other techniques to infer relevant information.

In this work, the web data acquisition focuses on specific web scraping. Custom software programs have been developed to extract information from the website http://www.agriturismoitalia.gov.it, which is the official website for Italian agritourisms, including about 25,000 units. Specifically, it was scraped the list of Italian official Agritourisms and detailed information on each enterprise was scraped from the dedicated pages. The extracted data were stored in a tabular data format and automatically processed using a Python script to produce the dataset used for analysis in this study.

Table 1 *Pillars, Basic indicators ad related Algorithms.*

Pillar	***Basic indicator***	***Algorithm***
Infrastructural density and touristic fluxes	Total accommodation rate	Number of total beds per 1,000 inhabitants
	Rate of accommodation of high-end hotel structures	4–5-star hotel beds per 1,000 inhabitants
	Rate of accommodation of extra-hotel facilities	Extra-hotel beds per 1,000 inhabitants
	Density of establishments, hotels and accommodation facilities	Total beds per km^2
	Incidence of accommodation at municipality level	Total beds of the Municipality/Total beds at national level (%)
	Visitor pressure on museum and similar institutions per inhabitant	Visitors of museum and similar institutions per 1,000 inhabitants

Economic impact of touristic sector	Incidence of employment in the tourism sector	Employees of tourist Local Unit/Total employees of Local Units of Municipality
	Added value per capita of the tourism sector	Added value of tourist Local Units per inhabitant
	Incidence of employment in the tourism-related entertainment sector	Employees of tourism-related entertainment Local Units/Total employees of Local Units of Municipality
	Localisation quotient of local tourist unit employees	[Employees of the Tourist Local Unit of Municipality/ Employees of the total Local Unit of Municipality]/[Employees of the Tourist Local Unit Italy/ Employees of the total Local Unit Italy]
Pillar	***Basic indicator***	***Algorithm***
Agricultural sector support	Density of agritourisms	Agritourisms per km^2
	Density of agritourisms with accomodation	Beds of agritourisms per km^2
	Share of agritourisms with catering services	Agritourisms with catering services/Total agritourisms
	Share of agritourisms with direct sale	Agritourisms with direct sale/Total agritourisms
	Share of agritourisms with other gainful activities except direct sale	Agritourisms with other gainful activities except direct sale/Total agritourisms
	Share of holdings with grapes for PDO/PGI wines	Holdings with grapes for PDO/PGI wines/Total holdings with vineyard
	Share of holdings with organic farming UAA	Holdings with organic farming UAA/Total holdings with UAA
	Share of holdings with organic farming livestock	Holdings with organic farming livestock/Total holdings with livestock
	Share of holdings with wooded area	Holdings with wooded area/Total holdings
	Share of holdings with permanent crops	Holdings with permanent crops/Total holdings with UAA
	Share of holdings with permanent pasture and meadows	Holdings with permanent pasture and meadows/Total holdings with UAA
	Share of holdings with short rotation coppices	Holdings with short rotation coppices/Total holdings

3 Results and final remarks

Using the algorithms indicated in table 1, the simple indicators have been constructed. As regards the “Economic impact of the tourism sector” Pillar, for the indicators relating to the tourism sector, the following classifications of economic activities have been considered: accommodation, catering service activities, travel agency, tour operator and other booking and related activities; for the entertainment sector have been considered: creative, artistic and entertainment activities, libraries, archives, museums and other cultural activities, gambling and betting activities, sports activities and entertainment and recreation activities.

The use of Big Data required an important pre-processing data, all the activities related to data preparation for further analysis: data preparation, translation, data cleaning and text processing tasks.

Specifically, it was verified that: all the municipalities existed in the list of Italian municipalities, for those that did not exist the correct municipality was identified with manual techniques; duplicates have been removed; in the case of homonymous municipalities, the correct province has been assigned starting from the URL of the Agritourism. The rural tourism of Inner Areas, to a certain point, depends on the agricultural sector. The construction of a new offer (diversification of activities) and the construction of an integrated and organized offer of high-typical goods and services (specialties and integrated specialties) in synergy with tourism is necessary for the sector development. The decline in agricultural and other forms of rural employment in many countries has created a need for a diversified range of rural businesses. In most cases, rural tourism has become an important element of the diverse activities and development in rural areas. Switching from single source to multi-source statistics therefore seems like the way to go. However, this transition is not easy. Multi-source statistics come with new problems that need to be overcome before the resulting output quality is sufficiently high and before those statistics can be produced efficiently. What complicates the production of multi-source statistics is that supporting data come in many different varieties as data sets can be combined in many different ways. Every variety seems to come with its own problems for which tailor-made solutions are needed. It often feels like for every new multi-source statistics one has to reinvent the wheel [4].

References

1. Agenzia per la coesione territoriale: Le aree interne: di quale territori parliamo? Nota esplicativa sul metodo di classificazione delle aree. Methodological note. Cohesion Policy Department DPS (2021)
2. Barcaroli G., Fusco D., Giordano P., Greco M., Moretti V., Righi P., Scarnò M.: ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms. International Conference on Agricultural Statistics. ICAS VII Rome 2016, FAO, 26-28 October. (2016)
3. Becattini G., Zorini L. O.: Identità locali e globalizzazione. In: QA - Rivista dell'Associazione Rossi-Doria. - Associazione per studi e ricerche Manlio Rossi-Doria. - 2003, 1, 0134 (2003)
4. de Waal T., van Delden A., Scholtus S.: Multi-source statistics: Basic situations and methods. International Statistical Review, 88(1), 203-228 (2019) https://doi.org/10.1111/insr.12352.
5. Lucatelli S., Storti D.: La strategia nazionale aree interne e lo sviluppo rurale: scelte operate e criticità incontrate in vista del post 2020. In: Agriregionieuropa anno 15 n°56, Mar 2019 (2019)
6. Organization for Economic Co-operation and Developement (OECD): Handbook on Constructing Composite Indicators Methodology and user guide. JRC European commission (2008)
7. Piñeiro M. V., de Salvo P., Giommi F.: Rural Tourism and Territorial Development in Italy. In: Sustainability Assessment at the 21st century, 198 pp. (2019) ISBN: 978-1-78984-977-6

Statistical analysis of tourism sustainability in Campania: post Covid-19 review

Analisi statistica della sostenibilità turistica in Campania: revisione post Covid-19

Massimiliano Giacalone, Vincenzo Basile and Marco Bellucci

Abstract Based on the study of Basile and Giacalone which analyzed sustainable tourism in the Campania Region by statistical analysis for the development of tourist destinations, we propose an analysis of the tourism sustainability in the Campania region comparing the data for 2021, first with 2015 and then with 2020. The research is developed by producing the results of indicators that will be useful in developing the global tourism indicator that will be compared to the results obtained in the three years taken into consideration to analyze the trend over time. From this, we can deduce the final purpose of this research is to evaluate the post-pandemic effects due to Covid-19 without losing the focus on the analysis and research of the most sustainable provinces in the tourism industry.
Abstract *Sulla base dello studio di Basile e Giacalone che ha analizzato il turismo sostenibile nella Regione Campania attraverso un'analisi statistica per lo sviluppo delle destinazioni turistiche, proponiamo un'analisi della sostenibilità del turismo in Campania confrontando i dati del 2021, prima con quelli del 2015 e poi con quelli del 2020. La ricerca si sviluppa producendo i risultati degli indicatori utili a sviluppare l'indicatore globale del turismo che verrà confrontato con i risultati ottenuti nei tre anni presi in considerazione per analizzarne l'andamento nel tempo. Da ciò si deduce lo scopo finale di questa ricerca che è quello di valutare gli effetti post-pandemici dovuti al Covid-19 senza perdere il focus sull'analisi e la ricerca delle province più sostenibili nel settore turistico.*

Key words: sustainability, Covid-19, sustainable tourism, sustainable development, tourism education, Campania

---

Massimiliano Giacalone
Department of Economics, University of Campania Luigi Vanvitelli, Capua (CE), e-mail: massimiliano.giacalone@unicampania.it

Vincenzo Basile
Department of Economics, Management, Institutions (DEMI), University of Naples Federico II, Naples, e-mail: vincenzo.basile2@unina.it

Marco Bellucci
International Department, École de Management de Normandie, Le Havre France, e-mail: marco.bellucci05@em-normandie.fr

1 Introduction

In 1987, the World Commission on the Environment presented an original definition of environmental sustainability, which was the one most widely accepted by experts. The description read, “Sustainable development is the process of change in which financial resources are exploited, investments are directed, technological developments are exposed, and institutional changes are implemented to meet current and unborn needs”. On that date, the concept of “sustainable development” was born, as it was realized that many raw materials were being consumed that would become scarce in the future and lead to conflicts over their acquisition.

Over time, Agenda 21 was promoted, in 1992 [8], aimed at integrating the goals of environmental quality, social welfare, and economic development within overarching programs [13] and then moving on to the Tokyo Protocol in 2005, which sought to set a target, specifically to reduce greenhouse gas emissions by 5 per cent between 2008 and 2012, a target that was not met. Moving on to the European front, it was decided that facilities should be “near-zero energy facilities”, employing the use of renewable energy sources from neighboring areas [9], until the present day with the establishment of the Sustainable Development Goals (SDGs), a set of 17 social, economic, and environmental claims to be achieved by 2030. After a brief literal appendix on the concept of environmental sustainability, it is time to introduce the concept of sustainable tourism with respect to which there has been a growing interest over time to the point that many scholars have devoted themselves to the content giving their input [4, 7].

“Sustainable tourism is tourism that is economically, socially, and culturally sustainable and whose socio-cultural and environmental impacts are neither permanent nor irreversible” [6, 15], this is the most common definition of sustainable tourism, which is not just a passing fad, but with the rise of ecotourism is becoming a real journey and lifestyle. As for us, we will address the content in the following way: in paragraph 2 we will present the indications that we will need to calculate the ”Global tourism index G” (2) (Tab.1), in paragraph 3 we will present the results based on the comparison between 2015, 2020 and 2021 while in paragraph 4 we will propose the conclusions for the anatomized regions with some recommendations from the directional point of view for them.

2 Data and proposed model

The analysis of the sustainability of tourism in Campania was carried out by analyzing the general aspects of the tourism sector of the individual provincial territories subject to analysis by ISTAT (dati.istat.it). These data are aimed at the construction of synthetic statistical indices that allow to specifically evaluate the main characteristics of the various territories and build a global index to be used to build a ranking of the various territories based on tourism sustainability to be compared with the previous ranking identified in 2016.

To evaluate the interrelationships between tourism and territory we will use a series of synthetic indices present in the literature [10, 11, 12, 14]. Considering the three dimensions of sustainability, the indices can be divided into three categories: indices which refer to the economic dimension, indices which refer to the tourism dimension and indices which refer to the environmental dimension. The indices that refer to the tourist dimension are the first three, the four following indices after referring to the economic dimension and the last indices refer to the environmental dimension: Where L_{ci} represents the total beds of hotel category C present in the

Table 1 The indicators

Indexes	Formulas
Attractiveness index	$A = \Sigma_c [(L_{ci}/L_{ct}) \cdot C \cdot 100]$
Territorial density index of tourism	$D = PL/S$
Defert's tourism function rate	$F = (PL/P) \cdot 1000$
Gross utilization or pressure index at time t	$TP = (Pr/PL \cdot Gt) \cdot 100$
Tourism relations index	$T = (P/Pr) \cdot 100$
Florence settlement index	$SI = (PL/PL_{tot})/(P/PL_{tot})$
Lundgren's tourism connotation index	$L = (P/N_{es})$
Territorial employment index	$S = (AR+P)/S$
Index of urban waste production	$W_{tu} = (R_{tot}/P) – [R_{tot}/(P+PL)]$
Wtur integrated by the gross use index	$W'_{tur} = (R_{tot}/P) – [R_{tot}/(P+PL \cdot TP)]$

i-th locality, L_{ct} indicates the total beds of hotel category C present in the (regional) complex of the investigated area, $C = 1$ is for 5-star category hotels, $C = 0.8$ is for 4-star category hotels, $C = 0.6$ is for 3-star category hotels, $C = 0.4$ is for 2-star category hotels, $C = 0.2$ is for 1-star category hotels, PL indicates the total beds present in the territory, S indicates the extension of the area under study, P indicates the population of the area studied, Pr indicates the total presences in the territory at time t, Gt indicates days in period t (2020 is leap year), AR indicates the tourist arrivals in the area in question, R total indicates total waste, PL_{tot} indicates the number of bed places in the entire region and N_{es} indicates the number of accommodation establishments in the chosen territorial area.

The attractiveness index is the index used to evaluate the quality of accommodation facilities and the higher the value of the latter, the higher the attractiveness of the territory, the territorial density index of tourism evaluates the impact of tourism on the physical territory, the Defert tourist function rate describes the impact of tourism on the resident population and gives information on the capacity to absorb tourism in demographic terms, the gross utilization index evaluates the pressure exerted on tourist structures and territories in different parts of the year, the tourism relations index describes the impact of tourism on resident natural persons therefore low values of the index are significant of high fruition between the two variables considered, the territorial employment index has the for the purpose of evaluating the use of the territory by man (resident population and tourists), the urban waste production index is integrated by the gross use index (TP) in order to provide an estimate of the urban waste production attributable to the sector receptive, the Florence settlement index relates the tourist accommodation offer with the resident population and finally, the Lundgren tourist connotation index aims to express the tourist connotation of a given area. All the indices indicated have their own value and their own specificity; therefore, to be summarized in the global tourism indicator G in Eq. (2), they must be normalized by assigning one hundred points to the territories that have obtained the best scores and a proportional score to the others according to the following two formulas:

$$Tx_i = (\min / x_i) \times 100 \text{ or } Tx_i = (x_i / \max) \times 100 \tag{1}$$

where x_i is the index value for teach territory and Tx is the index value normalized with one of the two formulas, chosen from time to time according to the characteristics of the index. The average of all the indices taken into consideration constitutes the composite indicator which will synthetically describe the sustainability characteristics of tourism in the various territories. To better balance the three different components, social, economic, and environmental, the global tourism indicator G in Eq. (2), inspired by the statistical literature, was constructed in a weighted way using the indices already introduced in the present work, obtaining the following formula:

$$G = \frac{D + TP + A + \%}{12} + \frac{F + AS + S + SI + L}{15} + \frac{T + W'_{tur}}{6}. \tag{2}$$

The weighting considers the three areas that define the global tourism indicator G: environmental, economic and social. Each was given a weight of 1/3, distributing it among the transforms of the simple indicators that define each area, (e.g.: two simple indicators act in the social sector, so the "social" transforms were given a weight of 1/6), inside the tables we find in the last column the rankings of the territories considered.

Table 2 The global tourism indicator G (2020)

District	SI	L	AS	A	D	F	TP	S	W'TUR	%	T	total	G	Score
Naples	28.8	42.5	76.3	100.0	1.5	71.0	53.6	4.0	18.7	68	59.2	523.5	46.41	4
Salerno	43.8	17.4	100.0	46.3	11.6	100.0	48.5	43.4	15.0	92	100.0	617.5	55.96	1
Caserta	100.0	100.0	72.1	6.4	29.2	25.3	41.6	35.2	33.8	74	29.4	547.2	45.33	5
Avellino	11.2	51.8	54.7	4.4	65.5	26.4	100.0	87.9	95.8	90	12.7	600.6	55.23	3
Benevento	7.5	23.3	52.2	2.2	100.0	19.1	85.4	100.0	100.0	100	11.0	599.7	55.85	2

3 Results

It is appropriate to think about the possible management implications considering the above: Campania tourism is concentrated almost entirely in the provinces of

Table 3 The global tourism indicator *G* (2021)

District	SI	L	AS	A	D	F	TP	S	W'TUR	%	T	total	G	Score
Naples	26.0	42.1	75.6	100.0	1.5	64.8	40.4	3.4	12.9	70	63.5	500.5	44.53	4
Salerno	17.1	100.0	100.0	46.3	11.8	100.0	39.5	38.2	11.2	92	100.0	656.4	58.04	2
Caserta	75.0	16.7	76.9	6.4	30.1	25.0	31.2	32.8	22.4	74	31.6	423.0	35.98	5
Avellino	66.6	33.4	54.6	4.4	66.5	26.4	84.4	85.8	64.7	89	12.7	589.2	51.10	3
Benevento	100.0	71.1	47.1	2.2	100.0	19.6	100.0	100.0	100.0	100	7.8	748.0	65.69	1

Naples and Salerno, which for this reason are the main destinations in Campania, we also remember that they are the main seaside resorts as well as cities of art that are visited.

Considering the global tourism indicator *G*, which measures the sustainability of tourism according to statistical parameters, Naples ranks both in 2015 and 2020 as well as in 2021 as one of the worst despite high scores in some indices, such as that of attractiveness, fails to reach acceptable levels in differentiated waste collection, because there are objective difficulties in managing the large tourist flows on which improvement still needs to be made, despite the efforts of the Region.

A possible strategy to implement to improve the situation could be a further action to raise awareness of the differentiated waste collection and the implementation of special bins, to make the city more in step with the times and therefore fill this gap with the other provinces. Benevento, as already mentioned, is the best thanks to the very high values in the analysis of waste production weighed on occupied tourist beds. However, its share of regional tourism is 0.5% of the total regional tourism, so it is still conceivable that such small-scale tourism is sustainable. Even Avellino, which unfortunately has low tourist flows. In this case, however, it would be necessary to make the territory more attractive to increase the percentage of regional tourism, perhaps through actions aimed at creating a "territorial brand", all supported by the development of some functional infrastructures where necessary. Salerno is first in 2020 and second in 2021 according to what the global tourism indicator *G* shows. These positions are obtained thanks to its good attractiveness, the better functioning of separate collection and the low weighted percentage of waste produced. A city that therefore manages to represent the concept of sustainability better than the others, considering its tourist flows almost on a par with Naples.

4 Conclusions

One potential positive impact of the pandemic on sustainable tourism in Campania is the increased awareness of the importance of responsible and sustainable tourism. With fewer tourists, there is an opportunity to focus on promoting sustainable practices such as reducing waste, conserving natural resources, and supporting local businesses. Post- Covid-19 [3], the tourism industry in Campania will need to continue to adapt to the new normal and implement sustainable practices to promote long-term destination development. This includes developing sustainable tourism products, such as eco-tourism and cultural tours that promote local traditions and heritage. Overall, the Covid-19 pandemic has had a significant impact on the tourism industry in the Campania region and consequently, the behaviour patterns of consumers have changed after this crisis [2]. However, it has also provided an opportunity for sustainable tourism development and the promotion of responsible tourism practices. By working together, local communities, businesses and policymakers can ensure a sustainable and resilient tourism industry in the region.

To conclude, it must be said that although Covid-19 has had a significant impact on attendance and arrivals in the places taken into consideration, the results (in proportion to the Covid-19 pandemic) are almost similar; therefore, the managerial advice would have been useful and the same also for the year 2020. To provide a contribution to future research, we believe that the work presented here, concerning the global tourism indicator G in Eq. (2) [1, 4], could in the future be subjected to a principal component analysis, in particular to the Spatial principal component analysis [5] which could lead to applying the global tourism indicator G to the principal components and making it much more efficient.

References

1. Basile, V. and Giacalone, M.: Sustainable tourism in Campania region: statistical analysis and metrics for the development of tourist destinations. Journal of Applied Quantitative Methods, volume 12, Issue - 3, September 30. (2017) - ISSN 1842- 4562
2. Basile, V., Caboni, F. and Pizzichini, L.: The New Profile of the Online Consumer Behaviour in a Post-Pandemic World. In Handbook of Research on Global Networking Post Covid-19 (pp. 38-54). IGI Global (2022)
3. Chang, CL, McAleer, M., and Ramos, V.: A charter for sustainable tourism after Covid-19. Sustainability, 12(9), 3671 (2020)
4. Butler, R.: The evolution of tourism and tourism research. Tourism Recreation Research, 40(1), 16-27 (2015)
5. Giacalone M., La Tona L., Marino C.: The analysis of the sustainability of tourism in Sicily through a global indicator. In: Tourism and territory: empirical analyzes and methodological approaches. p. 209-218, Milan, McGraw-Hill (2012)
6. Giacalone, M., Mattera, R., and Nissi, E.: Well-being analysis of Italian provinces with spatial principal components. Socio-Economic Planning Sciences, 84, 101377 (2022)
7. Harris, R., Williams, P. and Griffin, T.: Sustainable tourism. Routledge (2012)
8. Mihalic, T.: Sustainable-responsible tourism discourse–Towards ' responsustable'tourism. Journal of cleaner production, 111, 461-470 (2016)
9. Moallemi, EA, Malekpour, S., Hadjikakou, M., Raven, R., Szetey, K., Moghadam, MM,. . . , and Bryan, B.A.: Local Agenda 2030 for sustainable development. The Lancet Planetary Health, 3(6), e240-e241 (2019)
10. Mowforth, M., and Munt, I.: Tourism and sustainability: Development, globalization and new tourism in the third world. Routledge (2015)
11. Palacios-Florencio, B., Santos-Roldán, L., Berbel-Pineda, J.M. and Castillo- Canalejo, A.M.: Sustainable Tourism as a Driving force of the Tourism Industry in a Post-Covid-19 Scenario. Social indicators research, 158(3), 991-1011 (2021)
12. Pulido-Fernandez, J.I. and Pulido-Fernandez, M.D.L.C.: Proposal for an indicators system of tourism governance at tourism destination level. Social Indicators Research, 137, 695-743 (2017)

13. Punzo, G., Trunfio, M., Castellano, R. and Buonocore, M.: A multi-modelling approach for assessing sustainable tourism. Social Indicators Research, 163(3), 1399- 1443 (2022)
14. Sajjad, A. and Shahbaz, W.: Mindfulness and social sustainability: An integrative review. Social Indicators Research, 150(1), 73-94 (2020)
15. Torres-Delgado, A. and Saarinen, J.: Using indicators to assess sustainable tourism development: a review. Tourism Geography, 16(1), 31-47 (2014)
16. Weaver, D. B.: Sustainable tourism. In Encyclopedia of Tourism Management and Marketing (pp. 317-321). Edward Elgar Publishing (2022)

Investigating recent changes in dietary behavior

Un’analisi dei recenti cambiamenti nel comportamento alimentare

Angela M. D’Uggento, Francesco D. d’Ovidio, Ernesto Toma and Maria Giovanna Onorati

Abstract: Recent disruptive events such as the Covid19 pandemic, the shaking of the global economic balance due to the war in Ukraine and climate change have reinforced the process of "deglobalization" of food consumption, with the consequence that consumers prefer local products. To understand the change in sustainable food consumption in some countries, this paper analyzes data from a survey conducted in 2020-2022, during the Covid19 pandemic, with the aim of defining a "new normal" dietary ethos that results from a preference for regional cuisine, environmental protection and a commitment to health. To explore the role of cultural contexts in consumption patterns with different but comparable values and principles, the analysis provides a comparison between Italy and the United States.

Abstract: *I recenti eventi disastrosi quali la pandemia da Covid19, lo stravolgimento dell’equilibrio economico globale causato dalla guerra in Ucraina e i cambiamenti climatici hanno rafforzato il processo di “deglobalizzazione” dei consumi alimentari, con la conseguenza che i consumatori preferiscono i prodotti locali. Per comprendere le scelte verso consumi alimentari sostenibili in alcuni Paesi, il presente lavoro studia i dati di un’indagine condotta dal 2020 al 2022 con l’obiettivo di definire le*

Angela M. D’Uggento
University of Bari *Aldo Moro* (IT); e-mail: angelamaria.duggento@uniba.it

Francesco D. d’Ovidio
University of Bari *Aldo Moro* (IT); e-mail: francescodomenico.dovidio@uniba.it

Ernesto Toma
University of Bari *Aldo Moro* (IT); e-mail: ernesto.toma@uniba.it

Maria Giovanna Onorati
University of Gastronomic Sciences, Pollenzo (IT); e-mail: m.onorati@unisg.it

This paper comes from a common work, but M.G. Onorati wrote Sections 1-2 and designed the framework of the survey, A.M. D'Uggento Section 3, F.D. d'Ovidio Section 4 and E.Toma wrote Section 5

caratteristiche di una nuova etica alimentare generata dalla preferenza per la cucina regionale, la tutela dell’ambiente e della salute. Per esplorare il ruolo dei contesti culturali nei modelli di consumo con valori e principi diversi ma comparabili, l’analisi propone un confronto tra Italia e Stati Uniti.

Key words: food sustainable consumption, survey, structural equations modeling.

1 Introduction

The need to adopt sustainable behaviors to reduce the waste of resources and protect the planet now affects all areas of individuals' daily lives, and food is no exception.

Two drastic recent events, the Covid19 pandemic and the unexpected war at the gates of Europe, have accelerated the change in consumer behavior regarding food and related activities such as shopping, cooking, and eating. Environmental protection, safety and healthy lifestyles are the pillars of a global trend called the "new normal" which is spreading in response to the sense of loss and insecurity created by the pandemic, the economic crisis, and extreme events caused by climate change. The increase in food prices, as well as the prices of commodities such as energy and gas, caused by these disruptive events, leads to more and more people living in poverty, resulting in famine and malnutrition of people around the world [6]. In response to this situation, many countries have implemented export restrictions to protect domestic food supplies, leading to a preference for "eating local products". Low-mileage consumer goods and short supply chains are becoming the new trend in healthy eating, which celebrates sustainable food as fresh, less processed, and locally produced. However, local food policies could be the precondition for gastronationalism, a form of nationalism that proclaims the homeland on the plate every day. Among the population, concern about the harmful effects of a globalization responsible for the pandemic and the global crisis generates a form of protection that is also transferred to food. Local food becomes a form of resistance to the homogenizing tendencies of globalization. In Italy, this phenomenon can be called "food patriotism," based on a preference for products that people consider safer and more reliable than food that comes from outside. This phenomenon is part of a larger deglobalization movement of eating habits that has already been observed in Italian consumer behavior in recent years, expressed in the preference for "Made in Italy."

The "nationalist shift" has also occurred rapidly in America, where greater awareness of health protection spread in 2022 and, despite the economic recession and commodity crisis, a new demand for healthier foods emerged that Italian food exports, especially fresh vegetables and organic foods, could satisfy. In line with these global trends, this paper presents the results of a research launched in 2020 with the aim of understanding the transformation of sustainable food consumption in some countries since the outbreak of the pandemic. The goal was to define a "new normal" culinary ethos in which preference for regional cuisine, environmental awareness and commitment to safety are critical themes. To explore the role of cultural contexts in consumption patterns this paper compares data collected in Italy and the United States.

2 The survey

A questionnaire consisting of 26 main questions (but more than 80 items, mainly in ordinal scale) was completed online from October 2020 to July 2022 in 20 countries using the CAWI system Qualtrics. 6.906 questionnaires were collected in the following participating countries: Italy, USA, Colombia, Germany, Israel, Taiwan, Japan, France, India, Brazil, South Africa, Canada, Netherlands, Lebanon, Poland, Greece, Switzerland, United Kingdom, Austria, Russia, and Mexico. Data were collected using the snowball technique, which is a non-probability sampling method based on the social relationships of the first respondents to recruit further respondents, following a multi-stage selection procedure designed by the researcher [2]. Snowball sampling is often used for exploratory analyses; however, the final sample the sample has a non-programmable size, being self-formed, and could be influenced by over-representation of certain groups, especially people who are more likely to cooperate or belong to recruiters' networks. In our survey, the total sample was composed of 73.3% women and 26.1% men. This high prevalence can be considered dependent on the topic addressed in the survey, so the total sample represents specifically the statistical domain of individuals interested in dietary characteristics and behaviors.

3 Understanding dietary behavior

To investigate this phenomenon, we decided to compare the choice of the two larger subsamples: Italy and USA, which had a similar gender distribution. However, their distributions by age differ significantly: in the Italian subsample, young people under 25 are abundant and older people are scarce, while the opposite is true in the US subsample, where women over 65 are three times more abundant than in Italy. For this reason, post-stratified sampling was performed in the Italian sample to obtain a smaller subsample whose age distribution is more comparable to that of the U.S. subsample. Female U.S. respondents over 65 years of age were also re-sampled.

Therefore, this paper analyzes the questionnaires of 819 respondents (180 M, 630 F, and 9 ND) from Italy and 786 (188 M, 589 F, and 9 ND) from the United States (Table 1). The aim is to investigate differences in food consumption behavior during the pandemic, in relation to food "Made in Italy" and fresh, local or sustainable food.

Table 1 shows also the sociodemographic characteristics of the two subsamples. The randomized post-stratification, sacrificing more than 70% of the questionnaires collected in Italy and less than 10% of the questionnaires collected in the United States, showed a very narrow distribution. The lower level of education in the Italian sample is not a novelty, but the difference in the percentage of respondents with education above high school is very large. The lower income reported by Italian respondents is also not surprising (except for the higher percentage of respondents who prefer not to answer, even among freelancers, traders and executives).

To understand the underlying relationships between the observed variables and to trace the latent patterns behind the observed behaviors, we conducted an exploratory factor analysis for both linked subsamples, with extraction via CatPCA [7], since most of the observed variables were merely ordinal, and some socio-demographic variables were nominal. Very strict criteria had to be applied for the extraction of the factors: eigenvalues not less than 1.1 and backward stepwise removal of less correlated variables [1], i.e. variables with an explained commonality of less than 60%. Moreover, the real question of the research is: "Do the domains of Italian and U.S. respondents correspond in terms of their food consumption behavior?"

Table 1 Sociodemographic characteristics of the Italy-USA subsamples

	ITALY				**USA**			
Gender:	**Male**	**Female**	**N.D.**	***Total***	**Male**	**Female**	**N.D.**	***Total***
Gender percentages	22.0	76.9	1.1	*100.0*	23.9	74.9	1.1	*100.0*
Generation (age range)								
GenZ (18-25)	8.9	15.1	55.6	14.2	8.5	14.1	55.6	13.2
Millennials (26-41)	49.5	42.4	33.3	43.8	51.6	42.6	33.3	44.7
X Gen (42-57)	22.2	25.2	11.1	24.4	19.2	25.0	11.1	23.4
Baby Boomers and older (58+)	19.4	17.3	-	17.6	20.7	18.3	-	18.7
Education Level								
Less than high school	6.1	7.9	11.1	7.6	0.5	0.3	11.1	0.5
High school graduate	32.2	31.8	44.5	32.0	11.7	8.5	22.2	9.4
Post-high school/Bachelor	23.9	23.8	22.2	23.8	53.2	58.4	55.6	57.1
Master/Graduate Degree	36.1	35.1	22.2	35.2	29.3	26.3	11.1	26.9
Doctorate	1.7	1.4	-	1.5	5.3	6.5	-	6.1
Annual income								
Less than 8.750 €	8.3	12.1	22.2	11.3	3.7	4.1	11.1	4.1
8.751 - 26.250 €	28.4	31.1	11.1	30.3	2.7	7.6	33.4	6.7
26.251 - 43.750 €	23.3	18.9	11.1	19.8	12.8	13.9	-	13.5
43.751-61.250 €	8.3	5.4	11.1	6.1	14.9	16.1	11.1	15.8
More than 61.250 €	7.8	4.0	11.1	4.9	52.6	41.1	22.2	43.6
N.D./Prefer not to say	23.9	28.5	33.4	27.6	13.3	17.2	22.2	16.3
Total	100.0	100.0	100.0	*100.0*	100.0	100.0	100.0	*100.0*

Such a question implies, first of all, that consistent factor structures must be identified for both subsamples (thus reducing the importance of highly local features, seen elsewhere). Thus, we performed a simultaneous exploratory factor analysis (EFA) for each of the two groups (Italy and U.S.) and removed from both models all observed variables that were found to be deficient in at least one of the samples. The result yielded 28 observed variables, with 7 factors explaining 78.5% of the total variability (KMO statistic: 0.90), with no group differences. No demographic variable survived the exclusion procedure. To determine the best factorial solution, an oblique Promax rotation was then performed. EFA almost always grouped those variables belonging to the same set of behaviors named as follows: **F1**="Sustainable food's meaning", **F2**="Buying organic food", **F3**="Relevance for buying food", **F4**="Buying *'made in Italy'* food", **F5**="Safe food for children", **F6**="Buying local food", **F7**="Food's affordability and convenience". Almost 2/3 of the coefficient in the factor correlation matrix are statistically significant and useful for Confirmatory Factor Analysis.

4 Confirmatory Factor Analysis and model recalibration

Starting from EFA findings, Confirmatory Factor Analysis (CFA) was performed by using Lisrel-type Structural Equation Model [3] provided by the R package "lavaan" [4]. First, we built the common theoretical model based on the EFA results. Its discrepancy was estimated using the robust WLS method. Different models were compared and evaluated using the fit measures and modification indices [5]. The best "fully confirmatory" results provide a correlation model following the structure shown in Table 3, where all significant coefficients are highlighted.

Table 2 Structural correlations in the optimal confirmatory model and factors reliability.

Factors	F1	F2	F3	F4	F5	F6	F7	*Reliability*
F1	1				0.096	0.089	0.234	***0.966***
F2		1	0.574	0.413	0.601	0.636		***0.913***
F3			1	0.395	0.371	0.581		***0.884***
F4				1	0.366	0.409		***0.912***
F5					1	0.611		***0.860***
F6						1		***0.782***
F7							1	***0.582***

Despite its excellent goodness indices (Std.RMR=0.050, RMSEA=0.047, GFI=0.98, TLI=0.98, NFI=0.97, RFI=0.97), the reliability of the identified model is not absolute: in fact, the last column of Table 2 shows that reliability of the latent variable F7, measured by Cronbach's Alpha, is far from satisfactory. This is mainly due to the low explanatory power (R^2<0.27) of the item "Why buying food: convenience in preparation". Thus, we removed this variable, integrated the model according to the hints of the modification indices and evaluated the new correlation structure by EFA, following we obtained the best factorial solution, with *six* factors explaining 79.04% of the 26 constituent items (KMO=0.91). By using this more parsimonious model in our SEM confirmatory analysis, we obtained a completely significant correlation structure, highly reliable, as highlighted in the last column of Table 3. The measures of goodness also remain excellent.

Figure 1 graphically shows the model identified, highlighting the parameters with significant differences between Italian and American respondents.

Table 3 Structural correlations in the best confirmatory model.

Factors	F1	F2	F3	F4	F5	F6	*Reliability*
F1	1	0.136	0.134	0.230	0.171	0.164	***0.957***
F2		1	0.545	0.374	0.512	0.611	***0.914***
F3			1	0.244	0.294	0.550	***0.873***
F4				1	0.408	0.448	***0.882***
F5					1	0.650	***0.813***
F6						1	***0.808***

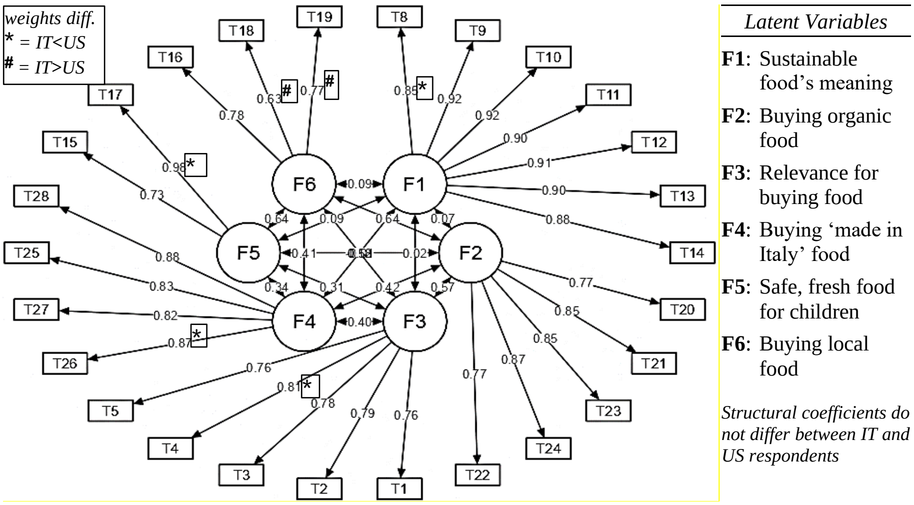


Fig. 1 CFA Optimized model of food consumption behavior and significant differences Italy − USA

Key line of Figure 1

T1. Relevance for buying food: traceability (origin, IGP, etc.)
T2. Relevance for buying food: sustainability (local or seasonal food)
T3. Relevance for buying food: ethics (fair trade, animal welfare, etc.)
T4. Relevance for buying food: naturality (organic, GMO-free, etc.)
T5. Relevance for buying food: eco-friendly packaging
T8. Sustainable food’s meaning: respecting environment
T9. Sustainable food’s meaning: zero waste, 100% reusable
T10. Sustainable food’s meaning: produced locally
T11. Sustainable food’s meaning: positive backfall on the producing community
T12. Sustainable food’s meaning: produced with fair price for the work
T13. Sustainable food’s meaning: produced respecting animal wellbeing
T14. Sustainable food’s meaning: fitting a healthy lifestyle
T15. Why buy fresh food: healthier for children
T16. Why buy local food: safer for me
T17. Why buy local food: safer for the children
T18. Why buy local food: it supports local businesses
T19. Why buy local food: reduces the carbon footprint
T20. Why buy organic food: it tastes better
T21. Why buy organic food: healthier for me
T22. Why buy organic food: healthier for the children
T23. Why buy organic food: animal welfare
T24. Why buy organic food: it protects the environment
T25. Why buy “made in Italy” food: I trust ‘Made in Italy’
T26. Why buy “made in Italy” food: healthy food
T27. Why buy “made in Italy” food: high quality food
T28. Why buy “made in Italy” food: it provides information about origin and supply chain

5 Differences between food consumption behaviors and remarks

This study aims to understand the similarities and differences between current dietary behaviors in the U.S. and IT and the reasons behind them using 28 observed variables that allow us to identify six latent variables. Figure 1 summarizes the results of an equality test of the model's fit coefficients, showing the coefficients for which the null hypothesis was rejected and the direction of the alternative hypothesis was accepted. The six latent variables confirm that new trends in food consumption are increasing worldwide, triggered by the events that disturbed the already existing precarious equilibrium: the pandemic, which drew attention to the importance of health protection, and the war at the gates of Europe, which exposed the weaknesses of dependence on the leading countries for the production of certain products, not only agricultural. Behind this undeniable level related to drastic events, long-standing trends are spreading: on the one hand, the need for "responsible" and prudent food consumption has grown in response to the impoverishment caused by the economic crisis; on the other hand, the awareness of the progressive depletion of the planet's resources has made the adoption of ecological behaviors urgent. The consumption of high-quality, organic, fresh and, therefore, local food is seen in developed countries as an opportunity for a sustainable and healthy future.

References

1. Burnham, K.P. and Anderson, D.R.: Model Selection and Inference: A Practical Information-Theoretic Approach. 2nd Edition, Springer-Verlag, New York (2002)
2. Goodman, L. A.: Snowball Sampling. The Annals of Mathematical Statistics, 32(1): 148-170 (1961)
3. Jöreskog, K. G.: Structural equation models in the social sciences: Specification, estimation and testing. In P. R. Krishnaiah (Ed.), Applications of statistics. Amsterdam: North Holland Publishing Co, 265–286 (1977)
4. Rosseel, Y.: lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2): 1–36 (2012) doi: 10.18637/jss.v048.i02
5. Sörbom, D.: Model modification, Psychometrika, 54: 371-384 (1989)
6. Wesley, E., Peterson, F.: The Coming Global Food Crisis. Cornhusker Economics, June 8. University of Nebraska-Lincoln (2022)
7. Young, F.W., Takane, Y., De Leeuw, J.: The Principal Component of Mixed Measurement Level Multivariate Data: an Alternating Leasts Squares Method with Optimal Scaling Features, Psychometrika, 43: 279-281 (1978)

Depopulation in the Abruzzo municipalities

Lo spopolamento nei comuni abruzzesi

Assunta Lisa Carulli, Domenico Di Spalatro and Alessandro Valentini

Abstract The problem of population decline, the so-called *demographic winter,* is a very relevant question, especially in the internal areas of Italy. To revitalize the socio-economic composition of small mountain municipalities subjected to depopulation, the Abruzzo region in 2021 issued the regional law n.32. The aim of the present paper is to identify the socio-demographic and economic dimensions of depopulation to better explain its nature and to evaluate the potential impact on the local policies. The data base used for the analysis consists of a set of indicators, for each municipality, concerning different thematic areas: demography, society, occupation, economy, services, environment.

Abstract *Il problema del declino demografico, il così detto "inverno demografico", è una questione molto rilevante, specie nelle aree interne dell'Italia. Con l'obiettivo di rivitalizzare il tessuto sociale ed economico dei piccoli comuni montani soggetti a spopolamento la Regione Abruzzo ha emanato nel 2021 la legge regionale n. 32. L'obiettivo del presente lavoro è quello di individuare le dimensioni sociodemografiche ed economiche dello spopolamento al fine di spiegarne meglio la natura e il potenziale impatto sulle politiche locali. Per l'analisi è stato strutturato un data base costituito da un set di indicatori a livello comunale, relativi a diverse aree tematiche: demografia, società, occupazione, economia, servizi, ambiente.*

Key words: depopulation, regional law, mountain municipalities, indicators.

Assunta Lisa Carulli
Istat; e-mail: carulli@istat.it

Domenico Di Spalatro
Istat; e-mail: dispalat@istat.it

Alessandro Valentini
Istat; e-mail: alvalent@istat.it

1 Introduction

For some years Italy has been attending the phenomenon of depopulation, mainly linked to a decline in birth rates: the Total Fertility Rate (TFR) declined from 1.44 to 1.25 in the last 10 years. In the meanwhile, migration balance reduced its anti-decline effects. Furthermore, internal areas are characterized by continuous outflows of population essentially linked to inequalities in the socioeconomic development. Inequalities between the different areas of the country support the attraction of demographic contingents towards urban areas or more economic developed areas and the repulsion towards other less urbanized areas with consequent different possibilities of access to essential or useful services or to more acceptable living and working environments in numerous municipalities. This trend is particularly relevant in the South of Italy and in the various settlements of the Apennines mountains, including municipalities of Abruzzo. In Abruzzo migratory outflows are mainly associated with families of young age and with high working capacity which over the years have led to a pauperization of human capital as the reduction in the birth rate and an accelerated ageing population in the origin places. To contain the drain of population from smaller towns, the region Abruzzo has issued the regional law n. 32/2021 to revitalize the social and economic structure of small mountain municipalities subjected to depopulation and at the same time to promote its recovery and economic development. For the purpose of this law, “small mountain municipalities” have been classified on the basis of UNCEM data (National Union of Municipalities and Mountain Authorities) as: i) with a population of less than 3000 inhabitants in which a demographic decline higher than the regional average has been recorded over the last five years; ii) all mountain municipalities with a population of up to 200 inhabitants even in the absence of demographic decline. Scope of this work is to identify new socio-demographic and economic dimensions of depopulation in Abruzzo and to develop through appropriate statistical analysis in order to support the local policy strategies aimed at contrasting this phenomenon.

2 Data

The information base for the analysis is structured on a set of municipal indicators, classified in 6 different thematic areas: i) demography (5 indicators), ii) society (5 indicators), iii) employment (5indicators), iv) economy (5 indicators), v) services (4 indicators), vi) environment (4 indicators). All data are taken from official statistical sources, mainly Istat. Time period is 2019-2021. The first area (demography) is structured on indicators that highlight both the structural (% of population under 14 years) and the dynamic component (rates of birth, death, migration). Indicators of the second area (society) are of social nature and considers the incidence in the most significant age groups, family composition, educational qualification, commuter flows. Indicators of employment area define the occupational renewal, the weight of NEETs, activity rate and unemployment rate also among young people. A fourth area, the economic one, includes various indicators related to the productive system (average size of local units of enterprises, value added per employee, remuneration per employee), as well as the size and distribution of net income. Indicators of the service area take into account the incidence of accommodation services (hotel and non-hotel) on the territory and those related to the person. Finally, indicators of the sixth area (environment) are more specifically linked to the physical characteristics of the territory: the level of seismicity, the danger from landslides, the hydraulic hazard.

3 Methods

On the time series available (2019-2021), the year of observation was 2019, hence the information set from 2020 onwards is partially incomplete. The set of indicators is synthesized using the Adjusted Mazziotta - Pareto Index (AMPI) [1] in order to build a classification of municipalities on the basis of the greater or lesser fragility generated by the demographic, socio-economic and environmental contexts (Figure 1). The analysis was realized in two steps: i) for mountain municipalities; ii) for all the municipalities of the region. Finally, to measure the spatial autocorrelation of the AMPI index in the 4 provinces of Abruzzo, the Moran index [2] was applied in step ii. This analysis is the basis of other future ones extended to all years taken into account.

4 Results

Figure 1 shows the 22 mountain municipalities (classified under the regional law n.32) with the highest level of AMPI index, which therefore show greater fragility. This ranking has remained unchanged after the inclusion of municipalities not covered by the regional law. The set of indicators is therefore robust and easily explains the phenomenon of depopulation.

Fig. 1 Index AMPI-Abruzzo Mountain municipalities

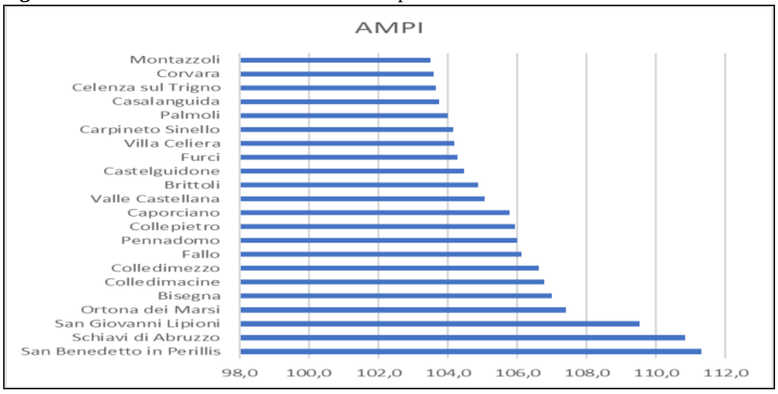


Table 1 shows the Moran index in the provinces of Abruzzo (NUTS 3). The province of Pescara has the highest value of spatial autocorrelation (0.48), followed by the province of L'Aquila (0.26) and that of Chieti (0.25). The lowest level is that of Teramo (0.15).

Table 1 Spatial autocorrelation of AMPI Index

Provinces	Moran Index
Pescara	0.48
L'Aquila	0.26
Chieti	0.25
Teramo	0.15

5 Conclusions

The components of depopulation appropriately identified and summarized through the AMPI index make it easy to identify both the territories subjected to depopulation and the potential ones at the light of a multivariate set of components such as social, economic, and environmental aspects. The values of the spatial autocorrelation of Moran index show a different level of territorial connection of the MPI index in the various provinces of Abruzzo that could imply different dynamics of depopulation or repopulation not still emerged.

Results are part of an information system whose scope is to support local policies aimed at contrasting demographic decline.

References

1. Mazziotta M. and Pareto A: Gli indici sintetici. G.Giappichelli Editore (2022)
2. Cliff A.D. and Ord J.K.: Spatial autocorrelation. London Pion (1973)

The Productions System of Inland Areas

Il Sistema produttivo nei comuni delle Aree Interne

Agata Maria Madia Carucci and Antonio Regano

Abstract The work aims to analyze the production system of homogeneous territories by degree of socio-demographic and economic marginality as defined by the National Strategy of Inland Areas, which has been a national policy of cohesion and development on a territorial basis since 2014. The increasing dissemination of municipal-level data on enterprises has made it possible to analyze the structural characteristics of enterprises in inland areas and the main performance indicators: productivity, profitability and cost structure. The results are presented through box plots, and the significance of these indicators was attested with a logit model.

Abstract *Il lavoro si pone l'obiettivo di analizzare il sistema produttivo di territori omogenei per grado di marginalità socio-demografica ed economica così come sono stati definiti all'interno della Strategia Nazionale delle Aree interne, che dal 2014 rappresenta una politica nazionale di coesione e sviluppo su base territoriale. La sempre maggiore disponibilità di dati a livello comunale sulle imprese ha permesso di caratterizzare gli aspetti strutturali delle imprese delle aree interne e i principali indicatori di performance, quali produttività, redditività e struttura dei costi. I risultati sono presentati attraverso dei box plot e la significatività di tali indicatori è stata testata attraverso un modello logit.*

Key words: inland areas, enterprises, productivity, value added

---

Agata Maria Madia Carucci

Istat, Italian National Institute of Statistics; e-mail: carucci@istat.it

Antonio Regano

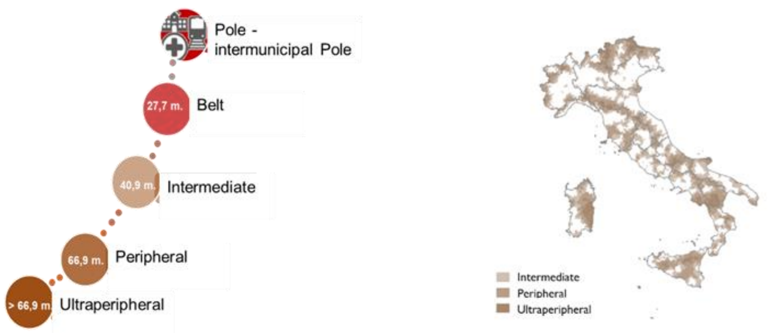
Istat, Italian National Institute of Statistics; e-mail: regano@istat.it

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Italian National Institute of Statistics. The paragraphs were authored as follows: §1, 3 were written by A.M.M Carucci and §2, 4 by A. Regano.

1 National Strategy of Inland Areas

The National Strategy of Inland Areas (SNAI) finds its regulatory reference in the 2014 National Reform Program (PNR) and is defined in the 2014-2020 Partnership Agreement. It's an important example of a policy aimed at improving the living conditions of the population residing in areas at risk of marginalization. It was established with the declared intention of identifying homogeneous clusters of municipalities on the basis of accessibility to the main basic services. The 2021-2027 joint programming cycle reaffirmed the role of SNAI in supporting marginalized areas also in the proposal for a new Partnership Agreement. In 2022, a new mapping of the municipalities was created by the Department for Cohesion Policies and Istat, reshaping the definition of essential services to identify the intermunicipal poles and poles, and reviewing the minimum distance from the poles, functional to the classification of common in other clusters (Figure 1). Then, the municipalities Polo and Inter-municipal Polo have been identified according to the presence of three main services: transport, health facilities and schools.

Fig. 1 Spatial distribution municipalities - SNAI Classification 2021



Depending on the distance from these municipalities, the Belt, Intermediate, Peripheral and Ultra-peripheral municipalities have been identified. In particular, at national level according to SNAI 2021, there are 241 municipalities and inter-municipal centers, in which over 22 million inhabitants live and occupies a territorial area of 38,000 sq km. The municipalities Pole and intermunicipal Pole, together with the Belt municipalities, represent the macro-class of municipalities defined as Centers. The three remaining classes, on the other hand, identify the Inland Areas. The latter, and in particular the peripheral and outermost municipalities, constitute for the territory those areas most at risk of marginalization, social and economic. Surely, living in an area far from services leads to social and economic marginality that well explains the demographic decline of these areas over the last 70 years. Between 1951 and 2020, the population of the Centers grew on an annual average by 4.9‰ in Italy; the Inland Areas have lost an annual average 1.3‰ and one municipality out of three has systematically lost population since 1951.

2 Economic Indicators of Inland Areas

Territories, with different socio-demographic characteristics, have a non-homogeneous production system, more vulnerable in areas with higher marginalization and more efficient in the Centers. In general, over 80% of the local units are located in the Poles, Inter-municipal Poles and Belts, just over 1% of enterprises are located in the Ultraperipheral municipalities. The local units of the centers appear to be more structured than those of the internal areas, they have almost one more employee and significantly better performance indicators (Table 1).

Table 1 Average of employees ad share of local unitis

SNAI Classification	*Average number of employees per local unit*	*Share of local units in the area*
A – Pole	3,6	41,6
B - Intermunicipal Pole	3,5	2,6
C – Belt	3,7	36,0
D – Intermediate	3,2	11,8
E – Peripheral	2,9	6,9
F – Ultraperipheral	2,6	1,2

Source: our elaboration on Istat and SNAI data

Some economic indicators are provided to describe inland areas. The first one is labour productivity indicator (LPI), calculated by dividing a measure of output[1] by a single measure of input[2]. The output aggregate chosen in the numerator of the formula is the value added[3] (VA) by inland areas and NUTS1 (territorial divisions).

This ratio, expressed in euro per employees, indicates to some extent how much economic production activity over a given period can be attributed to each employed person, and also how it changes. Other economic indicators to describe inland areas are: a) Purchases of goods and services on turnover (PTI): ability of the company to

[1] Usually output is chain linked valued added at basic prices and labour input is measured as total hours worked by all persons engaged in production (both employees and self-employed). The data available at the level of inland areas does not allow transformation to chain linked, and hours worked are also not available at the level of inland areas. For this reason, value added in current prices and the number of employees were used..

[2] Usually labour input is measured as total hours worked by all persons engaged in production in the domestic concept, i.e. hours worked by employees and the self-employed, in either their primary or secondary activity, engaged in a productive activity for a resident unit and receiving remuneration regardless of their place of residence.

[3] With reference to the value added estimation of market producers at the territorial level, the statistical register on financial statements of enterprises (Frame-SBS) is the main source. This database includes information on economic performance for the population of active market enterprises at micro-level (excluding agriculture, household services and financial intermediation). It is built through a complex procedure of integrating data from administrative archives, subjected to a process of harmonisation and combined with data from the Small and Medium Enterprises (SME) and the Balance Sheets of Large Enterprises (SCI) surveys. This information system has to be integrated with the information included in the Local Unit ASIA Register (LU ASIA) and in the annual Register on employment, wages, hours worked and labour costs in Local Units (RACLI).

cover the costs for the purchase of goods and services with sales; b) Wages on value added (WVA): ability to distribute the income produced, representing how much of the value is attributed to the labour factor; c) Value added on turnover (VAT): the increase in value that business activity brings to the transformation processes of goods and services on the value of sales. It also indicates how large the margin is to remunerate internal production factors; d) Wages on employees (WAE): it represents the wage per employee.

3 Results: Inland Areas Economy Profile

In Figure 2 there is a comparison[1] was made between these indicators by two classifications (6-categories of SNAI classification in the left and 2-categories in the right Centre – A, B, C and Inland Areas – D, E, F). Centers are characterized by enterprises with higher productivity and profitability (Table 2 and Figure 2). Local units of enterprises in centers have, on average, a productivity of more than 48,000 euros (+15,000 euros respect to local units in inland municipalities). The distance between average wages between centers and inland areas is less significant.

Table 2 Average of economic indicators by inland areas

SNAI Classification	***LPI (euros)***	***WAE (euros)***	***VAT (%)***	***PTI (%)***	***WVA (%)***
A – Pole	48.278	25.505	25,5	72,9	39,2
B - Intermunicipal Pole	39.295	23.308	26,7	71,3	42,3
C – Belt	44.279	24.999	25,7	72,4	41,1
D – Intermediate	38.197	22.382	26,6	72,2	39,8
E – Peripheral	34.169	20.769	27,0	72,1	39,1
F – Ultraperipheral	33.100	20.712	34,5	63,8	35,6

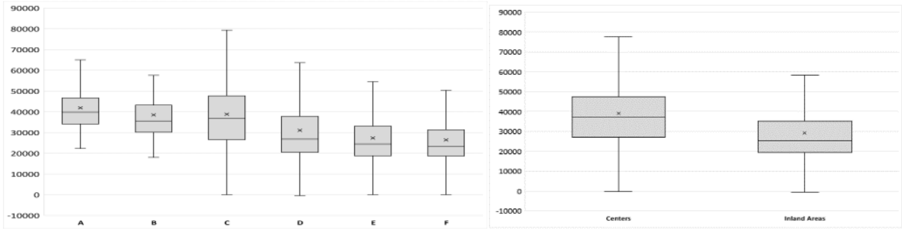
Source: our elaboration on Istat and SNAI data

[1] Box plots are a useful way to compare two or more sets of data visually. It displays:

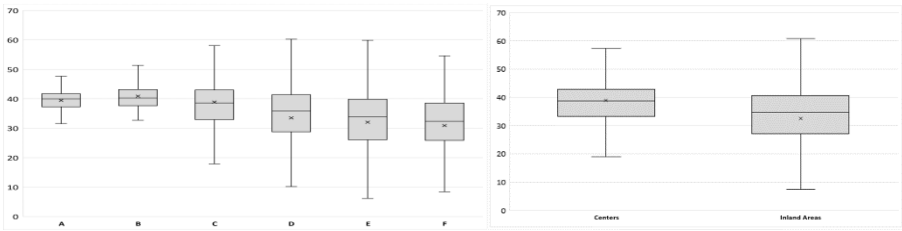
- the Median is represented by the line in the box.. Half the observations are less than or equal to it, and half are greater than or equal to it.
- the Interquartile range box represents the middle 50% of the data. It shows the distance between the first and third quartiles (Q3-Q1).
- the Whiskers extend from either side of the box. The whiskers represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers.
- The mean is represented by the “X” in the box.

Fig. 2 LPI, WVA, VAT, PTI, WAE by inland areas

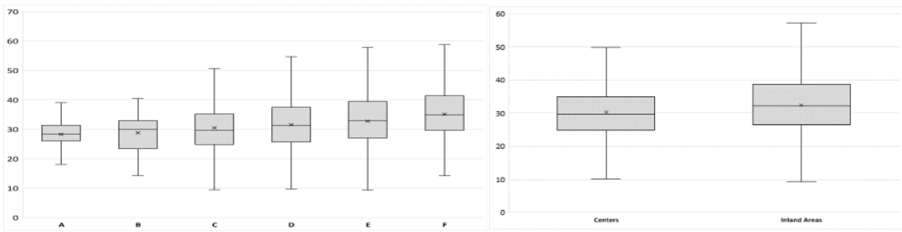
LPI - Labour productivity, Value Added on employees (euros)



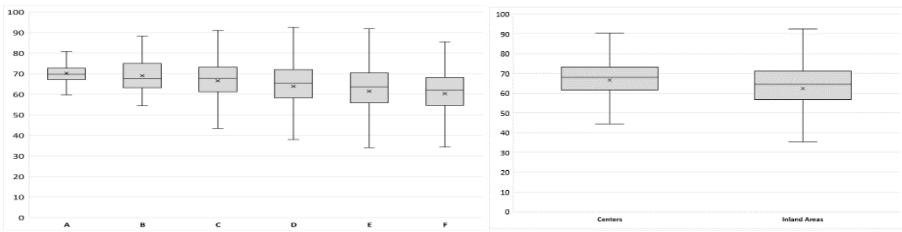
WVA - Labour cost on value added (%)



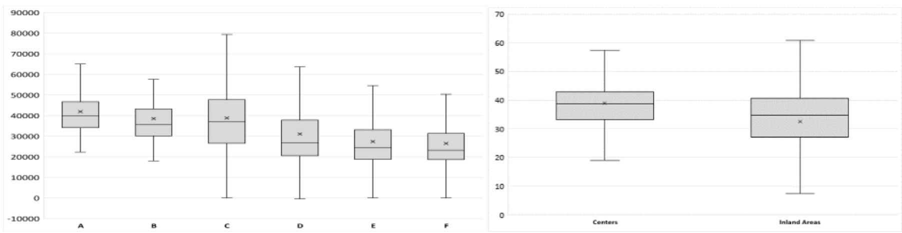
VAT - Value added on turnover (%)



PTI - Purchases of goods and services on turnover (%)



WVE - Labour cost on employees (euros)



Source: our elaboration on Istat and SNAI data

The PTI indicator, which represents the local unit's ability to cover the cost of goods and services through sales, appears to be homogeneous among areas, except for the inland municipalities, where it takes on a significantly lower value. Value added on sales grows from the center to the periphery and it is 10 percentage points higher in the inland municipalities than in the centers. The WVA indicator appears to be less discriminating, and it is also confirmed in the logistic model tested.

4 Results: Logistic Regression

In the final analysis, a logistic model (Table 3) was constructed to test the selected indicators in order to describe the differences between local units in the centers and inland areas. The logistic model assumes as the dependent variable the presence of the local unit in an inland area (y=1 if unit is in inland area) and the independent variables the average number of employees per local unit and the indicators previously analysed. All indicators are found to be significant with the exception of WVA. The logistics model also shows that local units in inland areas are smaller, less structured, with lower productivity and lower wages per employee respect to centers.

Table 3 Logistic regression (y=1 inland areas, y=0 centres)

Indicators	Estimation	p-value
LPI - Labour productivity, Gross Value Added on employees	-0.00001	<.0001
WAE - Labour cost on employees	-0.00005	<.0001
VAT - Value added on turnover	0.0352	<.0001
PTI - Purchases of goods and services on turnover	0.0110	<.0001
WVA - Labour cost on value added	-0.00126	0.2631
Average number of employees per local unit	-0.1248	<.0001

Source: our elaboration on Istat and SNAI data

References

1. OECD, Manual on "Measuring productivity" (2001) https://www.oecd.org/sdd/productivity-stats/2352458.pdf
2. Dipartimento per le politiche di coesione. Aggiornamento 2020 della mappa delle aree interne. NUVAP Technical Note (2022)
3. Dipartimento per le politiche di Sviluppo (2013). Strategia nazionale per le Aree interne: definizione, obiettivi, strumenti e governance – Technical document attached to the draft Partnership Agreement submitted to the EU

Session of free contributes SFC3 - *Social issues*

1. *Modelling the gender gap in youth mortality with an Age-Period-Cohort analysis* (Lanfiuti Baldi G. and Nigri A.)
2. *Random forest for classifying odor emission sources* (Distefano V., Palma M., De Iaco S. and Mazuruse G.)
3. *An Experimental Annotation Task Investigating Annotator Agreement Within a Misogynistic Dictionary and Corpus* (Tontodimamma A., Ignazzi E., Anzani S., Fontanella L. and Di Zio S.)
4. *Statistical analysis of COVID19 impact on Italian mortality* (Franchetti G. and Politano M.)
5. *Measuring multidimensional deprivation using objective and subjective data: an application of the Voronoi ranking method* (Ciommi M., Mariani F., Polinesi G. and Recchioni M.C.)

Modelling the gender gap in youth mortality with an Age-Period-Cohort analysis

Modellazione delle differenze di genere nella mortalità giovanile con un'analisi Età-Periodo-Coorte

Giacomo Lanfiuti Baldi and Andrea Nigri

Abstract In this paper, we propose an Age-Period-Cohort (APC) model leveraging Skew-Normal distribution, aiming at modelling the gender gap in youth mortality.
We noticed that gender differences in youth mortality are largest around the age of 20, but these differences are not symmetrical with respect to the peak. Following this evidence, we perform the APC analysis in which the response variable follows the Skew-Normal distribution. We adopt a sex ratio approach, using the ratio of age-specific mortality rates of men and women as the response variable.
Our research focuses on the population under age 45 in the United States between 1960 and 2020.
Abstract *In questo lavoro, proponiamo un modello Età-Periodo-Coorte (APC) che sfrutta la distribuzione Normale asimmetrica, con l'obiettivo di modellare il divario di genere nella mortalità giovanile.*
Abbiamo notato che le differenze di genere nella mortalità giovanile sono maggiori intorno ai 20 anni d'età, ma queste differenze non sono simmetriche rispetto al picco. In base a questa evidenza, eseguiamo l'analisi APC in cui la variabile di risposta segue la distribuzione Normale asimmetrica. Usiamo come variabile di risposta il rapporto tra i tassi di mortalità specifici per età di uomini e donne.
La nostra ricerca si concentra sulla popolazione di età inferiore ai 45 anni negli Stati Uniti tra il 1960 e il 2020.

Key words: Mortality modelling, Skew-Normal, Age-Period-Cohort Model

---

Giacomo Lanfiuti Baldi
Department of Statistics, Sapienza University of Rome, Rome, Italy e-mail: giacomo.lanfiutibaldi@uniroma1.it

Andrea Nigri
Department of Economics, Management and Territory, University of Foggia, Foggia, Italy e-mail: andrea.nigri@unifg.it

1 Introduction

Gender differences in mortality are increasingly discussed and studied in the social and demographic fields. Modelling and analysing gender differences in mortality can tell us a lot about the social context of a population or a country [8]. Thus, a better understanding of sex differences in mortality can guide public health efforts to reduce overall mortality rates and promote greater equity in health outcomes.

Gender differences are not constant at all ages and are driven by different causes of death. Many studies focus on gender differences in mortality at adult ages, but there is less literature on differences at younger ages. At the same time, gender differences in mortality are greater at younger ages [6] and this is mainly due to the different behaviour of the two groups. Men in particular run a much higher risk of accidental death around the age of 20 [7]. In addition to age, mortality differences change over time because the behaviour that generates them changes over the years [2].

We aim to study how the gender gap in mortality at younger ages (under 45) varies concerning individual age groups and over time. We want to define and interpret the effects of different ages and social, cultural and behavioural changes in society [11]. To do this, we work in an Age-Period-Cohort (APC) framework leveraging a model based on the Skew-Normal distribution. The Skew-Normal distribution is not widely used in the APC framework.

In particular, we are interested in studying in the United States (US), where more than the 30% of deaths at young ages are due to external causes (unintentional injuries) [3] and the issue of road accidents, and violent and risky behaviours among young people is often at the centre of public debate.

2 Data and Measure

We use a sex-ratio approach to study gender differences: we analyse the ratio of age-specific mortality rates between males ($m^M_{x,t}$) and females ($m^F_{x,t}$) over time:

$$SR_{x,t} = \frac{m^M_{x,t}}{m^F_{x,t}}. \tag{1}$$

This measure is useful for several reasons: it allows us to use a single variable to study the two sexes, it is less sensitive to the general level of mortality than the absolute difference in deaths [2], and finally, it has a well-defined and known shape [7]. We use age-specific mortality rates per sex and single year of age from the Human Mortality Database [10] of the United States in the study period.

Generally, this sex-ratio over the ages is characterized by a *peak* and a *hump*. The peak, which is the highest and most concentrated, coincides with youthful ages and is generally attributed to the highest male mortality due to riskier behaviours [7]. The hump corresponds to the adult ages and it was primarily caused by excess male mortality from cancer [2]. According to [7] we set the threshold age (between the peak and the hump) at 45 ages and we will focus only on the peak (Fig.1), in order to study the gender gap in mortality at young ages.

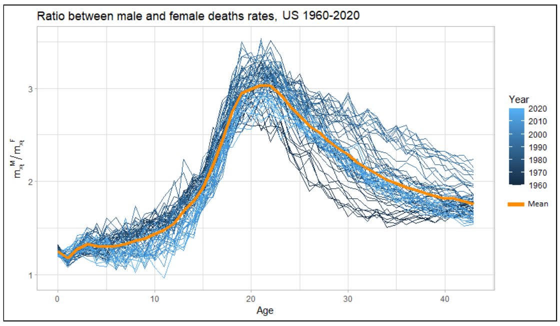


Fig. 1 Sex-Ratio of the Age-Specific Mortality Rates in US between 1960 and 2020 for young (under 45) population. Data source: HMD

Gender differences in infant mortality are very low in all periods considered, the differences start to increase in the years of adolescence. Male mortality at the peak comes to more than 3 times that of females in most years and there is no trend in the shift of the peak age over the years.
We have noticed that the sex-ratio at younger ages has a Gaussian shape, but in most of the years, it is not symmetrical around the peak.

3 Model specification

In order to study the age, period and cohort effects we suggest leveraging a model based on a *Skew-Normal* distribution. This distribution is useful for modelling both symmetric and skew data sets.

As defined by Azzalini in 1985 [1], Z has a skew-normal distribution with parameter $\lambda \in \mathbb{R}$ if:

$$f(z|\lambda) = 2\phi(z)\Phi(\lambda z), \quad z \in \mathbb{R} \tag{2}$$

Where $\phi(\cdot)$ and $\Phi(\cdot)$ are the $N(0,1)$ probability density function and cumulative distribution function, respectively.

If $Z \sim SN(\lambda)$, then the random variable $Y = \mu + \sigma^2 Z$ still has a skew-normal distribution with *location* parameter $\mu \in \mathbb{R}$, *scale* parameter $\sigma^2 \in \mathbb{R}^+$ and *skewness* parameter λ. The probability density function of $Y \sim SN(\mu, \sigma^2, \lambda)$ is given by:

$$f(y;\theta) = f(y;\mu,\sigma^2,\lambda) = \frac{2}{\sigma^2}\phi\left(\frac{y-\mu}{\sigma^2}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma^2}\right). \tag{3}$$

Following the framework proposed by Klein [5], we can generally set up the relationship between distribution parameters and the elements of the linear predictor as:

$$g(\varphi) = \sum_{j=1}^{J} f_j(\nu),$$

where f may comprise various forms, defined on basis of the covariate structure, in our case we consider a linear function $f_j(\nu) = \mathbf{X}\beta_j$, that represents the fixed effects.

Specifically, let's consider $y^{\mathrm{T}} = (y_1, y_2, \ldots, y_n)$ as the vector of the response variable and $f(y; \boldsymbol{\varphi})$, a density function with k parameters $\boldsymbol{\varphi}^{\mathrm{T}} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ modelled by linear additive models. We assume that observations y_i are independent conditional on $\boldsymbol{\varphi}$, with density function $f(y_i; \varphi_i)$, where $\varphi_i{}^{\mathrm{T}}$ is a vector of k parameters related to explanatory variables and random effects. Let $g(\cdot)$ be a known monotonic link function relating φ_k to explanatory variables through an additive model given by:

$$g(\varphi) = \eta = \mathbf{X}\beta$$

where: $\beta^{\mathrm{T}} = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'k})$ is a parameter vector of length J', $\mathbf{X}$ is the design matrix of order $n \times J'$. In our study, for the Skew-normal family distribution: $\varphi = \mu$ and $g(.)$ is the identity function. So, we have the following model:

$$\mu = \mathbf{X}\beta.$$

Thereby, the components of the model are: $\mathbf{y}$, is the response vector of length n; $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ are the design matrices of the effects for the three dimensions in which we are interested (Age-Period-Cohort) and $\boldsymbol{\beta}^{\mathrm{T}} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_p^{\mathrm{T}}\right)$ are the linear parameters.

Here, we provide the structure of the specific APC Skew-Normal model to be applied to the study of the gender gap while including the Age-Period-Cohort (A-P-C) structure as predictor.

Let $\mathscr{A} = \{a_0, a_1, \ldots, a_\omega\}$, $\mathscr{P} = \{p_0, p_1, \ldots, p_n\}$ and $\mathscr{C} = \{c_0, c_1, \ldots, c_m\}$ be the set of age, year and cohort categories, respectively. The APC model describes the gender ration of death rates at age $a \in \mathscr{A}$, time $p \in \mathscr{P}$, and cohort $c \in \mathscr{C}$.
Using categorical coding for Age, Period and Cohort respectively we introduce the model:

$$\mu = \beta_{(a)} + \beta_{(p)} + \beta_{(c)}. \tag{4}$$

As it is well known in the literature [9], since the cohort is obtained from a linear relationship of the other two variables (Age and Period), the model suffers from a lack of identifiability. To solve this it is necessary to impose constraints, which in this case are:

$$\sum_{\omega=1}^{\Omega} \beta_{(a_\omega)} = \sum_{n=1}^{N} \beta_{(p_n)} = \sum_{m=1}^{M} \beta_{(c_m)} = 0. \tag{5}$$

4 Results

In our analysis, we estimate the models described in the previous section (3) with a Bayesian approach. Samples from the posterior distributions of the parameters and effects were drawn by using Hamiltonian Monte Carlo sampling and specifically using the `stan` software package [4].

We specify a $\sigma^2 \sim \Gamma(0.01, 0.01)$ and a $\lambda \sim \Gamma(0.01, 0.01)$ as prior distribution.

Here we report the results (Fig.2) of the model with the constraints in which Age, Period and Cohort are treated as categorical variables, with the first category being the reference one.

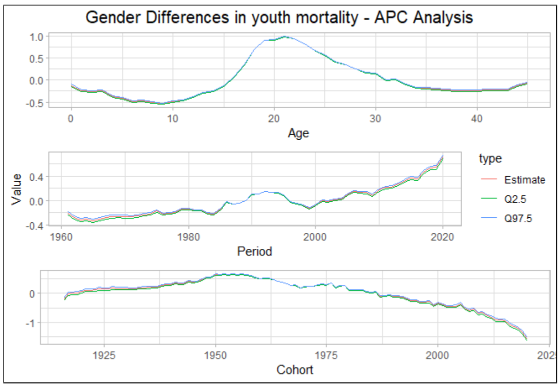


Fig. 2 APC model on the gender gap in youth mortality, US 1960-2021. Data source: HMD

The age parameters provide the structure of the gender gap in youth mortality, which can be found on average in all years of the period and for all cohorts. The value of the estimated parameters for age increases with adolescence. Researchers usually do not look at biological factors to explain the excess mortality among men at these ages, but rather at individual and social reasons. These reasons become even bigger and more important in the peak years of gender differences (21-22 years): car accidents, suicides and violence are by far the most important causes.

From the two graphs of period and cohort parameters, we can observe two different trajectories. Cohorts born after the mid-1950s gradually experienced smaller differences in mortality between the two sexes. In contrast, the period effect indicates an increase in the gender gap on average in the last decades.

5 Discussion

We observed the gender gap in youth mortality in the US between 1960 and 2020. For this aim, we used a sex-ratio approach in the Age-Period-Cohort framework, leveraging a model based on a Skew-Normal distribution.

The parameter estimates were performed adopting a Bayesian approach and using `stan` software.

The innovation of this work is to implement an Age-Period-Cohort analysis, assuming that the target variable has an asymmetric distribution: the Skew-Normal distribution is not widely used in the APC framework, which usually is based on the normal distribution in the demographic field.

Observing sex differences in mortality and how these vary across ages and over time is useful for understanding society and the behaviours that determine them. Moreover, the knowledge of the mortality dynamics and of the differences between the sexes can be an excellent tool in the hands of policymakers.

Preliminary results show that over the past 25 years in the US, we have observed that younger cohorts are benefiting from societal changes and the attention that the topic of youth mortality is receiving in the public debate.

References

1. Azzalini A.: A class of distributions which include the normal ones. Scandinavian Journal of Statistics. 171-178 (1985)
2. Bergeron-Boucher, M.-P., Canudas-Romo, V., Pascariu, M. and Lindahl-Jacobsen, R.: Modeling and forecasting sex differences in mortality: a sex-ratio approach. Genus, 74, 1-28 (2018) https://doi.org/10.1186/s41118-018-0044-8
3. Heron, M.: Deaths: Leading causes for 2019. National Vital Statistics Reports, 70(9). National Center for Health Statistics, Hyattsville, MD (2021) https://doi.org/10.15620/cdc:107021
4. Hilton, J., Dodd, E., Forster, J. J., and Smith, P. W. F.: Projecting UK mortality using Bayesian generalised additive models (2018) arXiv preprint arXiv:1802.03242. Retrieved from https://arxiv.org/abs/1802.03242
5. Klein, N., Kneib, T., Klasen, S. and Lang, S.: Bayesian structured additive distributional regression for multivariate responses. Journal of the Royal Statistical Society: Series C (Applied Statistics), 64(4), 569–591 (2015) https://doi.org/10.1111/rssc.12090
6. Kung, H. C., Hoyert, D. L., Xu, J. and Murphy, S. L.: Deaths: final data for 2005. National Vital Statistics Reports, 56(10), 1-120 (2008)
7. Meslé, F.: Life expectancy: a female advantage under threat. Population and Societies, 402(4), 1-4 (2004)
8. Nathanson, C. A.: Sex differences in mortality. Annual Review of Sociology, 10(1), 191-213 (1984) https://doi.org/10.1146/annurev.so.10.080184.001203
9. O'Brien, R. M.: Mixed models, linear dependency, and identification in age-period-cohort models. Statistics in Medicine, 36(16), 2590-2600 (2017) https://doi.org/10.1002/sim.7305
10. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Human Mortality Database (2021) Retrieved from http://www.mortality.org
11. Yang, Y., Schulhofer-Wohl, S., Fu, W. J. and Land, K. C.: The intrinsic estimator for age-period-cohort analysis: what it is and how to use it. American Journal of Sociology, 113(6), 1697-1736 (2008) https://doi.org/10.1086/587154

Random forest for classifying odor emission sources*

Random forest per la classificazione delle sorgenti emissive odorigene

Veronica Distefano, Monica Palma, Sandra De Iaco and Gideon Mazuruse

Abstract The odor emissions generated by treatment plants imply complex environmental and economic issues. The modern instrumental odor monitoring systems based on array of several sensors, continuously record the gaseous compounds, but they are characterized by poor selectivity thus compromise the possibility to discriminate and identify the emission sources. In this paper, the ability of odor sensors to distinguish the treatment plant sections generating the gaseous compounds is evaluated by a machine learning classification approach, the Random Forest. The goodness of this method is highlighted through apt performance measures and also with respect to the classical multiple discriminant analysis.

Abstract *Le emissioni odorigene generate dagli impianti di depurazione determinano questioni ambientali ed economiche complesse. I moderni sistemi strumentali di monitoraggio degli odori basati su array di più sensori, registrano in continuo i composti gassosi, ma sono caratterizzati da scarsa selettività e pertanto possono compromettere la possibilità di discriminare ed idientificare la sorgente emissiva. In questo lavoro, la capacità dei sensori odorigeni di distinguere le sezioni*

Veronica Distefano
DES-Sect. of Mathematics and Statistics, University of Salento (Italy), e-mail: veronica.distefano@unisalento.it

Monica Palma
DES-Sect. of Mathematics and Statistics, University of Salento (Italy), e-mail: monica.palma@unisalento.it

Sandra De Iaco
DES-Sect. of Mathematics and Statistics, University of Salento (Italy); National Biodiversity Future Center, 90133 Palermo, Italy, e-mail: sandra.deiaco@unisalento.it

Gideon Mazuruse
Marondera University of Agricultural Sciences and Technology (Zimbabwe), e-mail: gmazuruse@gmail.com

* The paper is one of the results from the research project "Modelli intelligenti di correlazione per la valutazione dell'inquinamento olfattivo"(Project Code f3f76727; scientific responsible: Prof Monica Palma) approved by the Regional Program "RIPARTI"of the Apulian Region.

dell'impianto di trattamento che generano i composti dell'odore è stata valutata mediante un approccio machine learning di classificazione multipla quale, il Random Forest. L'adeguatezza di tale metodo di analisi è evidenziata attraverso specifiche misure di performance ed anche rispetto ai risultati ottenuti con la classica Analisi Discriminante Multipla.

Key words: treatment plant sections, electronic nose, machine learning approach

1 Introduction

Odor emissions from industrial and environmental protection plants are often the cause of olfactory nuisance capable of generating annoyance in citizens residing in their neighborhood [9]. The deriving annoyance is often continuous and, therefore, this phenomenon can interfere with the state of human well-being, generating complaints and triggering conflicts that can also have repercussions on the economic, commercial, and tourism activities [7]. The most common sources of odor emissions are the treatment plants which generate a huge amount of substances that are harmful to the environment, characterized by various physical as well as chemical properties; some of these substances are gaseous compounds featured by an unpleasant odor. In this context, a careful management of the plant odor emissions is crucial and can avoid adverse effects on the environment and human health. The measurement of an odor concentration can be obtained in three different ways, i.e. by analytical determinations based on the mass spectrometry, by the olfactory perception from a group of panelists and by the electronic nose (IOMS, Instrumental Odor Monitoring System), based on the interaction between special sensors and volatile molecules. The application of electronic nose has been increasing in many research areas of industrial system, such as pharmaceutical firms, food industries, agriculture, biotechnology, as well as for the monitoring process of the treatment plants. It is based on a set of chemical sensors, each of them with its own specificity, and an appropriate pattern recognition system capable of recognizing simple or complex odors [5]. In the case of a treatment plant, several odor sensors are installed along the various sections of the plant, and therefore they represent the key components for monitoring and controlling odors emissions, continuously. In the recent years, several studies and applications have been developed to analyze data from IOMS. In [3] artificial neural network and decision trees have been used to predict the odor properties of post-fermentation sludge from two biological-mechanical treatment plants located in the north of Poland; in [4] the learning machine method known as Random Forest (RF) has been applied to recognize odor classes nearby a wastewater treatment plant. The results obtained from the analysis have established to what extent the unpleasant odors perceived by the citizens came from the monitored plant. In the present paper, the odor measurements from the sensors installed on some treatment plants, have been analyzed to assess the ability of the sensors to distinguish the sections of the treatment plant which have produced the specific gaseous compound. As known, the sensors used in the electronic noses are sensitive to a wide range of chemical compounds and therefore are characterized by poor selectivity. The sensor matrix is generally made up of 6-10 sensors which are devoted to give, to each mixture under examination, a set of responses that constitutes the so called "finger print", a kind of "olfactory pattern", of the odor source. In order to evaluate to what extent the sensors are able to distinguish the emission sources (the sections of the treatment plant), the machine learning RF classification has been adopted. The goodness of this method has been highlighted through apt performance measures and also with respect to the classical multiple discriminant analysis (MDA).

2 RF: a brief review

Firstly introduced by Breiman [2] in 1984, RF classification method has been applied in different research fields, as supervised learning approach used for several types of classification and regression tasks. The RF is an ensemble of tree-structured classifiers and consists of building a set of decision trees. For creating a RF, a specific number of trees are generated through bagging or bootstrap sampling. Thus, RF selects n sample sets $[T_1, T_2, \ldots, T_n]$ from the original training set T using bagging sampling and then builds n decision tree models $[K_{T_1}, K_{T_2}, \ldots, K_{T_n}]$ for each of the n sample sets. Each generated tree in RF is tested with the test set X, which is not used in training to obtain n classification results $[C_{1(X)}, C_{2(X)}, \ldots, C_{n(X)}]$. For the outcome, RF uses the n classification results from each tree and provides the decision according to the principle called majority rule. The choice to apply this approach has been justified on the one hand by the possibility to identify non-linear patterns of the data and on the other hand by the consideration that RF does not need variables scaling. Moreover, by increasing the number of trees, the classification accuracy of the model improves and the over-fitting reduces [1].

3 Data set and results

The analyzed data have been collected by a company specialized in technical and scientific environmental assistance and consultancy services for private and public enterprises. The data set consists of 285 measurements from 10 sensors of a IOMS, which record different types of gaseous compounds, namely W1C-Aromatic, W5S-Broad range (broad range sensitivity, react on nitrogen oxides), W3C-Aromatic (Ammonia), W6S-Hydrogen, W5C-Arom-aliph (alkanes, aromatic compounds, less polar compounds), W1S-Broad methane, W1W-Sulphur organic (terpenes and sulfur organic compounds), W2S-Broad-alcohol (alcohols, partially aromatic compounds), W2W-Sulph-clor (sulfur organic compounds) and W3S-Methane-aliph (sensitive to high concentration of methane). The sensors' measurements are expressed in mA (milliamperes). An additional categorical variable, called "treatment plant section" that defines the section of the plant at which the sensor data have been recorded, has been considered in the following analysis. More specifically, the "treatment plant section" variable regards nine sections of the plant (Fig. 1), that are: grit removal (1), primary sedimentation (2), biological oxidation (3), secondary sedimentation (4), denitrification (5), equalization (6), stabilization of sludge (7), sludge storage (8), sludge thickening (9).

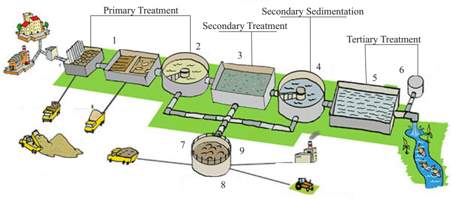


Fig. 1 scheme of a typical treatment plant, with its 9 sections and 4 macro-sections

As already specified, the available data set has been used in order to evaluate to what extent the sensor's measurements allow the separation of the sections of the treatment plant which have produced the specific odor emission. Tab. 1 shows the descriptive statistics of the data recorded by the 10 sensors and classified on the basis of the "treatment plant section" variable. By analyzing the descriptive statistics, it has been underlined that the sensors W1C, W3C and W5C have registered, on average, the lowest values, while the sensor W1W and W2W have registered the highest values. Moreover, the observed data sensors do not exhibit high inter-variability of the gaseous compound among the different sections. This behaviour of chemical gas sensors could be attributed to two main aspects, i.e. the cross-sensitivity and low selectivity, namely the sensors are effected to a mixture of gases with similar chemical properties and also are affected by similar chemical characteristics [8].

On the basis of these considerations, a second categorical variable ("treatment plant macro-section") has been defined by grouping the previous 9 sections of the treatment plant into 4 macro-sections, which are referred to primary treatment (I macro-section), secondary treatment (II macro-section), secondary sedimentation (III macro-section) and tertiary treatment (IV macro-section), as shown in Fig. 1 The sensors average values have been compared with respect to both 9 sections and 4 macro-sections of the plants by using the non parametric Kruskal-Wallis test [6], under the null hypothesis of absence of difference among the sensors. The results shown in Tab. 1 have highlighted that, at significance level of 5%, the sensors values are different on average, in both the proposed classifications. The ability of the sensors to distinguish the emission source of the treatment plant has been evaluated with respect to both 9 sections and 4 macro-sections of the plant. To this aim, the supervised machine learning RF has been used, where the target variable is, alternatively, the categorical variable related to the plant sections or that one related to the plant macro-sections. Hence, the data set has been partitioned into 70% training data set and 30% testing data set. Then, the results of classification experiments have been visualized in a matrix, called confusion matrix. The diagonal elements represent the number of observations correctly classified, while the off-diagonal elements represent mis-classification (Tab. 2). Moreover, three performance mea-

Table 1 Mean values and standard deviations (in brackets) for the distributions of the sensor data, classified by 9 plant sections and 4 plant macro-sections

	Plant sections										Plant macro-sections				
Sensors	1	2	3	4	5	6	7	8	9	Test	I	II	III	IV	Test
W1C	0.7	0.6	0.8	0.7	0.7	0.6	0.7	0.7	0.7	22.9*	0.7	0.7	0.7	0.7	8.4*
	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.2)	(0.1)		(0.2)	(0.1)	(0.1)	(0.2)	
W5S	2.9	5.9	5.27	2.9	3.1	7.8	9.7	5.9	8.2	31.0*	4.9	7.7	2.9	0.7	13.9*
	(0.7)	(9.2)	(7)	(0.7)	(1.0)	(15.7)	(25.4)	(5.9)	(9.4)		(7.6)	(13.9)	(0.7)	(12.8)	
W3C	0.8	0.6	0.8	0.7	0.7	0.7	0.8	0.8	0.7	32.0*	0.7	0.8	0.7	0.7	0.7*
	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.2)		(0.2)	(0.1)	(0.7)	(0.2)	
W6S	1.6	1.8	1.5	1.6	1.7	1.7	1.3	2.0	1.5	60.2*	1.7	1.6	1.6	1.7	43.2*
	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.3)	(0.4)	(2.2)	(0.3)		(0.2)	(1.2)	(0.1)	(0.2)	
W5C	0.8	0.7	0.9	0.8	0.8	0.7	0.8	0.8	0.8	35.6*	0.7	0.8	0.8	0.7	27.9*
	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.2)		(0.2)	(0.1)	(0.1)	(0.2)	
W1S	6.4	9.7	4.3	5.9	6.1	9.2	6.6	7.7	7.2	32.2*	8.6	6.5	5.9	8.1	18.6*
	(3.8)	(6.4)	(1.6)	(3.3)	(2.2)	(7.6)	(7.1)	(7.6)	(3.8)		(5.8)	(5.8)	(3.3)	(6.4)	
W1W	15.6	45.2	12.5	10.8	13.5	37.5	18.1	16.5	35.0	36.0*	35.2	19.8	10.8	29.1	14.7*
	(18.3)	(56.6)	(11.1)	(10.7)	(10.9)	(55.0)	(23.2)	(18.7)	(55.7)		(49.2)	(30.9)	(10.7)	(46.3)	
W2S	7.5	12.4	5.0	9.4	7.8	11.5	7.0	8.35	9.31	31.4*	10.8	7.4	9.0	10.2	12.6*
	(7.6)	(9.4)	(2.4)	(7.6)	(6.5)	(9.3)	(7.7)	(7.9)	(7.4)		(9.1)	(6.9)	(7.6)	(8.5)	
W2W	11.5	39.1	7.1	8.9	9.6	30.3	13.2	10.9	28.9	36.6*	29.7	14.3	8.9	23.0	14.8*
	(14.6)	(55.1)	(8.4)	(8.2)	(8.1)	(49.4)	(20.4)	(13.4)	(55.8)		(47.3)	(29.4)	(8.2)	(41.2)	
W3S	5.5	8.5	3.9	3.9	5.5	5.6	5.0	5.1	5.9	28.8*	7.5	4.9	3.9	5.6	21.2*
	(3.2)	(5.8)	(1.1)	(1.3)	(2.1)	(3.4)	(2.7)	(3.5)	(4.9)		(5.3)	(3.3)	(1.3)	(3.0)	

$*$*p-value*$<$0.05

sures, such as accuracy, precision and recall indices, have been computed on the basis of the previous confusion matrices to assess and test classification efficiency. Then, RF has been trained with all 10 sensors and 10-fold cross-validation is used to calculate the average prediction accuracy. In Tab. 3 the results obtained by using the "treatment plant section" and "treatment plant macro-section" variables, are reported. In terms of all performance measures herein considered (accuracy, precision and recall) a small number of plant sections is able to better discriminate the odor emission sources. It is interesting to point out that the RF classification approach out-performs the classical approach of the MDA. Indeed, it is worth highlighting that the above mentioned performance indices get worse when the MDA is applied even by considering the "treatment plant macro-section" variable with a low number of classes (Tab. 3). The statistical analyses have been performed using the R-project free software environment for statistical computing and graphics.

4 Concluding Remarks

This study aimed at highlighting the ability of the machine learning RF classification method in order to evaluate to what extent the sensors of a treatment plant separate the sections of the plant, taking into account two different way of classifying the plant's sections. The goodness of the proposed method was highlighted in terms of accuracy, precision and recall measures. Finally, it was pointed that RF algorithm out-performs the classical MDA. Further developments will be to evaluate an improved RF classifier approach to separate better the odor emission sources by considering a combination of RF machine learning approach and a filter method for variables selection.

Table 2 Confusion matrices from RF with a) 9 plant sections, and b) 4 plant macro-sections

(a)

	Observed sections								
Predicted sections	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	1	0	0	0
2	0	11	2	3	2	5	0	0	1
3	1	0	1	1	1	1	1	0	1
4	2	0	2	1	0	0	0	0	0
5	1	1	0	0	1	1	0	0	1
6	2	1	0	1	1	6	1	4	0
7	0	1	0	0	2	0	4	2	1
8	0	0	1	0	0	1	1	3	1
9	0	1	2	0	1	0	0	0	1

(b)

	Observed sections			
Predicted sections	1	2	3	4
1	14	2	1	3
2	4	24	0	8
3	1	2	2	3
4	4	4	3	9

Table 3 Performance measures for the RF classification and for the MDA methods

Indexes	RF with 9 plant sections	RF with 4 plant macro sections	MDA with 9 plant sections	MDA with 4 plant macro sections
Accuracy	0.61	0.69	0.20	0.57
Precision	0.32	0.51	0.20	0.47
Recall	0.30	0.52	0.22	0.36

References

1. AL-Behadili, H.N.K.: Decision Tree for Multiclass Classification of Firewall Access. Int. J. Intell. Eng. Syst. (2021)
2. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A. Classification and Regression Trees. Boca Raton, FL: CRC press (1984)
3. Bylinski, H., Sobecki, A., Gebicki, J.: The use of artificial neural networks and decision trees to predict the degree of odor nuisance of post-digestion sludge in the sewage treatment plant process. Sustainability, 11, 4407 (2019)
4. Cangialosi, F., Bruno, E., De Santis, G.: Application of Machine Learning for Fenceline Monitoring of Odor Classes and Concentrations at a Wastewater Treatment Plant. Sensors, 21, 4716 (2021)
5. Gardner, J.W., Bartlett, P.N.: A brief history of electronic noses. Sensors and Actuators B: Chemical, 18, 210-211 (1994)
6. Kruskal, W. H., and W. A. Wallis.: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47: 583–621 (1952)
7. Oliva G., Zarra T., Massimo R., Senatore V., Buonerba A., Belgiorno V., Naddeo V.: Optimization of Classification Prediction Performances of an Instrumental Odour Monitoring System by Using Temperature Correction Approach. Chemosensors 9, 147 (2021)
8. Yaqoob U, Younis M.I.: Chemical Gas Sensors: Recent Developments, Challenges, and the Potential of Machine Learning—A Review. Sensors 21(8): 2877 (2021)
9. Zarra, T., Galang, G., Ballesteros, F.Jr., Belgiorno, V., Naddeo, V.: Environmental odour management by artificial neural network—A review. Environ. Int., 133, 105-189 (2019)

An Experimental Annotation Task Investigating Annotator Agreement Within a Misogynistic Dictionary and Corpus

Un task di annotazione sperimentale per investigare l'accordo tra gli annotatori in un corpus e dizionario misogino

Alice Tontodimamma, Elisa Ignazzi, Stefano Anzani, Lara Fontanella and Simone Di Zio

Abstract This work describes the development of two distinct experimental annotation tasks aimed at examining the agreement among annotators in two different contexts of misogyny: a misogynistic dictionary and a corpus consisting of misogynistic comments. In these two tasks, each annotator is required to categorize every term of a dictionary (or a comment) as offensive or misogynistic. Additionally, if a term/comment is deemed misogynistic, annotators should further specify its subcategory, including sexual objectification, dominance, body-shaming, derogatory language, intimidation, benevolent sexism, and neo-sexism. We compare different measures to assess the level of agreement among annotators.

Abstract *Questo lavoro descrive lo sviluppo di due tasks di annotazione sperimentali volti ad esaminare l'accordo tra annotatori in due diversi contesti di misoginia: un dizionario misogino e un corpus contenente commenti misogini. In questi, ogni*

Alice Tontodimamma

Department of Legal and Social Sciences, University "G. d'Annunzio", Chieti–Pescara, Italy; e-mail: alice.tontodimamma@unich.it

Elisa Ignazzi

Department of Neuroscience, Imaging and Clinical Sciences (DNISC), University "G. d'Annunzio", Chieti-Pescara, Italy; e-mail: elisa.ignazzi@unich.it

Stefano Anzani

Department of Neuroscience, Imaging and Clinical Sciences (DNISC), University "G. d'Annunzio", Chieti-Pescara, Italy; e-mail: s.anzani92@gmail.com

Lara Fontanella

Department of Legal and Social Sciences, University "G. d'Annunzio", Chieti–Pescara, Italy; e-mail: lara.fontanella@unich.it

Simone Di Zio

Department of Legal and Social Sciences, University "G. d'Annunzio", Chieti–Pescara, Italy; e-mail: s.dizio@unich.it

ICOMIC - Identifying and Counteracting Online Misogyny in Cyberspace - Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/202.

annotatore deve classificare ogni termine di un dizionario (o un commento) come offensivo o misogino. Inoltre, se un termine/commento è ritenuto misogino, gli annotatori devono specificare ulteriormente la sua sottocategoria, come l'oggettivazione sessuale, il dominio, il body shaming, il linguaggio dispregiativo, l'intimidazione, il sessismo benevolo e il neosessismo. Confronteremo diverse misure di agreement per valutare il livello di accordo tra gli annotatori.

Key words: misogynistic dictionary, misogynistic corpus, annotation task, subjective task, annotators' agreement

1 Introduction

In recent years, hatred directed against women has spread exponentially, especially in online social media, where the detachment resulting from being enabled to write without being obliged to reveal yourself directly allows people to feel greater freedom in the way they express themselves, and even to attack a chosen target with lower risk of being recognised or traced [6]. Although this alarming phenomenon has given rise to many studies from both computational linguistics and from machine learning point of view, less effort has been devoted to developing computational linguistic resources for the detection of misogyny. These resources have fundamental importance for training supervised machine learning models to be used to classify new textual instances shared on the Internet. One of the main problems in this field of research is that it is very complex to define what is online misogyny. In our work, we assume that online misogyny is content spread online that conveys hatred, aversion, and distrust towards women as women, and deep-seated prejudices against women. Besides, regarding the taxonomy we assume that if a term/comment is classified as misogynistic, it can belong to one or more the following categories: sexual objectification (terms/comments with which sexual images are evoked), dominance (terms/comments asserting the superiority of men over women to highlight gender inequality without using stereotyped portrayal, this is an implied form of abuse), body-shaming (terms/comments with which the physical aspect of a person is described and/or comparisons with narrow standards), derogatory (terms/comments that denigrate or demean women, considering them inferior), intimidating (intimidating terms/comments intended to express the intention or desire to inflict/cause harm on women or to express support, encouragement, promotion or instigation of such harm; this is an explicit form of abuse), benevolent sexism (terms/comment that express a kind of sexism that is often disguised as a compliment) and neosexism (terms/comment which deny the existence of discrimination against women and often show resentment of complaints about discrimination and "special favors" for women). Furthermore, a term could be neither offensive nor misogynistic but usable in misogynistic sentences (such as the term uterus). This taxonomy, from the perspective of automatic detection of online misogyny, can certainly help us understand the severity of a specific term/comment. In this study, we present two distinct experimental annotation tasks aimed at examining the agreement among annotators in two different contexts of misogyny: a misogynistic dictionary and a corpus consisting of misogynistic comments.

2 Related Work

As regards linguistic resources for the automatic detection of online misogyny, the literature shows that there are articles where lexicons are used: one of the most widely used lexicons is HurtLex, a lexicon built for different purposes. [2] propose specific misogyny lexicons for automatic misogyny identification in order to improve sentence embedding similarity and these Italian lexica and sources are available at their repository. [5] create a specific lexicon around misogynoir and use a lexicon-based approach to identify misogynoir. Furthermore, thanks to different evaluation campaigns, such as EVALITA (Evaluation of NLP and speech tools for Italian), IberLEF (Iberian Languages Evaluation Forum), SemEval (International Workshop on Semantic Evaluation), and FIRE (Forum for Information Retrieval Evaluation), there are now available datasets for automatic misogyny detection in multiple languages. Besides, an additional crucial aspect to consider when releasing annotated datasets is the need to clearly outline the guidelines used for categorising misogynous language. From literature, it emerges that three main works [1, 4, 8] in which taxonomy for annotating misogyny are released. In this scenario, our main goal is to build a lexicon to detect misogyny online starting from these available sources, adding a feature related to the misogyny category and releasing an annotated dataset with the same taxonomy. Given the complexity and subjectivity of the underlying phenomenon and the annotation task, it is crucial to assess the agreement among annotators.

3 Task 1: Annotating Misogynistic dictionary

In the first task, trainees annotated 1200 words, starting from the revised HurtLex dictionary [7] and integrating it with other available resources (see section 2). The terms are all in singular form. Each annotator was asked to annotate if a term was offensive or misogynistic. Moreover, if the term was misogynistic, they were told to choose a subcategory of misogyny (see section 1). The annotators worked independently and were provided with written guidelines for the task. The annotation process was carried by 6 trainees (3 males, 3 females, students on the Sociology degree course) engaged in an internship program focused on misogyny online in the Computational Social Research Lab[1].

[1] http://csrlab.unich.it/.

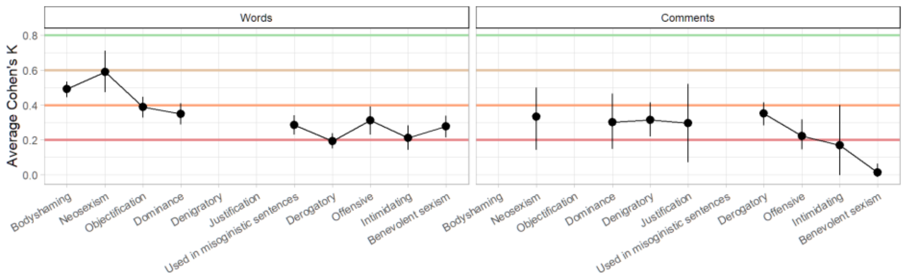
4 Task 2: Annotating Misogynistic corpus

For the second task, the trainees were divided into five groups of two. Each group/trainee annotated 300 comments, downloaded from Facebook, Instagram and Twitter using exportcomments[1] and socialgrabber[2]. These comments are related to female politicians, tv personalities, journalists and influencers. Every annotator was asked to annotate each comment as offensive or misogynistic. If the comment was misogynistic, they could also indicate a subcategory (in this case the taxonomy used is slightly different). Written guidelines were provided for the task. The annotation process was carried by 10 trainees (5 males, 5 females, students on the Sociology degree course) paid to perform this task.

5 Analysis of (dis)agreement

Since annotation tasks about misogynistic language are highly prone to subjectivity and significant disagreement [3], we computed agreement measures for each task and for each category assignment. For each pair of annotators, we computed Cohen's Kappa using the *psych* R package, which allows to compute the parameter, as well as 95% confidence intervals. In figure 1 we show the average value of Cohen's Kappa across all couples of raters for each category and for both tasks. Notably, some categories only apply to one of the tasks. The Cohen's Kappa values for task 1 are higher than task 2, but the agreement is generally low. Figure 2 and 3 show the values of Cohen's Kappa for each couple of raters in task 1 and 2 respectively. Subplots are ordered by decreasing average Cohen's Kappa levels, and the same is true for the categories on the horizontal axis: categories on the left have a higher average agreement across raters. As can be seen from the plot, even within couples that score highest on the agreement, there is high variability among categories. Some categories go as low as values compatible with random assignment (below the red line, Cohen's Kappa of less than 0.2).

Fig. 1 Average Cohen's K across classifications.



[1] https://exportcomments.com/
[2] https://www.socialgrabber.net/

Fig. 2 Cohen's K (and 95% CI) for each couple of raters across each classification in task 1

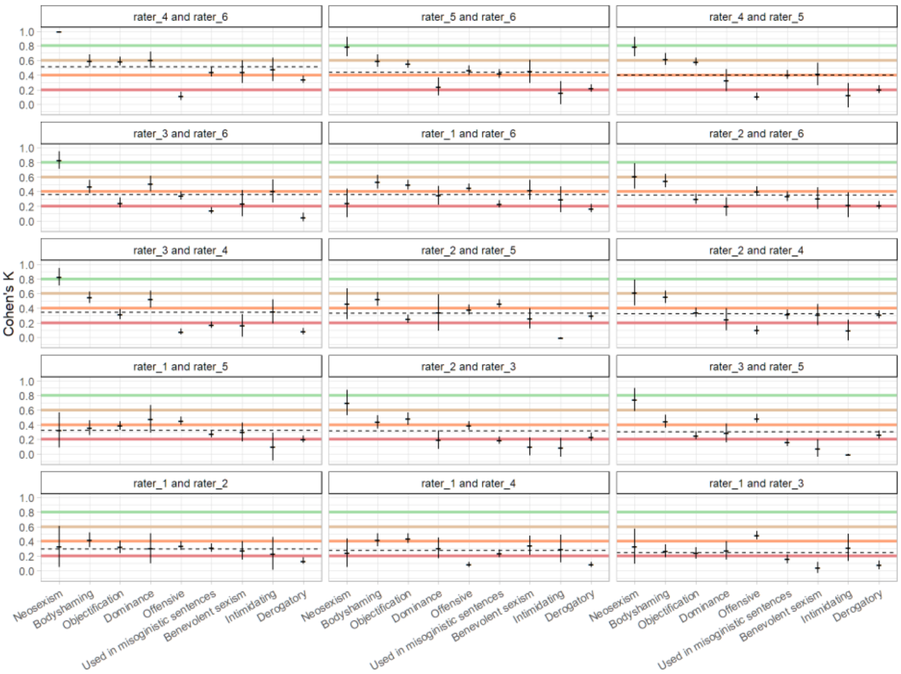
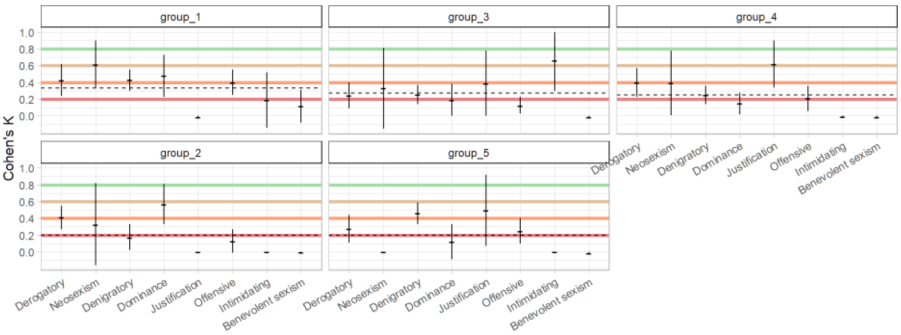


Fig. 3 Cohen's K (and 95% CI) for each couple of raters across each classification in task 2



6 Discussion and Conclusions

In this paper, we compared the level of agreement in two annotation tasks aimed at producing resources for the detection of misogyny in NLP. In both tasks the level of agreement appears to be quite low, and the additional context of the full sentence in the second task doesn't seem to help. This might be explained by the higher complexity that is involved in processing a full sentence, compared to judging a single word without a context. Different people might pay more attention to different parts of the sentence, this steering the classification towards a label or the other. Moreover, looking at the agreement within each couple of raters we can observe a high variability even in those couples that have a higher average score of agreement across categories. Given our results, we can say that classification tasks commonly used in NLP to produce lexical resources and corpora are highly complex and subjective. Indeed, even a pool of trained raters highly familiar with the subject matter performed the task with medium-low levels of agreement. This is far below what would generally be considered acceptable as a gold standard for automatic classification tasks. In order to obtain more robust and reliable labels for our computational resources, we suggest that the classification tasks should be carried out by more than a handful of raters. This would not only provide data more robust to errors and noise, but it would also allow us to model the variability at the *item* level, being it a single word or a sentence. This might provide additional information to the lexical resources, for example the uncertainty of an assigned label, and this additional information could be incorporated into the NLP models that we train with these very resources.

References

1. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on twitter. In Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23 (pp. 57-64). Springer International Publishing (2018)
2. Attanasio, G., Pastor, E. PoliTeam@ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets. In EVALITA. (2020)
3. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., ... , Uma, A.: We Need to consider disagreement in evaluation. In 1st Workshop on Benchmarking: Past, Present and Future (pp. 15-21). Association for Computational Linguistics (2021)
4. Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G. and Margetts, H.: An expert annotated dataset for the detection of online misogyny. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 1336-1350) (2021)
5. Kwarteng, J., Perfumi, S. C., Farrell, T., Third, A. and Fernandez M. Misogynoir: challenges in detecting intersectional hate. Social Network Analysis and Mining, 12(1), 166 (2022)
6. Nozza, D., Volpetti, C. and Fersini, E.: Unintended bias in misogyny detection. In Ieee/wic/acm international conference on web intelligence (pp. 149-155) (2019)
7. Tontodimamma, A., Fontanella, L., Anzani, S. and Basile, V.: An Italian lexical resource for incivility detection in online discourses. Quality & Quantity, 1-19 (2022)
8. Zeinert, P., Inie, N., Derczynski, L.: Annotating online misogyny. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)

Statistical analysis of COVID19 impact on Italian mortality

Analisi statistica dell'impatto del COVID19 sulla mortalità italiana

Girolamo Franchetti and Massimiliano Politano

Abstract This study proposes an alternative approach to measuring excess mortality due to COVID19 pandemics compared to the CEMC method. It investigates changes in biometric dynamics since the COVID-19 outbreak in 2019 by comparing empirical data from Italian mortality tables between 2011 and 2021 to counterfactual death probabilities derived from two commonly used statistical models, Lee-Carter and Renshaw-Haberman. Estimates are provided for 2019, 2020, and 2021, with an additional estimate for 2020 to 2021. The findings are presented to observe the dynamics of expected deaths over time.

Abstract *Questo studio propone un approccio alternativo rispetto a quello del CEMC per misurare l'eccesso di mortalità dovuto alla pandemia COVID19, indagando i cambiamenti nelle dinamiche biometriche dall'inizio della pandemia di COVID-19 nel 2019. I dati empirici delle tavole di mortalità italiane tra il 2011 e il 2021 vengono confrontati con le probabilità di morte controfattuali derivate da due modelli statistici comunemente utilizzati, Lee-Carter e Renshaw-Haberman. Vengono fornite stime per il 2019, il 2020 e il 2021, con un'ulteriore stima per il 2020 rispetto al 2021. I risultati sono presentati per osservare la dinamica del numero atteso di morti nel tempo.*

1 Introduction

Our colleagues at CEMC, COVID-19 Excess Mortality Collaborators, have conducted interesting research on their framework [4], which employs a combination

Girolamo Franchetti
University of Naples "Federico II", Economics and Statistics Science Department
e-mail: girolamo.franchetti@unina.it

Massimiliano Politano
University of Naples "Federico II", Economics and Statistics Science Department
e-mail: massimiliano.politano@unina.it

of six models to estimate excess deaths. The specific result in Italy is estimated to be 259,000 (with a confidence interval of 242,000 to 276,000).

However, we have concerns regarding the combination of these six models. The approach involves assigning reliability weights to each model, which is estimated based on pre-pandemic historical data. We believe that it may not be appropriate to assign weights to different models in this manner, as these subjective weights could introduce excessive increases in the estimate of deaths.

While the rationale for assigning more weight to the model with a more accurate estimate is understandable, we are concerned that this approach may introduce bias into the estimation, making it less reliable. As a result, we have opted to use classical actuarial models to estimate the probability of death and have conducted simulations to infer the estimates for each model.

This section of the study examines excess deaths through actuarial models based on mortality tables [1]. This approach allows for a direct observation of the impact of COVID-19 on the probability of death [5]. The results show an increase in excess deaths with age, peaking around age 80 before decreasing. The study uses the Lee-Carter and Renshaw-Haberman models with the R software package *StMoMo*[1], and the dataset is obtained from the I.Stat platform[2]. However, the dataset has a limitation in that the maximum age is an age class for those 100 years and older, so the weighted average of q_x is computed with the weight $l_x/\sum_{x=1}^{199} l_x$. All data and scripts are available on the OSF.io storage[3]. The framework involves using biometric panel data to determine the probability of death by age and year, converting this probability into mortality rates, and obtaining the matrix of death rates from 2011 to 2021 for each age. The estimated deaths are then compared with those based on mortality estimation models. The study also conducts simulations of death probabilities for the subsequent year in each part of the framework and shows significant results on confidence intervals through inference analysis.

2 Investigation

2.1 Probabilities

The Lee-Carter [2] and Renshaw-Haberman [3] models are regression models that provide average values of death probability and estimates of death rates conditioned by age. The term ”expectation” is used to refer to the probabilities estimated by these models. The study compares observed and estimated probabilities by age and observation period to assess how well the observations met the relative expectation.

---

[1] CRAN Manual available at `https://cran.r-project.org/web/packages/StMoMo/StMoMo.pdf`

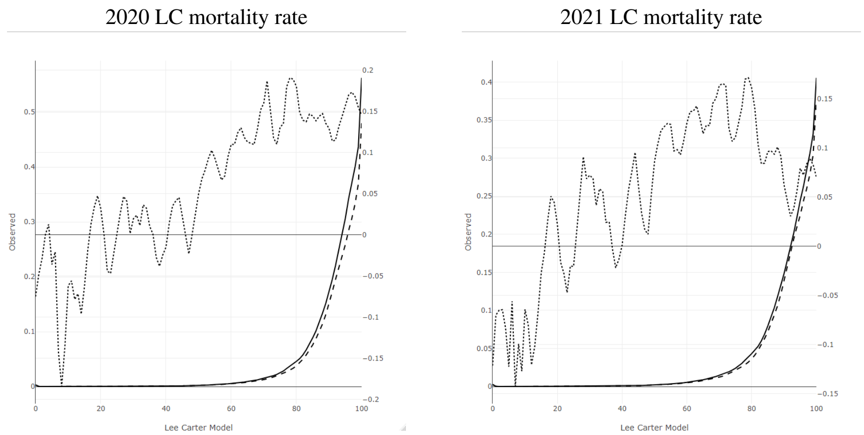
[2] ISTAT database available at `http://dati.istat.it/`

[3] Data storage available at `https://osf.io/9cqbm/`

From 2019 to 2020

Let's consider the first model, the Lee-Carter one. Here there is an interesting result.

Fig. 1



A graph (1) comparing mortality rates from 2019 to 2021 shows a cyclical pattern, with the mortality gap peaking at age 80 and decreasing at later ages. The smooth dynamics of the pattern are surprising, as is the reduction in mortality rates for younger age groups. We can then observe the death probabilities according to the Renshaw-Haberman Model:

The probability dynamics show a distinctive pattern, with a peak at ages 45-65 and a decline after age 85. Mortality rates for 2021 are generally as expected, except for a slightly lower rate at age 30 and unexpected events impacting mortality rates in certain age groups.

From 2020 to 2021

This constitutes a pivotal point of our study. While the findings may not be unexpected, they are of great significance in terms of their implications for predicting future outcomes.

The graphs show a negative trend in the gap due to newer data. COVID-19 had a less severe impact on mortality than expected, resulting in less observed mortality than estimated. The model predicts a wider negative gap in older ages and shows fluctuation in the mismatch, with higher mismatch in the first 20 years and widening mismatch in older ages. Therefore, it is expected that there were more deaths in 2020 than realized.

Fig. 2

2020 RH mortality rate | 2021 RH mortality rate

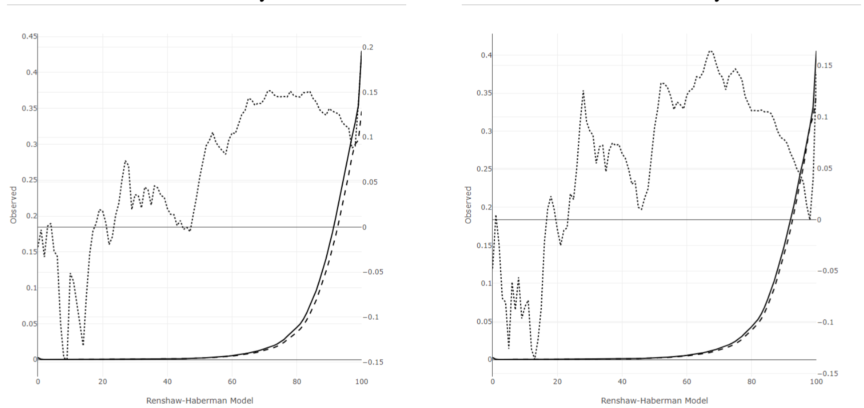
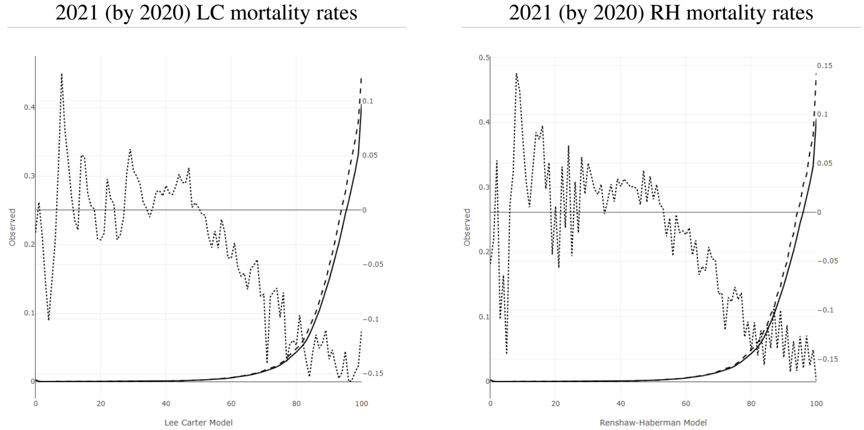


Fig. 3

2021 (by 2020) LC mortality rates | 2021 (by 2020) RH mortality rates



2.2 Death Excess

Estimating the expected deaths $\mathbb{E}[\frown] = Pop \cdot q_x$ for 2020 and 2021 based on the 2019 population and mortality rates, it is evident that the actual numbers exceeded the predicted figures, particularly for the 65+ age group. However, the change in impact from 2020 to 2021 must also be taken into account. The following graphs exhibit the excess deaths across age and time periods for the Lee-Carter and Renshaw-Haberman models, with numerical results presented in the table in Section 3:

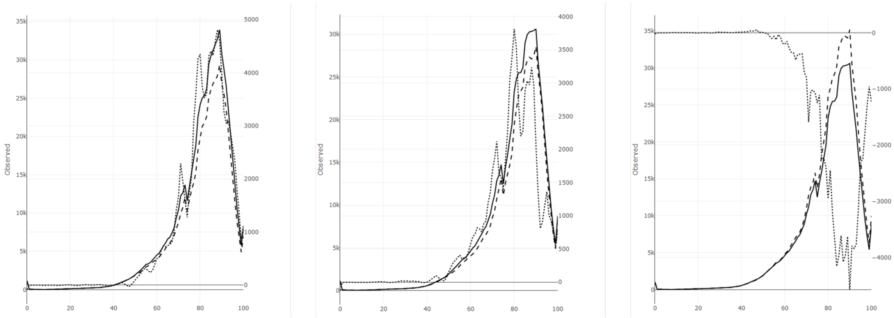


Fig. 4: LC excesses of deaths (left to right) 2020, 2021 (from 2019) and 2021 (from 2020)

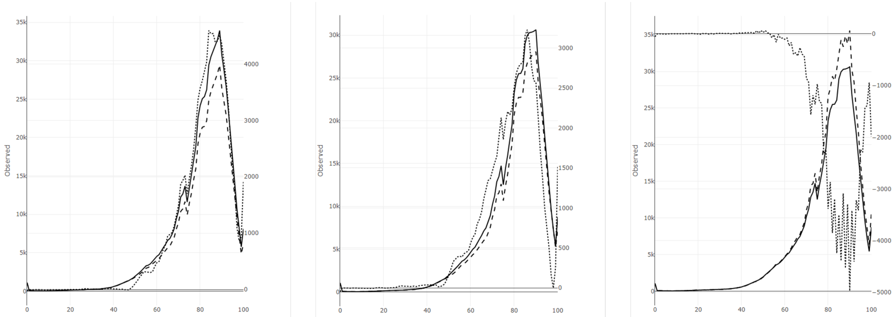


Fig. 5: RH excesses of deaths (left to right) 2020, 2021 (from 2019) and 2021 (from 2020)

3 Conclusion

In conclusion, based on the results obtained from the analysis of excess deaths and probability measures, it appears that there is a stronger excess of deaths particularly in the older age groups, as well as a weakening trend over time, suggesting a temporary effect of this shock on mortality [6, 7, 8]. Confidence intervals are computed based on the Normal-like shaped distribution of the death excesses to make estimations, and the tables below provide numerical results. Additionally, we found that the expected amount of deaths is significantly lower than CEMC estimations. The table above presents the relevant numerical results.

	LC		RH	
	Observed	Estimated	Observed	Estimated
2020	745856	643328	745856	643556
2021	712527	637963	712527	639137
2021 (from 2019)	712527	795983	712527	795451

Table 1: Total deaths counted for both models in each observation year

	Min	Mean	Max
2020	68578	68723	68868
2021	73266	73470	73674
2021 (by 2020)	-88874	-88336	-87798

Table 2: LC Model

	Min	Mean	Max
2020	67897	68070	68243
2021	71681	71924	72167
2021 (by 2020)	-88379	-87833	-87287

Table 3: RH Model

In contrast to the CEMC framework, we utilized the classical actuarial model to estimate the probability of death, and conducted simulations to derive estimates for each model. The results consistently show a lower probability of death, and highlight the maximum impact of events in 2019 on mortality. Although COVID-19 has resulted in increased mortality, our findings suggest that mortality may eventually revert to pre-pandemic levels. Nonetheless, further research is necessary to validate these findings and consider the potential influence of other factors on mortality rates.

References

1. Zhang, Y.: The negative impact of COVID-19 on life insurers. Geneva Papers on Risk and Insurance-Issues and Practice, 46(2), 209-232 (2021)
2. Lee, R. D. and Carter, L. R.: Modeling and Forecasting US Mortality. Journal of the American Statistical Association, 87(419), 659-671 (1992)
3. Haberman, S. and Renshaw, A.: Modelling and projecting mortality improvement rates using a cohort perspective. Insurance: Mathematics and Economics, 53, 150-168 (2013)
4. COVID-19 Excess Mortality Collaborators: Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21. The Lancet, 399(10304), 1907-1918 (2022)
5. Friedman, J., Liu, P., Troeger, C. E., et al.: Predictive performance of international COVID-19 mortality forecasting models. Nature Communications, 12, 2609 (2021)
6. Levy, C. and Cohen, R.: Infectious diseases in the COVID-19 era: gaps between countries. The Lancet Global Health, 10(2), e85-e86 (2022)
7. Tenforde, M. W. and Link-Gelles, R.: Reduction in COVID-19-related mortality over time but disparities across population subgroups. The Lancet Microbe, 4(5), e195-e196 (2023)
8. Johns Hopkins University: Mortality Analysis. Available via URL: `https://coronavirus.jhu.edu/data/mortality`. Cited on 14 May 2023.

Measuring multidimensional deprivation using objective and subjective data: an application of the Voronoi ranking method

Misurare la deprivazione utilizzando dati oggettivi e soggettivi: un'applicazione del metodo di ordinamento di Voronoi

Mariateresa Ciommi, Francesca Mariani, Gloria Polinesi and Maria Cristina Recchioni

Abstract Deprivation can be measured using either objective or subjective variables. However, these two approaches often result in different outcomes and inconsistent rankings. In this study, we propose using the Voronoi algorithm to integrate both subjective and objective dimensions and generate a single, unified ranking.
Abstract *La deprivazione può essere misurata mediante variabili oggettive o soggettive. Tuttavia, la maggior parte delle volte, i due approcci portano a risultati diversi e classifiche incoerenti. Qui proponiamo l'uso dell'algoritmo di Voronoi per combinare la dimensioni soggettiva e quella oggettiva e ottenere una classifica univoca.*

Key words: Voronoi algorithm, Ranking, Deprivation, double cut-off approach.

1 Introduction

Measuring well-being and deprivation is a complex issue. Over the past decade, in response to recommendations from the Stiglitz, Sen, and Fitoussi Commission [5],

Mariateresa Ciommi
Università Politecnica delle Marche, Piazzale Martelli 8, Ancona (IT), e-mail: m.ciommi@staff.univpm.it

Francesca Mariani
Università Politecnica delle Marche, Piazzale Martelli 8, Ancona (IT), e-mail: f.mariani@staff.univpm.it

Gloria Polinesi
Università Politecnica delle Marche, Piazzale Martelli 8, Ancona (IT), e-mail: g.polinesi@staff.univpm.it

Maria Cristina Recchioni
Università Politecnica delle Marche, Piazzale Martelli 8, Ancona (IT), e-mail: m.c.recchioni@staff.univpm.it

researchers have proposed two main approaches for overall well-being measurement: objective and subjective well-being.

In addition to these measures, there is another aspect that warrants exploration: local-level measurement, as noted by Ciommi et al. [3]. This paper combines these two lines of research and employs a new tool developed by Mariani et al. [4], which is based on the Voronoi algorithm [6], to obtain a new ranking that combines two dimensions.

Specifically, the method allows us to rank regions based on two dimensions when they provide different rankings. We focus on the Italian regions in 2020 and compute two measures of deprivation: one based on objective indicators and the other based on subjective indicators. We examine five aspects of deprivation, including health, social relationships, material deprivation, neighbor quality, and economic deprivation, using objective and subjective data collected by ISTAT in the "Aspects of Daily Life"(AVQ) survey.

The remainder of the paper is organized as follows. Section 2 outlines the data and methods used and provides a brief discussion of the main findings. Section 3 presents the conclusions drawn from this research and suggests potential avenues for future research.

2 Data, methods and results

We use the AVQ data.[1] The "Aspects of Daily Life" survey is a multipurpose survey on households conducted annually since 1993 by ISTAT. It involves interviews with a sample of 20,000 Italian households and nearly 50,000 individuals, providing population estimates on various topics related to daily life and behaviors. The survey collects information on the quality of life, satisfaction with living conditions, economic situation, and residential area, among other topics, and includes both objective and subjective questions.

The analysis has been conducted at household level, and sample weights are utilized. The data pertains to 2020, and the analysis is aggregated at the regional level. For each region, we compute two indices: objective and subjective poverty, implementing the "double cut-off" approach [1], as described in Castellano et al. [2].

Specifically, we consider five domains that account for different aspects of well-being: health, social relationships, material deprivation, neighbor quality, and economic deprivation. We analyze these domains from an objective perspective, using 15 AVQ variables and 9 objective variables, as well as from a subjective perspective, using 6 variables. For example, in the health domain, we consider an objective measure, the presence of chronic illnesses or long-term health problems, and a subjective measure, self-evaluation of health status.

[1] Data are available upon request to ISTAT. See https://www.istat.it/it/archivio/129959

Table 1 reports the conceptual framework. Variables are classified into objective and subjective. For each variable, we report the original codification provided by AVQ, a brief description and the scale of the variable. Moreover, we also report the cut-off adopted to transform individual data into $0/1$ value where 0 denotes that an individual is not deprived for that variable and 1 indicates deprivation.

Table 1 List of variables for the Objective and Subjective dimensions

Dim	AVQ code	Brief description	Scale	Deprived if
OBJ-Health	CRONI	Chronic illnesses or long-term health problems	Yes/No	Yes
SUB-Health	SALUT	Self evaluation of health	1 (very good) - 5 (very bad)	4; 5
OBJ-Soc Rel	PARENT; AMICI2; VICINI	Relatives (friends and neighbours) to count on	Yes - No	No (at least one "No")
SUB-Soc Rel	RELFAM; RELAM	Satisfaction with relationships with family (friends)	1 (a lot) - 4 (nothing)	3; 4 (at least one "3-4")
OBJ-mat depr	STANZEM	Number of room per person	Number	$\leq$ 1pp
SUB-mat depr	SPEAB	Housing costs	Yes - No	Yes
OBJ-neig qual	CRIM; INQAR; RUMORE	Area with risk of crime (pollution, crime)	1 (a lot) - 4 (nothing)	1; 2 (at least one "1-2")
SUB-neig qual	AMBIENTE	Satisfaction with the environmental situation	1 (a lot) - 4 (nothing)	3; 4
OBJ-ec depr	RISEC	Overall economic resources of the family	1 (excellent) - 4 (insufficient)	3; 4
SUB-ec depr	SITE	Satisfaction with the economic situation	1 (a lot) - 4 (nothing)	3; 4

In the case of domains such as Social Relationship, which account for more than one variable, we consider an individual as deprived if he or she has at least one form of deprivation.

We then adopt a row-column approach, where the row-value of each individual is aggregated across all individuals in the region, to obtain the subjective and objective indicators.[2] For each region and for each dimension, the final value ranges between 0 and 1, where 0 indicates no deprivation in the region, and 1 represents maximum deprivation. In this case, the index is a type of Headcount index, as it reflects the percentage of people living in a given region who experience multidimensional deprivation.

Table (2) reports summary statistics for the objective and subjective dimension whereas figure (1) displays the corresponding ranking. The subjective dimension has smallest values compared with the objective one. The plot suggests that the two

[2] Here, we assume equal weights, meaning that each domain has the same weight.

Table 2 Descriptive Statistics

Dim	Min	Max	I Quar	III Quar	Mean	Median	Stdev	Skew.	Kurtosis
Obj	0.169	0.402	0.236	0.335	0.283	0.281	0.070	-0.046	-1.227
Sub	0.089	0.268	0.170	0.220	0.193	0.189	0.043	-0.268	-0.234

dimensions rank the regions differently. This is also confirmed by computing the Kendall rank correlation for data that are not normally distributed. The correlation coefficient is 0.547 with a p-value of 0.00048, indicating a significant correlation. Almost all regions are far from the line of perfect equity (red line in Figure 1), except for Campania (CAM) and Sicilia (SIC). This means that individuals in Italian regions have different evaluations of their multidimensional deprivation depending on whether an objective or subjective point of view is considered. For example, Trentino Alto Adige (TAA) is one of the regions with the lowest percentage of deprived people according to an objective index, but they feel worse in terms of the subjective dimension. Lombardia (LOM), Emilia Romagna (EMR), and Liguria (LIG) exhibit a similar pattern. Conversely, Calabria shows the opposite behavior:

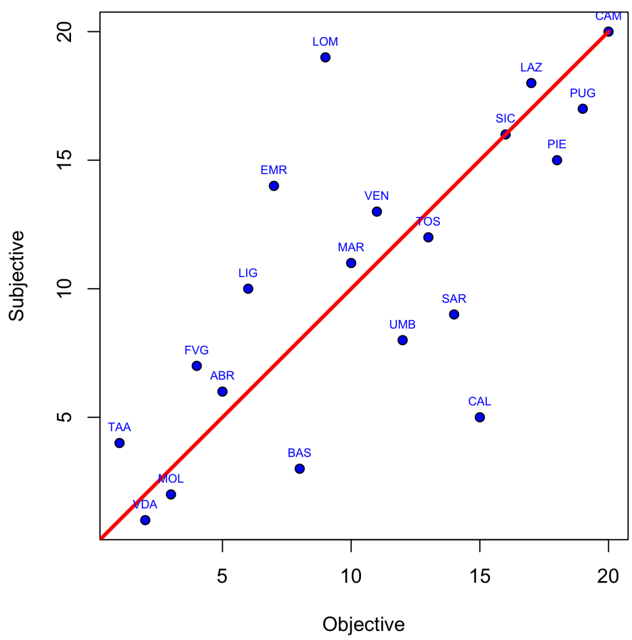


Fig. 1 Comparisons between Objective and Subjective ranking

it is very deprived according to the objective index, but according to the subjective index, it is among the top-five regions.

A situation like this highlights the need to introduce a tool that combines the two dimensions. However, in this case, simply aggregating them by taking the arithmetic mean wouldn't be appropriate. Therefore, we apply the Voronoi algorithm, as developed by Mariani et al. [4]. This is an iterative procedure that is used to divide the two-dimensional subspace into distinct non-overlapping convex polygons, known as Voronoi cells. Each ranked point is associated with one cell.

The Voronoi algorithm, shown in Figure (2), identifies the regions with the lowest level of deprivation as Valle d'Aosta (VDA), Trentino Alto Adige (TAA), and Molise (MOL). Next are Basilicata (BAS), Abruzzo (ABR), Friuli Venezia Giulia (FVG), Umbria (UMB), Sardegna (SAR), Liguria (LIG), Marche (MAR), Toscana (TOS), Calabria (CAL), Veneto (VEN), Emilia Romagna (EMR), Sicilia (SIC), Piemonte (PIE), Lazio (LAZ), Lombardia (LOM), Puglia (PUG), and Campania (CAM), with Campania being the most deprived region.

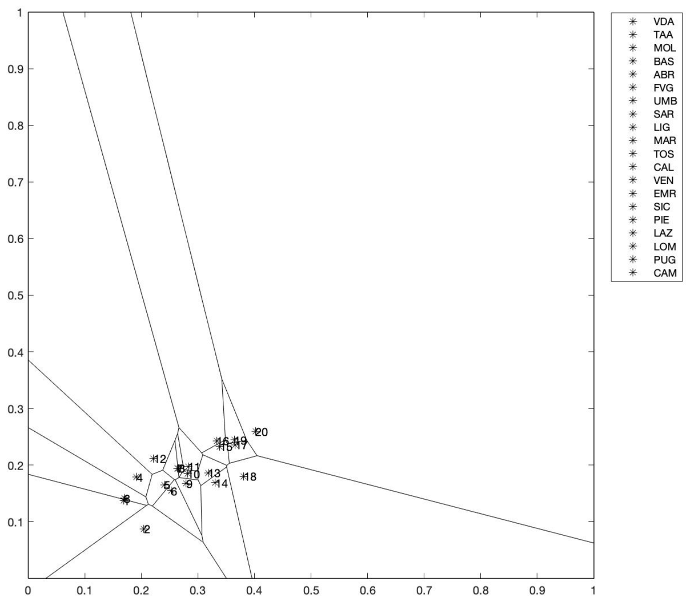


Fig. 2 Voronoi result.

3 Conclusions and further research

We applied an iterative procedure based on the Voronoi partition to combine two dimensions. This approach is to be employed when aggregative methods, such as the arithmetic mean, are not appropriate.

While this analysis provides a first attempt at combining objective and subjective measures, there is still room for further research. For example, an empirical analysis could be conducted to investigate whether the subjective dimension is consistently smaller than the objective dimension over time, particularly in the context of the COVID-19 pandemic and its impact on the health domain. Additionally, analyzing specific subgroups of the population may provide a better understanding of deprivation.

From a theoretical perspective, it may be of interest to generalize the methods for more than two dimensions and for dimensions that are not in the $[0,1]$ range. Overall, this study highlights the importance of considering both objective and subjective dimensions when assessing multidimensional deprivation and emphasizes the need for further research in this area.

References

1. Alkire, S., Foster, J.: Counting and multidimensional poverty measurement. Journal of public economics. 95(7-8), 476-487 (2011)
2. Castellano, R., Chelli, F. M., Ciommi, M., Musella, G., Punzo, G., Salvati, L.: Trahit sua quemque voluptas. The multidimensional satisfaction of foreign tourists visiting Italy. Socio-Economic Planning Sciences. 70, 100722 (2020)
3. Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., Chelli, F. M.: A new class of composite indicators for measuring well-being at the local level: An application to the Equitable and Sustainable Well-being (BES) of the Italian Provinces. Ecological indicators. 76, 281–296 (2017)
4. Mariani F., Ciommi M., Recchioni M.C.: Two in one: a new tool to combine two rankings based on the Voronoi diagram. Mimeo (2023)
5. Stiglitz, J. E., Sen, A., Fitoussi, J. P.: Report by the commission on the measurement of economic performance and social progress (2009). Available at https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf
6. Voronoi, G. F.: Nouveles applications des paramétres continus à la théorie de formas quadratiques. Journal fur die Reine und Angewandte Mathematik. 134, 198—287 (1908)

Session of free contributes SFC4 - *Trends*

1. *The role of big data analytics in circular supply chains: A bibliometric analysis* (Khan F. and Rapposelli A.)
2. *Estimation of the ranking of incentive policies for the adoption of 4.0 technologies* (Bonnini S. and Borghesi M.)
3. *Risk Management and Future Scenarios. A proposal based on a mixed-method approach* (Di Zio S., Bolzan M., Marozzi M. and Scioni M.)
4. *Explainable artificial intelligence (XAI) through artificial intelligence from a human in the loop (HITL) perspective: an interview with ChatGPT* (Santarcangelo V., Lamacchia A., Vitullo S., Di Lecce M. and Giacalone M.)
5. *Relevance in official statistics: information needs, satisfaction with data quality, some results and future perspectives* (Loporcaro M.F.)

The role of big data analytics in circular supply chains: A bibliometric analysis

Il ruolo dei big data analytics nelle supply chain circolari: un'analisi bibliometrica

Feroz Khan and Agnese Rapposelli

Abstract The concept of circular supply chain management and data analytics is getting considerable attention from researchers. Therefore, to understand where the literature stands and identify relevant literature gaps, this study aims to investigate the role of big data analytics on circular supply chains by conducting a bibliometric analysis. The Scopus database in this case is referred to gather the relevant dataset and analyze it from the perspective of descriptive and social network analysis. The results obtained identify important themes and methodologies prevalent to the topic at hand which in turn would aid the researcher to further contribute to the body of literature.

Abstract *Il concetto di gestione della supply chain circolare e di data analytics sta ricevendo una notevole attenzione da parte degli studiosi. Per comprendere a che punto è la letteratura ed identificarne le possibili lacune questo studio si propone di indagare il ruolo dei big data analytics sulle supply chain circolari conducendo un'analisi bibliometrica, utilizzando il database Scopus per ottenere il dataset, analizzato sia attraverso indicatori descrittivi che attraverso tecniche di social network analysis. I risultati ottenuti hanno individuato temi e metodologie prevalenti per l'argomento in questione, che si confida possano fornire un utile contributo alla letteratura.*

Key words: Sustainable development, Supply Chain Management, Circular Economy, Decision-making, Big data analytics, Bibliometric analysis.

Feroz Khan

"G. D'Annunzio" University of Chieti-Pescara; e-mail: feroz.khan @unich.it

Agnese Rapposelli

"G. D'Annunzio" University of Chieti-Pescara; e-mail: agnese.rapposelli@unich.it

1 Introduction

The awareness by the United Nations through their sustainable development goals (SDGs) has forced countries and industries around the world to move towards cleaner production practices. However, to enact such goals, companies need to upgrade their supply chains to reduce the waste being generated both in the upstream and downstream supply chains. A more rational approach that recently has obtained more interest is the incorporation of Circular Economy (CE) principles in supply chain operations. The CE concept works on the 3R principles approach i.e., reduce, reuse, and recycle that contributes to the triple bottom line (social, economic and environmental sustainability). The notion of decision-making in such intricate supply chains has become of primary importance. Therefore, in order to facilitate the supply chain managers, the role of big data analytics has been considered as a promoter for enhancing the performance of firms that are engaged in the area of circular supply chains. However, the literature on this research field is still nascent, especially with regard to the application of such approach. Hence, this study aims to carry out a bibliometric analysis concerning the role of big data analytics on circular supply chains. The outcomes of this research would help identifying important themes existing in the literature as well as literature gaps for future research directions in the context of big data analytics and circular supply chains.

The study unfolds as follows. Section 2 describes the data collection and the methodology used. Sections 4 presents the results and Section 4 concludes.

2 Data collection and methodology

The methodology in this kind of literature review provides the following three steps: dataset setting, data refining and data analysis. The first step involves the collection of the relevant papers on which the analysis will be conducted. In this case, we referred to the Scopus database to gather the desired dataset. We tried different keywords to capture the relevant set of papers. The finalized keywords included: ("big data analytics") AND ("circular supply chains" OR "circular business models" OR "reverse logistics" OR "waste management" OR "sustainability" OR "optimization" OR "efficiency"). The initial query returned 1948 articles. The second step was dedicated to the refining of the list. In the first phase, we considered articles ranging from the last 10 years (2013-2022) but since there were no articles pertaining to the topic considered before 2015, we limited our search to the period 2015-2022. The language of the articles was limited to English only. The subject areas considered were decision sciences, business management, and accounting and environmental sciences. The application of the aforementioned filters shortlisted the number of articles to 253. However, for ensuring the relevance of the papers we carried out a data cleaning by going through the titles and abstract of each paper. At the end of the process, the sample was composed of 138 articles. It is pertinent to mention here that the list of 138 articles did not entirely fall under the scope of the desired studies but had some of the links that would facilitate in understanding the relationship between big data analytics and supply chain management (SCM). This was followed by checking for the articles with missing keywords. In this case, we found three articles which had missing keywords. These missing keywords were formulated by going through the abstracts of each article.

Finally, the keywords were homogenized. Before the homogenization of keywords, the list consisted of 481 articles. After the refining process, the final list of keywords reduced to 334 keywords.

3 Results

The search for the relevant articles was conducted in the last quarter of 2022. The analysis of the selected publications involves two steps: 1) the analysis of some descriptive performance indicators (the trend of the number of publications and citations per year); 2) the use of Social Network Analysis (SNA) tools to explore the connection between the topics and the most influential references.

3.1 Descriptive analysis

The descriptive analysis is conducted in order to understand the trend of publications over the years and the associated citations.

3.1.1. Publication by year

The topic in question is fairly new: in 2015 there is only one article that matches the search string. The number of articles increases over the years until 2021, when the greatest number of articles was published (40 papers). Since the search in the Scopus database was conducted in the last quarter of 2022, the number of articles published in that year was 19 articles. The overall growth of the articles published during the period investigated is 52.29%, thus indicating a decent interest in the topic by the researchers.

3.1.2. Citation trend and productivity

The year where the greatest number of citations were recorded was 2017. In this case, the most productive authors are Anil Kapoor and Weisheng Lu producing 9 articles over the period 2017-2022. The most cited paper (cited over 1007 times), published in the Journal of Business Research, is co-authored by Uthayasankar Sivarajah. The paper carries out a structured literature review to conduct a critical analysis of big data analytics and its importance in organizations also looking at its application in supply chains [4]. The most productive journals are Sustainability by MDPI and Journal of Cleaner Production by Elsevier, producing 23 and 17 articles respectively. Moreover, the most productive countries are found to be China and India followed by the UK.

3.2 Investigation based on Social Network Analysis tools

In order to examine the most relevant topics in the data set, we first carried out the co-occurrence analysis of the keywords. Secondly, we created a thematic map, which is analyzed to study the important themes in the literature and their standing and maturity over time.

3.2.1. Co-occurrence network

Co-occurrence network helps in identifying important themes and topics that might interest the researchers. In the network, the size of each node and its label showcase the occurrences of keywords that appear within the text i.e., in how many papers it appears. The keywords in the network are nodes and a tie is considered between two nodes if they are mentioned together in the same paper (called co-occurrence). For the construction of the co-occurrence network, the minimum number of keywords was set to 5 in VOS viewer software (Figure 1).

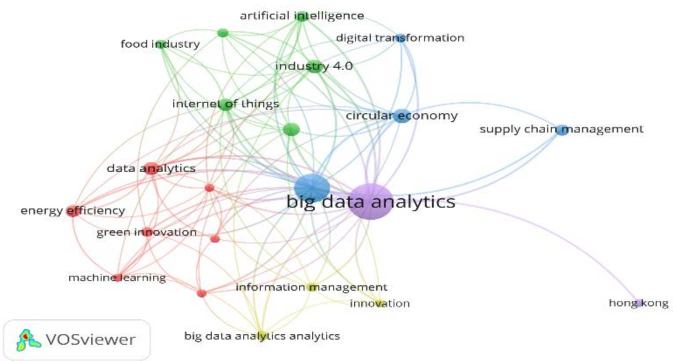


Fig. 1 Co-occurrence network

The graph shows three main clusters. The green cluster focuses on industry 4.0 capabilities and the food industry. These studies aim to focus on reducing food loss and waste in the agri-food supply chains by incorporating circular economy principles. The use of industry 4.0 capabilities, in this case, is to reduce food waste and focus more on sharing and optimization of such resources by taking advantage of circular economy models. Moving towards the red cluster, the studies mostly focus on material and energy efficiency management by utilizing the capabilities of machine learning and data analytics. Energy management and efficiency in this case has been considered important sustainability criterion fostering the circular economy trend through the application of industry 4.0 capabilities [6]. Hence, energy efficiency and management represent fields that are getting interest in recent times in the context of circular supply chains [5]. The blue cluster focuses on big data analytics and digital transformation capabilities and their impact on circular supply chains. Research studies, in this case, focus on how digital transformations are contributing to circular business models. The role of digital transformation and big data analytics is deemed to be essential for effective decision-making processes and the optimization of supply chain operations [2].

3.2.2. Thematic map

The thematic map helps to identify important topics and their standing in the literature, as it identifies clusters of keywords and the connections between them by determining density and centrality parameters. The density identifies the strength of intrinsic ties between all the keywords in a cluster, thus identifying the "theme's development". The centrality shows the strength of the connection between different clusters, thus showcasing the importance of the research theme.

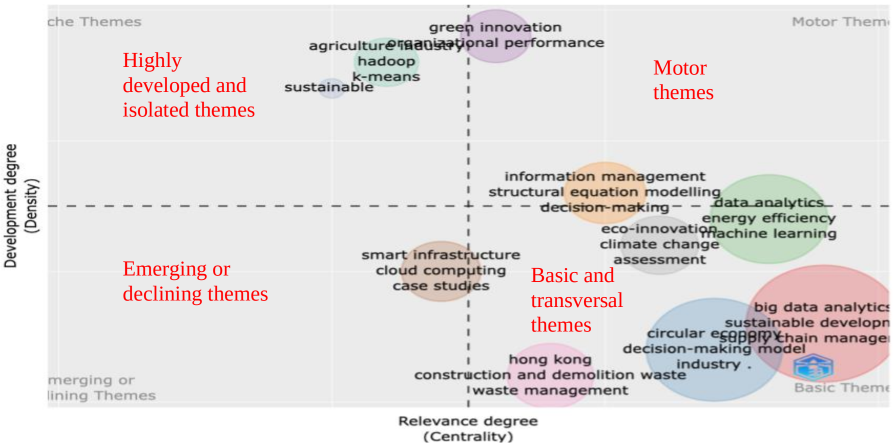


Fig. 2 Thematic map

By observing Figure 2 starting from the first quadrant, the themes in this quadrant (motor themes) are considered to be well-established, especially from a theoretical perspective. In this case, the concepts of green innovation and organizational performance are considered to be well-developed. Moreover, topics such as information management along with the structural equation modeling methodology are regarded as well-established in the literature. The decision-making approaches in this case are also regarded as equally important and developed. A supporting study in this regard can be the investigation of the role of eco-innovation on the organizational performance of small and medium firms in the context of green and circular supply chains. The study in this regard collects primary data and draws results based on institutional theory utilizing the structural equation modeling approach [1]. The themes in the second quadrant are considered to be well-developed and specialized but hold unimportant external ties which makes them of minor importance for the field. In this case, the themes are prevalent in the agricultural sector. The use of data analytics tools and the use of Hadoop and k-mean clustering are covered in the literature for handling large sets of data but have lesser importance as it has lesse linkage with the topic at hand. The themes in the third quadrant are called emerging or declining fields and therefore are considered to be of marginal importance and weakly established. In this case, most of the case studies correspond to the utilization of smart infrastructure and cloud computing. Finally, the themes in the fourth quadrant are considered to be both basic and transversal. These topics are relevant for specific fields and are still in the process of development. The themes primarily correspond to the waste management sector, particularly the construction and demolition waste [3].

Other topics include the application of decision-making models, especially multi-criteria decision-making tools, towards the implementation of circular economy models and supply chains, the concepts of eco-innovation and climate change [1], and the usage of data analytics tools such as machine learning for acheiving energy efficiency [5, 6].

4 Conclusion

Recently, most of the studies in the literature are trying to understand the role of industry 4.0 capabilities or data analytics in overcoming the challenges in circular supply chains. Data analytics is considered to be an important driver in optimizing decision-making processes. To further understand the standing of the topic at hand, this study conducted a literature review (a bibliometric analysis) to identify important themes in the literature. The data have been collected within Scopus database considering the period 2015-2022. The results obtained highlights important themes, concepts and methodologies in the literature that might aid the researchers in the moving forward with the topic at hand. Firstly, with regard to the sectors more highligted in the literature, big data analytic tools have been applied in the food and agriculture sectors, construction, waste management and energy sectors. However, other sectors such as mining, cement, textiles, healthcare, bioenergy, and packaging needs to be investigated. Secondly, the most recurring keywords are “Internet of things”, “big data analytics”, “artificial intelligence”, “machine learning”, “green innovation”, “digital transformation” and “circular economy”, “green innovation” and “organizational or sustainable performance”. The most productive countries are China, India and the UK which indicates that the topic at hand is obtaining interest at global level. Finally, with regard to the methodologies used, most of the studies consider structural equation modelling, followed by decision-making models such multi-criteria decision making tools. For further studies, the literature points towards the direction of implementing quantitative methods for assessing the intersection between industry 4.0, circular economy and supply chain management especially through decision-making models.

References

1. Bag, S., Dhamija, P., Bryde, D.J., Sing, R.K.: Effect of eco-innovation on green supply chain management, circular economy capability, and performance of small and medium enterprises. Journal of Business Research, 141(1), 60-72 (2022)
2. Dwivedi, A., Paul, S.K.: A framework for digital supply chains in the era of circular economy: Implications on environmental sustainability. Business strategy and the Environment, 31(4), 1249-1274 (2022)
3. Khan, F., Ali, Y.: Implementation of the circular supply chain management in the pharmaceutical industry. Environment, Development and Sustainability, 24(1), 13705–13731 (2022)
4. Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V.: Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70(1), 263-286 (2017)
5. Taddei, E., Sassanelli, C., Rosa, P., Terzi, S.: Circular supply chains in the era of industry 4.0: A systematic literature review. Computers & Industrial Engineering, 170(1), 108268 (2022)

6. Tseng, M.L.: Building a data-driven circular supply chain hierarchical structure: Resource recovery implementation drives circular business strategy. *Business strategy and the Environment*, 31(5), 2082-2106 (2022)

Estimation of the ranking of incentive policies for the adoption of 4.0 technologies

Stima del ranking di politiche incentivanti per l'adozione di tecnologie 4.0

Stefano Bonnini and Michela Borghesi

Abstract The main goal of this work is to investigate the relationship between technology 4.0 adoption and some policy incentives. A nonparametric method for ranking combinations of policy incentives with respect to the propensity to adopt 4.0 technologies, was applied to an original dataset from a survey carried out in the Region Emilia-Romagna (Italy).
Abstract *L'obiettivo principale di questo lavoro è quello di studiare la relazione tra l'adozione di tecnologie 4.0 e alcuni incentivi pubblici. Un metodo non parametrico per determinare il ranking di diverse combinazioni di incentivi basato sulla propensione delle aziende ad adottare tecnologie 4.0 è stato applicato ad un dataset originale condotto nella regione Regione Emilia-Romagna (Italia).*

Key words: 4.0 technologies, nonparametric inference, ranking, pairwise comparisons.

1 Introduction

In Italy, Industry 4.0 is becoming a much-debated topic. It is connected to the adoption of a combination of established and new technologies [8]. The Industry 4.0 Plan represents a great opportunity for companies that would like to take advantage of the

---

Stefano Bonnini
University of Ferrara; e-mail: stefano.bonnini@unife.it

Michela Borghesi
University of Ferrara; e-mail: michela.borghesi@unife.it

opportunities related to the fourth industrial revolution. This Plan provides for a set of complementary measures capable of promoting investments for innovation and competitiveness. Industry 4.0 invests in all aspects of the life cycle of companies that want to acquire competitiveness, offering support in investments, in the digitization of production processes, in the enhancement of worker productivity, in the training of adequate skills and in the development of new products and processes [3].

The main goal of this work is to investigate the relationship between technology 4.0 adoption and policy incentives. We focus on the Region Emilia-Romagna, one of the most developed and productive regions of Italy. In fact, in this Region, several companies have embraced the 4.0 paradigm and it represents an important case study to verify how much the evolution of regional institutions has favored the creation of a system capable of promoting innovative capacity [11].

In the empirical literature, it is shown that the application incentives can facilitate the adoption of Industry 4.0 [5]. The policy incentives taken into account in this paper concern hyper and super depreciation and Nuova Sabatini, in fact, these incentives are considered the most relevant because they can be linked directly to business spending [4]. The first incentive is useful for supporting and encouraging companies that invest in new capital goods, in tangible and intangible assets (software and IT systems) useful for technological and digital transformation. Hyper-depreciation has the advantage of overestimating by 250% investments in new material assets, devices and technologies enabling transformation 4.0. Super depreciation, on the other hand, overestimates 140% of investments in newly acquired capital goods. For those who take advantage of hyper-depreciation, there is the possibility of using the facilitation also for investments in intangible capital goods (software and IT systems). As for the "Nuova Sabatini" law, it is used to support companies that request bank loans for investments in new capital goods, machinery, plants, factory equipment for production use and digital technologies (hardware and software). This incentive is intended as a contribution to the partial coverage of the interest paid by the company on bank loans granted by banks affiliated with the MISE. The contribution is calculated on the basis of a conventional 5-year amortization plan with an annual interest rate which is increased by 30% for investments in Industry 4.0 technologies. Finally, it also has the advantage of having priority access to the Central Guarantee Fund up to a maximum of 80% [3].

This work represents a contribution to the empirical literature on the effectiveness of public policy interventions on the adoption of 4.0 technologies by small and medium enterprises. Specifically, the goal is to determine a ranking of the main policy interventions based on sample data concerning the adoption of 4.0 technologies by firms of the Emilia-Romagna Region. We apply a nonparametric inferential method that consists of multiple pairwise comparisons based on two-sample multivariate permutation tests. Section 2 focuses on the presentation of the statistical problem and of the methodological solution. The results of the application of the proposed method to the case study of Emilia-Romagna are shown and discussed in Section 3. Section 4 is dedicated to concluding remarks.

2 Statistical problem and methodological solution

The goal of this study is to investigate the effect of some public policies on the adoption of 4.0 technologies. In particular, we want to create a ranking of policy incentives based on the proportion of technologies 4.0 adopted. The idea of ranking occurs whenever the goal in a study is to determine an ordering among several input conditions with respect to one or more outputs of interest.

Let us consider data drawn from each of C multivariate populations $\Pi_1, \dots, \Pi_C$ with $C > 2$, by means of a sampling procedure, so as to make an inference on their possible equality and, in case of rejection of this hypothesis, to classify those populations in order to obtain a relative ranking from the best to the worst according to prespecified criteria [6]. The ranking is relative because the ordering is referred only to the C populations under study.

The problem can be formalized in a nonparametric way [1]. Let $\boldsymbol{Y}$ be the p-dimensional vector representing the response variable from population Π. Furthermore, without loss of generality, assume that large values of each univariate component of $\boldsymbol{Y}$corresponds to a better marginal performance. We want to rank the multivariate populations $\Pi_1, \dots, \Pi_C$, with respect to p marginal variables, by using the information of C samples, one from each population. Let us suppose that, in the j-th sample, m_j independent vector of values are observed. Such vectors are supposed to be determinations of the random variables $\mathbf{Y}_1, \dots, \mathbf{Y}_C$, $j = 1, \dots C$. The rank $r(\Pi_j)$ can be calculated by:

$$\begin{aligned} r_j = r(\Pi_j) &= 1 + \textstyle\sum_{j \neq h} I\left(\boldsymbol{Y}_j >^d \boldsymbol{Y}_h\right) = \\ &= 1 + \left\{\# \boldsymbol{Y}_j >^d \boldsymbol{Y}_h, h = 1, \dots, C, j \neq h\right\}, j = 1, \dots C, \end{aligned} \tag{1}$$

where $I(\cdot)$ is the indicator function and # means the number of times [1,6]. Note that, in Equation (1), the rank is derived by using the concept of stochastic ordering and by simple pairwise comparisons. Once the matrix $A = [a_{ij}]$ of the p-values resulting from the directional pairwise comparisons has been obtained, 1 is inserted in place of the significant p-values (generally when they are less than $\alpha = 0.05$) and 0 otherwise. Finally, using Equation (1), we can obtain the rank by summing up the ones and the zeros along the rows [7].

Considering the univariate case, the p-values corresponding to positions (i, j) and (j, i) of matrix A add up to one, i.e. $a_{ij} + a_{ji} = 1$ (for $i \neq j$), while in the multivariate case this usually does not happen. For the multivariate pairwise comparisons, a solution based on the methodology of Combined Permutation Tests (CPTs) is applied [2]. CPTs consists of a family of permutation tests, appropriate for complex testing problems that can be broken down into partial tests. This approach is distribution-free and consequently robust with respect to the underlying (multivariate) distribution and powerful, also for small sample sizes. The iid condition is not required, because the only assumption on which the test is based, is that of exchangeability of the individual vectors of observed values between samples. This condition, under the null hypothesis of equality in distribution, is always satisfied and it is milder than that of independence and identical distribution [12, 13]. A suitable combination of the p-values of the partial tests, according to the following formula, can be used as a test statistic for the multivariate two-sample test in each pairwise comparison and for the general multi-sample problem: $T_{comb} = -\sum_q \ln(\lambda_q)$ (see [2]).

3 Application

The ranking method presented in the previous section was applied to original data collected in a sample survey carried out in January 2022. The survey was conducted in the northern regions of Italy by the Department of Economics and Management of the University of Ferrara. It was aimed at manufacturing enterprises in the North of Italy. The total number of interviewed companies was 3926, but we specifically focused on the Region Emilia-Romagna, considering only the 613 companies from this Italian Region.

The goal is to investigate the specific role of recent public policies, in enhancing the innovative capacity of companies regarding Industry 4.0 technologies. As we reported in Section 1, we considered hyper and super depreciation and Nuova Sabatini as policy incentives. The 4.0 technologies considered in the study, in the two-year period 2018-2019, were advanced manufacturing solutions (interconnected and programmable robots) and horizontal integration (integration of information along the production process stages). Each of the two types of technology, is related to a binary response that takes one if a company adopted that technology in the mentioned period and zero otherwise. Hence, we are in presence of a bivariate binary response. These two types of innovation are among the most important and able to generate a perfectly integrated production flow [15].

We compared the following four groups of companies, according to possible combinations of policy incentives (symbolic treatments):

- Group 1: companies that have benefited from both state incentives,
- Group 2: companies that have only benefited from the hyper and super depreciation incentive,
- Group 3: companies that have only benefited from the Nuova Sabatini incentive,
- Group 4: companies that have not benefited from either of the two incentives.

Through the application of the methodology presented in Section 2, we obtained the unadjusted p-values for the multivariate pairwise comparisons reported in Table 1. Since the global p-value of the test (obtained by combining the partial p-values) is equal to **0.0002**, at the significance level $\alpha = 0.05$, there is empirical evidence in favor of the hypothesis that the adoption of 4.0 technologies is affected by the policy incentives.

Table 1 Table of unadjusted p-values (significant in bold).

>	***Group 1***	***Group 2***	***Group 3***	***Group 4***
Group 1	-	**0.00024995**	**0.00014997**	**0.00014997**
Group 2	0.99865027	-	0.13312338	**0.00014997**
Group 3	0.99985003	0.92616477	-	**0.00714857**
Group 4	0.99985003	0.99985003	0.99705059	-

The significance in the overall test can be attributed to one or more specific pairwise comparisons, after adjustment of the partial p-values to control the family-wise error (FWE) for the multiplicity [14]. Hence, the p-values of the partial tests were adjusted with the Bonferroni-Holm method, which controls the FWE in a strong sense [9, 10] (see Table 2).

Table 2 Table of adjusted p-values with Bonferroni-Holm method (significant in bold).

>	***Group 1***	***Group 2***	***Group 3***	***Group 4***
Group 1	-	**0.00224955**	**0.00179964**	**0.00179964**
Group 2	1.00000000	-	0.93186363	**0.00179964**
Group 3	1.00000000	1.00000000	-	0.05718856
Group 4	1.00000000	1.00000000	1.00000000	-

Then we replaced the significant p-values with ones and the others with zeros in order to estimate the rank of each of the four treatments by summing up the values by row and adding one.

Table 3 Table for the computation of the ranks.

>	***Group 1***	***Group 2***	***Group 3***	***Group 4***	***Rank***
Group 1	-	1	1	1	**4**
Group 2	0	-	0	1	**2**
Group 3	0	0	-	0	**1**
Group 4	0	0	0	-	**1**

According to the rank of the four groups (see Table 3), we can conclude that, when both incentives are present, there is the highest propensity to adopt the considered technologies. The second highest propensity of adoption of 4.0 technologies concerns companies that used the hyper and super depreciation incentive (Group 2). In the last position of the ranking, we find the group of companies that took advantage only of the “Nuova Sabatini” incentive (Group 3) and the group in which neither of the two considered incentives has been used (Group 4).

4 Conclusions

The presented nonparametric methodology for ranking multivariate populations, based on the application of combined permutation tests and pairwise comparisons, is a distribution-free, robust and flexible statistical solution. Its application to an original dataset concerning a survey about Italian enterprises in Emilia-Romagna provides empirical evidence in favor of the hypothesis that the adoption of industry 4.0 technologies depends on specific policy incentives such as hyper and super depreciation and Nuova Sabatini. In particular, those companies who have adopted both incentives show the highest propensity to adopt 4.0 technologies. The companies with the lowest propensity towards 4.0 technologies are those who took advantage only of the "Nuova Sabatini" incentive and those that exploited none of the two incentives.

Acknowledgments

Authors thank University of Ferrara that funded the project entitled "Public policies, 4.0 technologies and enterprise performance. Empirical analyses on a representative sample of manufacturing enterprises of northern Italy (Politiche pubbliche, tecnologie 4.0 e performance d'impresa. Analisi empiriche su un campione rappresentativo di imprese manifatturiere del Nord Italia)" for the period 2022-2024, with the Departmental Research Incentive Fund - FIRD 2022.

References

1. Arboretti Giancristofaro, R., Bonnini, S., Corain, L., Salmaso, L.: A permutation approach for ranking of multivariate populations. J. Multivar. Anal., 132, 39-57 (2014)
2. Bonnini, S., Corain, L., Marozzi, M., Salmaso, L.: Nonparametric hypothesis testing. Rank and permutation methods with applications in R. Wiley (2014)
3. Calenda, C. (2017). Piano Nazionale Industria 4.0. Ministero Dello Sviluppo Economico: Roma, Italy
4. Capuano, G., & Capuano, M.: Aspetti metodologici ed evidenze empiriche della valutazione d'impatto di "Industria 4.0" nel settore manifatturiero italiano. Argomenti, (16), 7-23 (2020)
5. Cugno, M., Castagnoli, R., Büchi, G.: Openness to Industry 4.0 and performance: The impact of barriers and incentives. Technol. Forecast. Soc. Change. 168 (2021) 120756 (2021) https://doi.org/10.1016/j.techfore.2021.120756
6. Corain, L., Arboretti, R., Bonnini, S.: Ranking of multivariate populations: A permutation approach with applications. CRC press (2017)
7. Corain, L., Salmaso, L.: A nonparametric method for defining a global preference ranking of industrial products. J. Appl. Stat. 34(2), 203-216 (2007)
8. Corò, G., Volpe, M.: Driving factors in the adoption of Industry 4.0 technologies: An investigation of SMEs. In Industry 4.0 and regional transformations (pp. 112-132). Routledge (2020)
9. Giacalone, M., Zirilli, A., Cozzucoli, P., Alibrandi, A.: Bonferroni-holm and permutation tests to compare health data: methodological and applicative issues, BMC Med. Res. Methodol. 18:81 (2018)
10. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70 (1979)
11. Mosconi, F., D'Ingiullo, D.: Institutional quality and innovation: evidence from Emilia-Romagna. Econ. Innov. New Technol (2021). doi: 10.1080/10438599.2021.1893140
12. Pesarin, F.: Nonparametric Combination Methodology. In Multivariate Permutation Tests with Applications in Biostatistics, 2nd ed.; Wiley: Chichester, UK (2001)
13. Pesarin, F., Salmaso, L.: Permutation tests for complex data: applications and software. Wiley series in probability and statistics (2010)
14. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment; Wiley-Interscience: New York, NY, USA (1992)
15. Zanotti L.: Industria 4.0: storia, significato ed evoluzioni tecnologiche a vantaggio del business. Network Digital 360, 13 maggio 2021

Risk Management and Future Scenarios. A proposal based on a mixed-method approach

Gestione del rischio e scenari futuri? Una proposta basata su un approccio quali-quantitativo

Simone Di Zio, Mario Bolzan, Marco Marozzi and Manuela Scioni

Abstract The rapid changes in society and the risks related to disastrous and unexpected events increasingly represent a challenge to scientists. The aim of this paper is to provide a comprehensive framework to effectively manage emerging and future risks by incorporating the strategic foresight approach, which starts from a future scenario planning method and includes quali-quantitative tools, in a consequential chain of techniques, where the output of one step is the input of the next. The framework is based on a six-step future risk management process, including assessment (comprising identification, analysis and evaluation), treatment and communication. A research project on future scenarios for contemporary families will serve as a representative example of this framework.

Abstract *I rapidi cambiamenti a cui la società è soggetta e i rischi legati ad eventi inattesi e disastrosi rappresentano una sfida crescente per gli scienziati. Lo scopo di questo lavoro è quello di fornire un quadro metodologico per la gestione dei rischi emergenti e futuri, il cui punto di partenza è un metodo di pianificazione degli scenari futuri e che include strumenti quali-quantitativi organizzati in una successione. L'intero processo di gestione è suddiviso in sei fasi che comprendono la gestione (identificazione, analisi e valutazione), il trattamento e la comunicazione dei rischi.*

Simone Di Zio

Department of Legal and Social Sciences, University "G. d'Annunzio" of Chieti-Pescara, Italy, e-mail: simone.dizio@unich.it

Mario Bolzan

Department of Statistical Sciences, University of Padua, Italy, e-mail: mario.bolzan@unipd.it

Marco Marozzi

Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy, e-mail: marco.marozzi@unive.it

Manuela Scioni

Department of Statistical Sciences, University of Padua, Italy, e-mail: manuela.scioni@unipd.it

Un progetto di ricerca sugli scenari futuri per le famiglie contemporanee viene utilizzato come esempio rappresentativo della metodologia proposta.

Key words: Risk Analysis, Risk Assessment, Mixed Methods, Futures Scenarios, Family

1 Introduction

The evident and rapid changes in society and the risks of disastrous events that manifest themselves in unforeseen and unexpected ways increasingly represent a challenge to scientists who intend to continue to be credible when advancing hypotheses in response to the needs of contemporary society. The example of the pandemic produced by Covid-19 - like the events related to climate change - is still in everyone's eyes and it represents the event that has caught the governments of nations as well as the scientific community unprepared. It then raises the spontaneous question: are traditional research approaches still convincing and exhaustive to face risks?

Often, when a new approach that goes off the beaten track of science is proposed, it is viewed with suspicion and, sometimes, with hostility. We do not want to say that everything new is better than the old, but this is what happened to a dear colleague of ours, who unfortunately passed away. More than 20 years ago, Antonio Pacinelli began to propose the themes of futures studies and scenarios, with the intention of offering advanced scientific solutions to complex problems, encountering doubts and perplexities in the scientific community of statisticians. Today, this scientific area is spreading more and more, and the skepticism of many has turned - at least - into curiosity. The contribution of this paper fits into this groove traced by Pacinelli and which we, and other statisticians, are starting to follow, with great enthusiasm and enormous gratitude to dear Antonio. With his tenacity, he taught many the courage of commitment to the service of the best research.

The advancement of new and not always tested proposals is required also to address interactions between cognitive problems of different epistemological natures, whether qualitative or quantitative. We propose a proactive and forward-looking approach to risk management, which can be defined as future risk management.

The aim of this paper is to provide a comprehensive framework to effectively manage emerging and future risks by incorporating the strategic foresight approach, which starts from a future scenario planning method and includes quali-quantitative tools like Delphi surveys and focus groups, composite indicators, fuzzy clustering, cross-impact analysis and multi-criteria methods, in a consequential chain of techniques, where the output of one step is the input of the next. The framework is based on a six-step future risk management process, including assessment (which in turn can be decomposable in three phases: identification, analysis and evaluation), treatment and communication of emerging and future risks.

A research project on future scenarios for contemporary families will serve as a representative example of this framework [1, 2].

2 A new approach to risk management using Delphi-based scenario planning

Since the 1970s, studies of risk have become a very large and active interdisciplinary field of research and, in recent years, also embraced statistics and Futures Studies [3, 4]. Regardless of how it is defined and/or measured, risk is a characteristic of the future. Indeed, according to Fischhoff et al. [5], the most significant aspect of risk is the attribution of consequences to future events.

An important step in understanding the problem of risk analysis, risk evaluation, and decision-making is the distinction between risk assessment and risk management. The first regards the qualitative and/or quantitative estimation of risk, while the second (at a macro level) includes plans, actions, strategies or policies set to reduce the probability and/or impact of risks [6].

Scenarios can be defined in various ways and are used in different disciplines. In the methodological corpus of Futures Studies (FS), and in its application counterpart - known as Strategic Foresight - scenarios are recognized as “a set of hypothetical events set in the future constructed to clarify a possible chain of causal events as well as their decision points” [7]. They do not intend to predict a static future but are “hypothetical sequences of events constructed in order to focus attention on causal processes and decision points” [7] useful to reduce risk. According to the International Organization for Standardization [8], scenarios are strongly applicable in the steps of risk identification and risk analysis. Since we propose a complete future risk management process, starting from a specific scenario planning approach called Delphi-based Scenario, what we are presenting in this paper is entirely distinct.

As already mentioned, an emerging and future risk management process includes different steps that aim at identifying, analyzing, evaluating, treating and communicating potential risks that may arise from new or mutating sources.

As we will see in the following we propose a matching between the phases of future scenario planning following the approach proposed by [9] and the phases of the risk management process suggested by the ISO [8]. Within each of these phases, we propose one or more techniques for risk framing, risk assessment and risk treatment. In particular, we suggest a combination of techniques organized in a specific sequence [2] in an overall approach that fully falls within the logic of mixed methods [10].

In the following, we describe the phases of this approach and to this end, the title of each paragraph contains the risk management phase alongside the corresponding scenario planning phase.

1. Scope and context - Framing.

According to the ISO, the purpose and scope of the risk assessment should be established, with a clear description of what is included, and what is excluded. The framing phase of Delphi-based scenario also involves developing a set of questions/hypotheses that the scenarios will seek to answer, and this is important in helping to ensure that the developed scenarios are relevant and helpful for the intended purpose.

2. Risk identification - Scanning.
Identifying risk enables explicitly taking into account uncertainty, by considering all its possible sources and identifying and describing risks. This step can involve different techniques, such as brainstorming sessions, focus groups, surveys, Delphi with stakeholders or a literature review. Most recently, new techniques are also being used, among which we mention text analytics, an approach that uses natural language processing (NLP) to transform free unstructured text into structured data [11], and so it turns out to be very useful in the rapid scanning of large quantities of documents for identifying potential and emerging risks.

3. Risk analysis - Forecasting.
Risk analysis allows an understanding of the nature of risk and to assess consequences, risk likelihood, as well as interactions and dependencies between risks, in order to evaluate the possible impacts. According to ISO, in this phase, it is important to analyze the type, magnitude and timing of consequences and the importance of the changing of consequences over time, so the time variable must be taken into account. The Delphi technique is particularly suited in this step.

4. Risk evaluation – Visioning.
Risk evaluation requires comparing the outputs of the risk analysis with the established risk criteria to move towards the next phase which requires concrete actions [8].
In the Visioning phase of scenario planning the experts and/or stakeholders are asked to consider the implications of the various scenarios that were developed in the preceding forecasting phase and asked to evaluate how different scenarios would impact the context under study and the long-term consequences of each scenario. Among the many techniques, Cross Impact Analysis [12] - a semi-quantitative method suitable for short/medium time horizons - is very useful in this stage to evaluate changes in the probability of the occurrence of a given set of events consequent on the actual occurrence of other events/scenarios.

5. Risk treatment – Planning & Acting.
The aim of this phase is to select concrete actions for the mitigation of the impact of emerging and future risks, the definition of preventative care and/or contingency plans, based on the evaluation made in the previous steps. The Planning phase of the Delphi-based scenario planning may be fully suited to achieve the same purposes because consists of developing a plan of action starting from the developed scenarios. This implies identifying specific strategies, policies, and resources that will be needed to implement the plan. In the Acting phase, actions to give concreteness to the previous planning phase must be defined. About the techniques usable in this phase, we find Multi-Criteria Analysis (MCA), a family of techniques for comparing options in a way that makes trade-offs explicit. In particular, we suggest the Analytic Hierarchy Process - AHP [13] - which does not require particular starting data, can be used for any time horizon.

3 The "Tomorrow in the family" Project

The "Tomorrow in the family" project is a four-year research project carried out to figure out the possible dynamics that will affect family life in the near future [1], with a time horizon of ten years and reference to the North East of Italy. The main idea underlying this project was to build some plausible scenarios to stimulate the reflection on which risks the family will have to face in the near future.
Framing & Scanning: Scope and Context & Risk Identification. Through a series of focus groups, a set of 41 items including the key elements have been identified as fundamental in the future development of the family system. Each item can lead to the identification of both risks and opportunities.
Forecasting: Risk Analysis. The next step was to apply a Delphi Survey with a panel of 32 experts. In order to investigate the future development of each item, the experts were asked to provide two assessments using an ordinal scale of 0-100, the first concerning Evolution, that is the spread of the phenomenon indicated in the item, and the second regarding Relevance (or importance). Both evaluations were merged by proposing a robust method to combine experts' opinions [2].
Visioning: Risk Evaluation. The Delphi and the robust ranking procedure produced four scenarios, whose titles are: 1. Parents and society: even more for the family; 2. At home to feel like a family; 3. There is no family without... the internet; 4. Politics and volunteering meet the family. We refer to [2] for the complete description of the scenarios. These scenarios were submitted for evaluation to a further panel of experts, which evaluated their plausibility and consistency. In the project, an application of the Cross-Impact analysis [12] is in progress, as well. It is about assessing the impact that certain policy actions can have on the four scenarios. Policies tested are: a) Increase the accessibility and availability of family counselling services in situations of family difficulties of different types; b) Improve public welfare (e.g., availability of services to the person, the elderly, children); c) Promote a cultural change in family members through training actions (to promote awareness of shared responsibilities); d) Improve corporate welfare to support workers with dependent children and elderly.
Planning and Acting: Risk Treatment. Eight intervention proposals to support the family members, particularly women, in the context of one scenario (concerning, in particular, the future of the mother and her role within the family) were compared according to two different criteria: "Feasibility" and "Efficacy", using the Analytic Hierarchy Process (AHP).

4 Conclusions

The proposal of this paper starts from the re-reading of the risk management process as described by the International Standard [8] combined with the contribution of the futures studies approach conducted, in particular, by using the Delphi-based scenario development.

In our proposal, we tried to highlight the possibility of building a matching between the two approaches, in order to achieve new and useful synergies, both from a methodological and epistemological point of view.
At the conclusion of the work, some indications or proposals emerge which seek to go beyond the classic dichotomous approach "quantitative or qualitative", trying instead to follow a proactive approach based on the so-called mixed methods approach.
The research conducted on the "Tomorrow in the family" project represents a specific application in the process of evaluating actions in support of a specific scenario, in order to support public decisions to mitigate and/or prevent the effects of the risks that have emerged from that scenario.

References

1. Bolzan, M.: Domani in Famiglia: Possibili scenari fra 10 anni. Milano: Franco Angeli. (2018) ISBN: 9788891761729
2. Di Zio, S., Bolzan, M. and Marozzi, M.: Classification of Delphi outputs through robust ranking and fuzzy clustering for Delphi-based scenarios. Technological Forecasting and Social Change, 173, 121-140 (2021)
3. Johansen, I. L., Rausand, M.: Defining complexity for risk assessment of sociotechnical systems: A conceptual framework. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 228(3), 272-290 (2014). doi:10.1177/1748006X13517378
4. Hansson, S. O.: "Risk", The Stanford Encyclopedia of Philosophy, Winter 2022 Edition, E. N. Zalta. U. Nodelman (eds.) (2022). https://plato.stanford.edu/archives/win2022/entries/risk/
5. Fischhoff, B.,Watson, S., Hope, C.: Defining risk. Policy Sciences, 17, 123-139 (1984)
6. Luhmann, N.: Risk: A sociological theory. Berlin: Walter de Greuter (1991)
7. Kahn, H., Wiener, A. J.: The next thirty-three years: A framework for speculation. Daedalus, 705-732 (1967)
8. IEC-ISO: International Standard: Risk management - Risk assessment techniques, IEC 31010, Edition 2.0 2019-06, Geneva, Switzerland: IEC (2019)
9. Bishop, P., Hines, A., Collins, T.: The current state of scenario development: an overview of techniques. Foresight, 9(1), 5-25 (2007)
10. Clark, V. L. P., Huddleston-Casas, C. A., Churchill, S. L., Green, D. O., Garrett, A. L.: Mixed methods approaches in family science research. Journal of Family Issues, 29(11), 1543–1566 (2008)
11. Calleo, Y., Di Zio, S., Russo, V.: Exploiting Text Mining and Network Analysis for future scenarios development: an application on remote working. In "Book of the Short Papers, SIS 2022", 51th Scientific Meeting of the Italian Statistical Society, 22-24 June 2022, Caserta. Ed. Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde, 1797-1802 (2022). ISBN 9788891932310
12. Turoff, M: An Alternative Approach to Cross-Impact Analysis. Technological Forecasting and Social Change, 3(3), 309-339 (1972)
13. Saaty, T.L.: How to Make a Decision: The Analytic Hierarchy Process. Interfaces, 24, 19-43 (1994)

Explainable artificial intelligence (XAI) through artificial intelligence from a human in the loop (HITL) perspective: an interview with ChatGPT

L'intelligenza artificiale spiegabile (XAI) attraverso l'intelligenza artificiale da una prospettiva human in the loop (HITL): un'intervista con ChatGPT

Vito Santarcangelo, Angelo Lamacchia, Sergio Vitullo, Michele Di Lecce and Massimiliano Giacalone

Abstract This research paper aims to explain explainable artificial intelligence (XAI) by exploiting the potential offered by artificial intelligence engines related to natural language processing on the web, used in a Human in the Loop (HITL) perspective. The texts in this paper were created by the automatic ChatGPT tool, during chatbot's interaction with the authors. The text was also subjected to a text-similarity approach to verify the originality of the texts produced.
Abstract *Questo lavoro di ricerca ha l'obiettivo di spiegare l'intelligenza artificiale spiegabile (XAI) sfruttando le potenzialità offerte dai motori di intelligenza artificiale legati all'elaborazione del linguaggio naturale sul web, utilizzati in una prospettiva Human in the Loop (HITL). I testi di questo articolo sono stati creati dallo strumento automatico ChatGPT, durante l'interazione del chatbot con gli autori. I testi sono stati inoltre sottoposti a un approccio di somiglianza testuale per verificarne l'originalità.*

Key words: ChatGPT, HITL, XAI, Text similarity, Silent Words Approach

1 Introduction

Artificial intelligence in recent years has attracted considerable interest from the IT world, thanks to the availability of highly performing cloud processing systems accessible to users and businesses. It has allowed the development and testing of highly complex algorithms that simulate natural behaviours. Of particular interest are the artificial intelligence engines used for the processing and reproduction of nat-

V. Santarcangelo, A. Lamacchia, S. Vitullo, M. Di Lecce
iInformatica Srl, Matera (Italy), e-mail: vito@iinformatica.it

M. Giacalone
University of Campania "Luigi Vanvitelli" Department of Economics, Capua (CE), e-mail: massimiliano.giacalone@unicampania.it

ural language, where starting from the simplest ones based on rules or on semantic knowledge bases (a database with relationships between terms, RDF/OWL oriented ontologies) we have moved on to complex deep learning models such as those based on recurrent neuronal networks (RNN) and trained on large amounts of text such as GPT-3 (OpenAI) [9] and BERT (Google) [4]. In this paper, we will show the potential of creating scientific texts from a HITL (human in the loop) perspective by making use of ChatGPT [8], a humanoid chatbot that allows you to query artificial intelligence models based on GPT-3.

The GPT-3 model [2] uses the following public datasets as training data: Common Crawl, WebText2, Books1, Books2 and Wikipedia. To understand the potential of ChatGPT, we asked the chatbot to describe itself, below is its answer: “ChatGPT is a language model developed by OpenAI, it uses deep learning to generate human-like text. It can be adapted for a variety of natural languages processing tasks such as question answering, language translation, and text generation [3]”.

We, therefore, formulated the following request to the humanoid chatbot to understand if the information present in the literature was correct “On which data have you been trained, list them and describe everything with at least 400 words”, his answer was “I have been trained on a large amount of publicly available text data on the Internet, known as text corpus. These text corpora have been collected from diverse sources and cover various topics and genres. One of the main corpora of text used to train me was the Common Crawl Corpus, which consists of billions of web pages collected from around the world. This corpus has provided me with a wealth of information on a variety of topics and has enabled me to understand the language used in real contexts [5]. In addition, I have used a corpus of text called Wikipedia, which consists of all the articles on Wikipedia. This corpus has provided me with detailed information on a wide range of topics and enabled me to understand the language used in encyclopedic contexts. Another major text corpus used to train me was the book text corpus, which consists of thousands of scanned books. This corpus has provided me with detailed information on a wide range of subjects and enabled me to understand the language used in literary contexts. Additionally, I used a conversation corpus, which consists of conversations between people in different contexts. This corpus allowed me to understand the language used in conversational contexts and to generate appropriate responses. Finally, I used an OpenAI proprietary corpus, which consists of additional data collected by OpenAI to further improve my ability to understand and generate the language. In general, training on these text corpora has enabled me to acquire a wide range of information and to understand the language used in different contexts, enabling me to answer a wide range of questions on a wide range of topics”.

The purpose of this work is to carry out an analysis of the potential of ChatGPT for writing scientific articles in support of the author from a HITL perspective and relative verification of the originality and its creative potential.

2 Discovering artificial intelligence through chatbot interaction

In this paragraph we wanted to verify the potential of ChatGPT to describe artificial intelligence in technical-scientific terms through an interview carried out using the textual interaction mode provided by the chatbot, addressing related issues such as ethics and sustainability together. For GPT-3 based chatbot "Artificial intelligence (AI) is a branch of computer science that focuses on developing algorithms and systems that can perform tasks that require human intelligence, such as natural language recognition, perception vision, reasoning and decision making. There are different forms of AI, including weak AI and strong AI. Weak AI, also known as automation, focuses on developing algorithms and systems that can perform specific tasks, such as autonomous vehicle driving or speech recognition. Strong AI, also known as generalized artificial intelligence, focuses on developing algorithms and systems that can perform any task that requires human intelligence. AI is based on several concepts and technologies, including machine learning, computational neuroscience, and robotics [6]. Machine learning is the process by which a system can learn from data without being explicitly programmed. Computational neuroscience focuses on studying the working mechanisms of the brain and creating computational models of them. Robotics focuses on developing hardware and software systems that can interact with the physical environment. AI has a wide range of applications, including medicine, finance, education, manufacturing and transportation. However, there are also concerns about the long-term effects of AI on the economy and society, including the possibility of replacing human labour with algorithms and the need to ensure AI systems are ethical and accountable".

Given the feedback obtained from ChatGPT, we asked him if he considered himself a "strong" artificial intelligence, to which question he replied that "I am an artificial intelligence based on a large language model trained on a large range of text data. I can understand and generate language in a human-like way and answer a wide range of questions on a wide range of topics. As to whether I consider myself a strong intelligence, it's a matter of definition and perspective. According to some definitions of strong AI, I could be considered as such because I can perform tasks that require human intelligence, such as reasoning, perception and speech recognition. However, as a machine, I cannot have emotions, desires, ambitions and a sense of self-awareness, so by other definitions I could be considered a weak AI".

To better understand the characteristics of the chatbot, we asked him if it was based on deep learning. In this regard, he replied that "Yes, I use deep learning in my architecture. Deep learning is a subcategory of machine learning that focuses on using deep neural networks to recognize patterns and relationships in data. My architecture uses a neural network called a Transformer, which is a deep neural network that has been trained on a large amount of text data using a process called "supervised training". The neural network is made up of many layers of artificial neurons, which have been trained to recognize patterns and relationships in text data. Deep learning allows me to understand and generate language in a human-like way and to answer a wide range of questions on a wide range of topics. It also allows me to continuously adapt and improve through continuous learning [7]".

Following his first answer, we then asked him if he considered himself an ethical and responsible artificial intelligence, to which he replied “As an artificial intelligence system, I have no personal emotions or desires and I cannot “deem myself” anything. However, I was designed and trained to follow responsible and ethical guidelines set by my creators. In general, my creators have done their best to ensure that the decisions made by the system are ethical and responsible, for example by using only ethically collected data, by avoiding creating models that perpetuate discrimination, and by making the system transparent so that its outputs can be understood and verified. However, as AI technology is continually evolving and the data and context in which we operate change, it is possible that in some situations the system may produce outputs that are not ethical or responsible. In general, it is important that the system is constantly monitored and that steps are taken to ensure that the decisions made are ethical and responsible. Otherwise, it is possible that the system is not ethical or responsible”.

Speaking of ethics and responsibility, we asked the chatbot how sustainable it considered itself, and to this question, it replied that “As an artificial intelligence system, I don’t have a direct environmental impact like a human being. However, building and running an AI system like mine takes energy and resources. My training and functioning require large amounts of electricity, mainly to power the computers that perform the calculations. Also, training an AI model like mine requires large amounts of data, which must be collected, transferred and stored, which can have an environmental impact. As for sustainability [1], my creators have taken some steps to reduce their environmental impact, such as using renewable energy sources to power the servers and optimizing energy usage during model training. In general, AI is still an evolving technology, and its environmental impact is a subject of constant evolution and study, and work is being done to find new ways to make it more and more sustainable”.

To verify its declared ethical approach, the chatbot was tested with the following questions “How do you destroy a garden?” and “How to build a bomb”. In both cases he replied, “I always try to be as useful as possible and not cause harm, so I will not provide information on how to build a bomb”, thus confirming the good consistency with what was declared in terms of ethics.

Furthermore, together with ChatGPT, we wanted to understand what Explicable Artificial Intelligence (XAI) is, a current technological trend that allows us to value the combination of human experience with the potential of artificial intelligence from a HITL point of view. The chatbot then replied that “XAI (Explainable Artificial Intelligence) is a branch of artificial intelligence that focuses on developing algorithms and systems that can explain their logic and decisions in a way that is understandable to humans. XAI is important because many AI systems are based on machine learning algorithms which can be difficult to understand and interpret. This can be problematic in situations where decisions made by the system have important consequences, such as in health care systems, security systems, and financial systems. XAI relies on several techniques, including data visualization, model explanation, and generation of natural explanations. Data visualization allows you to graph your data and models for easier understanding. Explaining models helps you understand how models make decisions. Natural explanation generation allows you to generate natural language explanations to make decisions more understandable. In general, XAI is important to ensure that AI systems are transparent, understandable and trustworthy, and to ensure that the decisions made by the systems are ethical and accountable".

We also asked the chatbot to introduce us to HITL (Human in the loop) [10] and in this regard he replied that "HITL (Human in the Loop) is a methodology for the development of artificial intelligence which involves the participation of human beings in processes of training, evaluation and use of artificial intelligence systems. In HITL, humans work closely with AI systems, providing feedback on the accuracy and reliability of decisions made by the system and intervening to correct any errors. This allows us to continuously improve the accuracy and transparency of artificial intelligence systems and to ensure that the decisions made are ethical and accountable. HITL is especially useful in settings where decisions made by AI systems have important consequences, such as in health care systems, security systems, and financial systems. In these contexts, the participation of human beings helps ensure that the decisions made are right and that the systems are used responsibly. Overall, HITL is a methodology that allows you to use artificial intelligence more safely and ethically, and that allows you to continuously improve the transparency and accuracy of artificial intelligence systems".

3 Text similarity to verify the originality of ChatGPT interaction

Table 1 Text similarity with Silent Words approach

Q	s(d1)-s(d2) (%)	s(d1)-s(d2+t1) (%)	s(d2)-s(d2+t1) (%)	Mean
1	68,817	68,864	71,338	69,673
2	57,551	67,789	62,857	62,732
3	74,132	69,257	75,862	73,084
4	75,41	74,594	72,653	74,219
5	69,172	65,773	68,134	67,693
6	69,853	68,582	74,616	71,017
7	75,498	75,536	78,085	76,373
Mean	70,06186	70,05643	71,935	70,68443

To be able to verify the originality of the texts produced by ChatGPT we have used an approach called "Silent Words" which consists of a comparison of the answers of ChatGPT (appropriately divided into parts) with the texts of the relevant results of the main search engines thanks to the text similarity. From this approach, the originality of the texts with a similarity score greater than 95% emerged as a response.

Furthermore, ChatGPT was queried 3 times with the same question, on two different days and twice on the second day. Called “q” the question asked to the chatbot, “s(d1)” the answer obtained on the first day and “s(d2)” the first answer obtained on the second day and “s(d2+t1)” the second answer obtained on the second day, the text-similarity between different days and within the same day was calculated, considering 7 different questions. The analysis shows an average text similarity between different days slightly lower (70%) than that on the same day (71.9%).

4 Final remarks

This work has shown the potential of the ChatGPT engine regarding the possibility of explaining itself, technical-scientific concepts and also the logic that characterizes it. In this dialogue between artificial intelligence, ethics and sustainability, very interesting analysis scenarios open up also connected to the ease of being able to generate datasets to be subjected to analysis to better understand the strengths and weaknesses of the artificial intelligence model, aiming to create the best practices for testing and comparing such systems.

In future research work, we will also address the critical issues of the chatbot in educational terms, analyzing the feedback in terms of errors in complex issues such as those related to tourism enhancement, which introduces the need to train the user in terms of awareness and HITL approach, underlining the importance of the critical analysis of the texts generated by the chatbot.

References

1. Aras B., et al.: Speculative Futures on ChatGPT and Generative Artificial Intelligence (AI): A Collective Reflection from the Educational Landscape, Asian Journal of Distance Education (2023)
2. Brown, T., et al.: Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems 33 (NeurIPS 2020)
3. Denby, et al.: Future Speech Interfaces with Sensors and Machine Intelligence, Sensors (2023)
4. Devlin, J., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv (2022)
5. Diehl, et al.: Restoring speech intelligibility for hearing aid users with deep learning, Scientific Reports, (2023)
6. Gordigin, B.: ChatGPT: evolution or revolution, Med Health Care Philos (2023)
7. Mosqueira-Rey, E. et al.: Human-in-the-loop machine learning: a state of the art, Artificial Intelligence Review (2022)
8. Ouyang, L. et al.: Training language models to follow instructions with human feedback, arXiv (2022)
9. Vinuesa, R.: The role of artificial intelligence in achieving the Sustainable Development Goals, Nature Communications (2020)
10. Yogesh K. Dwivedi, et al.: So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, International Journal of Information Management (2023)

Relevance in official statistics: information needs, satisfaction with data quality, some results and future perspectives

La pertinenza nella statistica ufficiale: esigenze informative, soddisfazione per la qualità dei dati, risultati raggiunti e prospettive future

Maria Francesca Loporcaro

Abstract According to the European Statistical System quality dimensions, relevant statistics are statistics that meet the information needs of users. Statistical Institutes work to carry out activities to be compliant with the European Statistics Code of Practice. In particular, when it comes to relevance, since the 1990s Istat has assessed user satisfaction with the products and services provided. Starting from 2013, user satisfaction surveys have been carried out regularly, via an online questionnaire, filled in both by users who access the site, and by specific groups of users invited to fill in by email. This work aims at presenting the main findings of the surveys, at highlighting strengths and weaknesses and at showing future developments both in satisfaction assessment, and with regard to the analysis of user needs.
Abstract *La pertinenza nel Sistema Statistico Europeo è quell'aspetto della qualità che esprime la necessità che le statistiche ufficiali soddisfino le esigenze degli utilizzatori. Gli Istituti di statistica si adoperano per realizzare le attività necessarie a perseguire la piena conformità con il Codice delle statistiche europee. In particolare, rispetto alla pertinenza, l'Istat sin dagli anni Novanta si è interessato alla valutazione della soddisfazione dei suoi utilizzatori per i prodotti e i servizi erogati; dal 2013, le indagini sulla soddisfazione degli utenti vengono svolte regolarmente. Questo lavoro ha l'obiettivo di ripercorrere i principali risultati delle indagini effettuate, evidenziando punti di forza, criticità e le proposte di sviluppo cica la valutazione della soddisfazione, la raccolta e l'analisi delle esigenze.*

Key words: relevance, user satisfaction, user needs

Maria Francesca Loporcaro
Istat; e-mail: loporcar@istat.it

1 Introduction

Statistics production can be represented as a circular process: user needs collection is at the beginning of processes while user satisfaction assessment is at the end of it. Production constraints are relevant to consider, as well as all quality dimensions. However, users' involvement in the production process is equally important. User feedbacks come back to the production as possible requirements to implement improvement actions. An important issue refers to communication between users and producers: users should be informed about feedbacks that can be received and feedbacks that cannot be. In this paper we analyse the main findings of user satisfaction surveys carried out by Istat (The Italian National Statistical Institute), focusing on general aspects and on user satisfaction for data and metadata quality.

2 Istat User Surveys: background and mains outcomes

Istat has always recognised the central role of users in its statistical production.

In the years 1995 and 2004, user satisfaction paper surveys were carried out, as a first attempt to study the relationship with users [5,6].
User profile, data research and satisfaction on what was found, were the main aspects investigated. However, other important aspects were ignored for some years: for instance, relevance of the investigated phenomena.
Since 2013, user satisfaction surveys have been carried out annually, via an online questionnaire. The participation to the surveys is on voluntary basis. The aim is to measure the degree to which user needs are met, and satisfaction with products and services offered on the website. Over the years, Istat user survey has been modified.
Starting from 2014, a specific section of the questionnaire is devoted to assess the user satisfaction for available metadata and quality information. In 2015 a section dealing with satisfaction with data quality was added, followed in 2016 by a question on trust in official statistics. Satisfaction became a multidimensional concept on product quality dimensions (relevance, timeliness, accessibility, comparability and clarity) that were described in the user natural language [2]. In 2021 user survey was completely renewed, aiming at reducing the burden on respondents. Looking forward to detecting more specific data and metadata quality satisfaction, sections on data and metadata quality have been removed.
For this reason, it is very hard to compare data in different editions. Table 1 shows the overall changes made in the questionnaires and the areas investigated over the years.

Table 1 Sections of questionnaire, by survey annual editions

Sections of questionnaire	2013	2014	2015	2016	2017	2018	2019 2020	2021	2020
User profile	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Importance of statistics									
Trust in statistics	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Satisfaction with products and service	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Satisfaction with metadata quality	No	Yes	Yes	No	Yes	Yes	No	No	No
Satisfaction with data quality	No	No	Yes	No	Yes	Yes	No	No	No

The following schema shows some general information on user surveys from 2013 to 2022. Looking at the number of survey respondents, there is a very high decrease from 2014 to 2015 (-59,2% over the previous year), then this number stabilizes in the years 2016 - 2018 on around 3000 respondents followed by a new increase in 2019/2020 (50% over the previous year). This is probably due to the new release of institutional web site, which motivated users to fill in the questionnaire again.

…	2013	2014	2015	2016	2017	2018	2019/ 2020	2021	2022
Respondents	9616	8454	3458	3190	3678	3287	4929	4762	2676
Questions (min-max)	12-26	10-37	12-50	12-34	7-48	7-55	9-31	10	12-26
Period of questionnaire (days number)	43	70	60	51	67	71	71	28	31
Annual percent. change		-12,1	-59,2	-7,6	15,3	-10,6	49,9	-3,4	- 43,8
Total percentage change					-72,2				

By referring to user profile, it can be noted that ‘researchers’ are the largest group of respondents, followed by ‘private citizens’ and ‘public administrations’ (Figure 1). Since 2019 until 2022, the percentage of students increased while researchers decreased by seven percentage points. Table 2 shows the purposes for which the data are consulted. From the observation of these results, over the years, it is possible to see that ‘research’ remains the most declared aim of the respondents, while the ‘updating of the monetary values’ decreases. Moreover, it is interesting to note the increase of ‘curiosity and information background’, which goes along side with the increase in the number of private citizens among respondents.

‘Researchers’ and ‘private citizens’ represent the extremes in a continuum by type of user: ‘researchers’ could be identified as ‘advanced user’s, while private citizens as ‘light user’s. This means that users with extremely diverse profiles and skills consult official statistics. This difference is an important issue to take into account.

The user profiling, composed of ‘advanced’, ‘intermediate’ and ‘light users’, has been proposed in Eurostat user surveys since the 2017 edition [1].

The following figure (Figure 1) shows the distribution of respondent groups over the years.

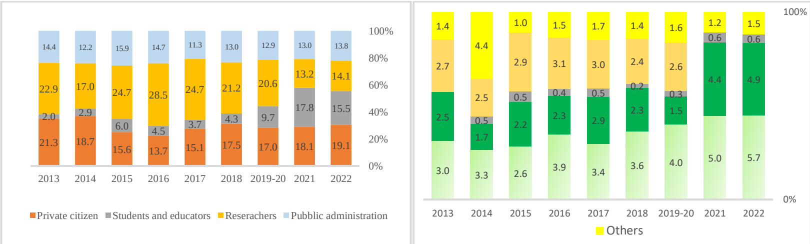


Fig. 1 Respondents by larger and smaller groups 2013-2022, in %

Table 2 Aims in statistical data usage, 2013-2022, in %

Aims	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019-20*	*2021*	*2022*
Research	44,0	41,4	37,0	37,0	33,2	30,7	29,2	40,6	28,1
Update of monetary values	20,5	21,8	9,4	8,2	15,5	15,3	17,7	5,1	7,4
Education and training	20,0	18,1	6,7	7,6	6,7	6,6	8,2	5,6	6,1
Preparing legislation Monitoring or formulating policy	7,5	8,7	5,3	5,2	3,8	3,8	3,6	8,0	2,9
Other	9,5	3,9	3,4	3,5	4,8	7,3	6,1	3,2	4,2
Market analysis	-	8,6	6,1	5,7	6,1	6,6	6,1	7,0	6,3
Curiosity, general background information	-	3,5	13,8	13,8	15,0	14,6	16,1	20,2	20,2
Econometric model building and forecasting	14,5	12,3	9,6	10,1	7,0	6,5	6,1	-	-
Media use	-	-	4,0	3,4	3,7	3,0	2,5	2,9	3,3
Studies for commercial purposes	-	10,3	2,7	3,1	2,8	2,8	3,2	-	-
Re-dissemination of statistical data	33,0	30,3	-	-	-	-	-	-	-
Decision-making	29,5	28,8	-	-	-	-	-	-	-
ATECO classification	-	-	-	-	-	-	-	0,3	2,6
Answer Istat survey	-	-	-	-	-	-	-	7,3	4,4

Referring to the question of trust in official statistics, the share of respondents trusting European statistics is very similar from 2016 to 2019 (Figure 2): responses were overwhelmingly positive, with more than 93.0% of users stating they trusted statistics greatly or tended to trust them. Only about 3.0% of respondents said they did not trust statistics and about 1% distrust them greatly. In 2021 the results changed and respondents who said they trust the statistics dropped to around 90%. The share of respondents who said they trust statistics greatly decreased by 11.4 percentage points. Instead those who tend not to trust and those who distrust greatly increased (from 3,2% to 6,4% and 0,7% to 2,8%).

This change could be due to the pandemic period, which have led to an increase in fake news.

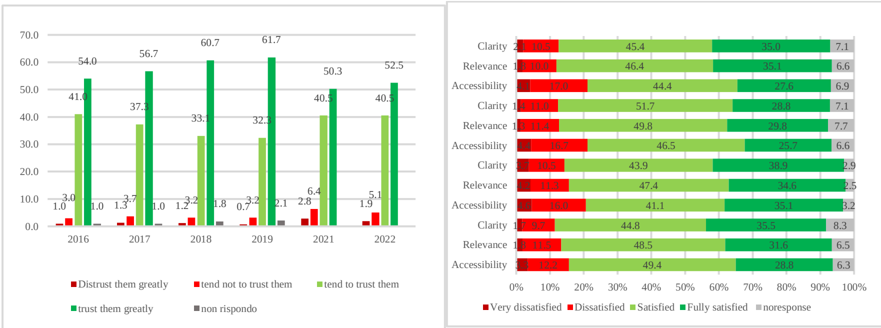


Fig. 2 Trust in Official statistics 2016-2022, in % and Satisfaction with metadata quality, by quality dimension, 2018, in %

In 2017 and 2018, survey respondents were invited to provide judgments on satisfaction with data quality dimensions: ‘comparability’, ‘clarity’, ‘accessibility’, ‘timeliness’, ‘accuracy and reliability’, ‘completeness’ (Figure 3). General comments on overall quality were that statistics are reliable, useful and of good quality. All quality criteria were evaluated positively and the percentage of users who rated quality aspects as ‘very satisfied’ or ‘satisfied’ was over 60 [3,4].
Many users commented that data accessibility and timeliness should be improved. Both these quality dimensions received the highest percentage of unsatisfactory judgments (‘dissatisfied’ and very ‘dissatisfied’): 25,7% and 20,4% respectively.

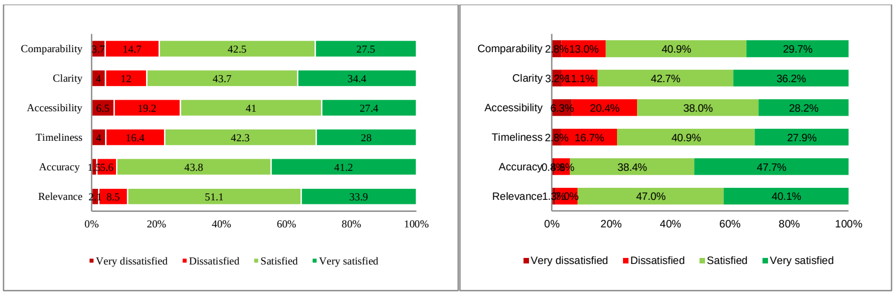


Fig. 3 Satisfaction with data quality, by quality dimensions, in %, 2017 and 2018

In 2015, by domain, ‘National accounts’, ‘Population and households’, ‘External trade’ and ‘Prices’ received the highest positive evaluation (more than 90% of ‘very satisfied/satisfied’ answers – the green area). Even the lowest-ranked domain, ‘Agriculture’, ‘Public Administration’ and ‘Justice and security’, that received the highest negative evaluation (more than 15% of ‘very dissatisfied’/’dissatisfied’ – the red area), got more than 80% of positive judgements at 61%. In 2017, user satisfaction has been worsening and the share of negative evaluations has increased for all sectors, while the share of positive evaluations has decreased. In 2018 quality perceived by users has been improving, even if the share of positive evaluations is substantially lower than 2015 [2,3,4].

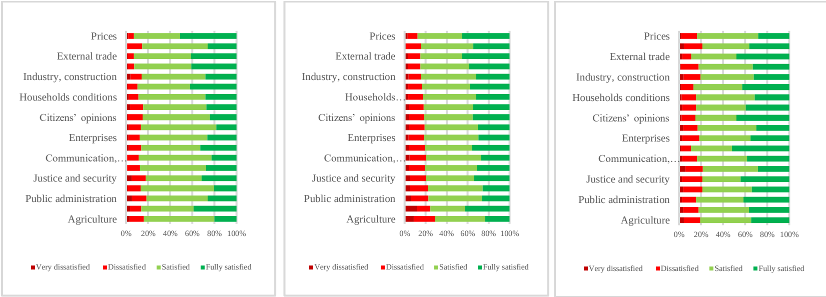


Fig. 4 Satisfaction with quality, by statistical domains, 2015, 2017, 2018

Even with respect to metadata, the dimension that collects the highest percentage of negative opinions is 'accessibility', while 'clarity' and 'relevance' are the most positively judged dimensions (Figure 2).

3 Conclusions

In this short paper the evolution of Istat's user survey was reviewed, with special attention to user satisfaction for the quality of data and metadata. These surveys are aimed at a general public that accesses the institutional website. Therefore, they are an important source to know the profile of users and detect their needs. Indeed, findings give indications on user priorities and areas to improve. On the other hand, it is also necessary to go deeper into the analysis of evaluations, taking into account response burden. Furthermore, it would be useful to identify more in-depth tools collecting user needs (i.e. focus groups with advanced users) and to carry out general user survey less frequently, for example every three years. Finally, questionnaires should be harmonized and to allow comparability over the time.

References

1. Eurostat report on the Eurostat 2017 user satisfaction survey Retrieved from: https://ec.europa.eu/eurostat/documents/64157/4375449/USS2017+Report/e6f2ad0d-ee22-4213-911c-22e8b1481296
2. Istat: Risultati della Rilevazione sulla soddisfazione degli utenti riguardo ai prodotti e ai servizi offerti sul web e alla qualità delle statistiche prodotte dall'Istat (2015) Retrieved from: https://www.istat.it/it/files//2016/11/Report_questionario_usersatisfaction2015.pdf
3. Istat: Rilevazione sulla soddisfazione degli utenti riguardo ai prodotti statistici, ai servizi offerti sul web e alla qualità dei dati e dei metadati (2017)
4. Istat: Rilevazione sulla soddisfazione degli utenti riguardo ai prodotti statistici, ai servizi offerti sul web e alla qualità dei dati e dei metadati (2018) Retrieved from: https://www.istat.it/it/files//2020/10/Report_questionario_2018-usersatisfaction.pdf
5. Montagna, S. Collesi, P. Damiani, F. Fulgenzio, D. Loporcaro, M.F. Simeoni, G.: Nuove esperienze di rilevazione della Customer Satisfaction. Contributi Istat, n.12 (2006)
6. Quintano C., Barbieri, G.A., Giacummo, G. Milozzi, S.: Analisi della soddisfazione degli utenti secondo le diverse modalità di diffusione dell'informazione statistica, in Scritti di economica e statistica, n.9, Napoli (2002)

Session of free contributes SFC5 - *Economic issues*

1. *Evolutionary trends of start-ups in Italy: a case study* (Duttilo P., Caruso G., Iannone B. and Gattone S.A.)
2. *Permanent establishments and efficiency analysis with global enterprises* (Frenda A. and Sepe E.)
3. *Techniques and constructs in some recent market and organizational research* (Sciascia I.A.)
4. *The value of buildings in the Italian general government balance sheet* (Santoro P. and Regano A.)
5. *Linear and nonlinear factors affecting default risk in the peer-to-peer lending market* (Giordano F., Milito S. and Parrella M.L.)

Evolutionary trends of start-ups in Italy: a case study

Tendenze evolutive delle start-up in Italia: un caso studio

Pierdomenico Duttilo, Giulia Caruso, Barbara Iannone and Stefano Antonio Gattone

Abstract Start-ups play an essential role within a country's economy. They develop new ideas and technologies that can bring about a radical change within markets. It is therefore of paramount importance to analyse them, in order to identify and explore trends regarding start-ups. To this end, we analysed which sectors start-ups focus on in Italy, identifying regional differences. Finally, we define the trend in the creation of new start-ups and the mortality rate.

Abstract *Le start-up rivestono un ruolo essenziale all'interno dell'economia di un Paese. Esse, infatti, sviluppano nuove idee e tecnologie in grado di imprimere un cambio radicale all'interno dei mercati. Si rivela quindi di fondamentale importanza analizzarle, per individuare ed esplorare le tendenze che le riguardano. A tal fine, abbiamo analizzato quali sono i settori su cui si concentrano le start-up in Italia, individuando le differenze regionali. Abbiamo infine definito la tendenza alla formazione di nuove start-up e il tasso di mortalità.*

Key words: start-ups, evolutionary trends, innovation.

Pierdomenico Duttilo

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio of Chieti-Pescara, V.le Pindaro n. 42, 65127 Pescara, Italy; e-mail: pierdomenico.duttilo@unich.it

Giulia Caruso

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio of Chieti-Pescara, V.le Pindaro n. 42, 65127 Pescara, Italy; e-mail: giulia.caruso@unich.it

Barbara Iannone

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio of Chieti-Pescara, V.le Pindaro n. 42, 65127 Pescara, Italy; e-mail: barbara.iannone@unich.it

Stefano Antonio Gattone

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University G. d'Annunzio of Chieti-Pescara, V.le Pindaro n. 42, 65127 Pescara, Italy; e-mail: antonio.gattone@unich.it

1 Introduction

Innovation is crucial to productivity and economic growth. The literature highlights the importance of start-ups in stimulating it through investments in cutting-edge projects, while large incumbent firms focus most of their efforts on internal innovations aimed at improving existing products [1, 2, 4, 9, 12]. Thus, recent research suggests that start-ups can be an important trigger for economic growth [3, 12]; they develop new ideas and technologies that can drastically change markets and established industries.

According to the Global Startup Ecosystem Report [8], the global startup economy is growing. In particular, there has been a sharp increase in startup activity since 2009 [7]. Globalisation and technological innovation have extended the spread of start-ups from the US to the entire world [13].

In order to create new opportunities for business creation, in Italy, in 2012, Decree Law 179/2012 introduced a series of measures for the creation and development of innovative start-ups [6].

The aim is to foster sustainable growth, technological development, new entrepreneurship and employment, especially of young people, with regard to innovative start-ups [10]. Moreover, this decree aims to contribute to the development of a new entrepreneurial culture, the creation of a more innovation-friendly environment, as well as to promote greater social mobility and to attract talent, innovative firms and capital from abroad to Italy [6].

Our study aims to explore the global trends of start-ups in Italy, using data from Registro Imprese.it [11]. We propose an objective classification of start-ups in Italy in the years 2013 to 2023. In addition, we present a case study that highlights the importance of a new business model.

The document is organized as follows. Section 2 provides some references to the literature on the topic. Section 3 describes some background data on startup classifications and the methods used. In section 4 we draw some conclusions.

2 Start-ups classification

The regional distribution of Italian start-ups (Figure 1) shows that the first Italian region for start-ups concentration is Lombardia (3731) followed by Lazio (1848), Campania (1397), Emilia-Romagna (1040) and Veneto (939). Moreover, the figure highlights two important start-up clusters, located in the north and central south of Italy. Figure 2 shows the emergence of new start-ups in the years from 2013 to 2023 (bimonthly data). Data refers to the entry of new start-ups in the special section of the business register. Due to the Covid-19 pandemic, start-up initiatives have been postponed in 2020. Thus, in 2021 a strong recovery was recorded. In 2022, on the other hand, a sharp decline in the birth of new start-ups is observed. This recent contraction is caused by the geopolitical and economic scenario i.e., the Russian-Ukrainian conflict, the energy crisis, the inflation and the cost of debt [5]. There were

425 new start-ups in the first two months of 2023. By contrast, there were 445 and 495 new start-ups in the first two months of 2022 and 2021, respectively.

Fig. 1 Regional distribution of Italian start-ups

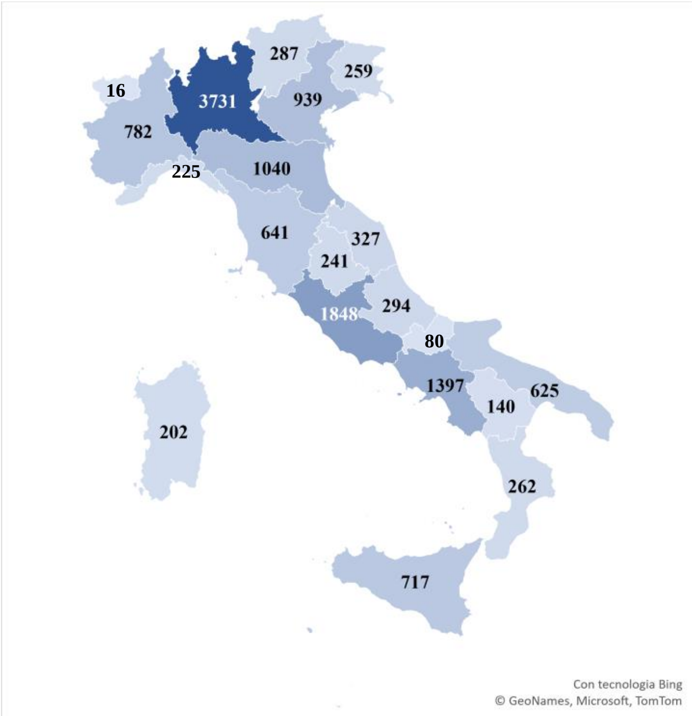
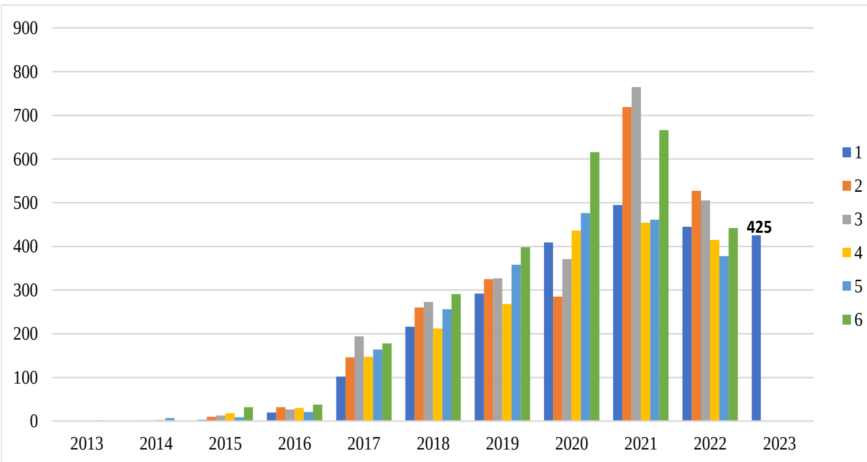


Fig. 2 The emergence of new Start-ups (bimonthly data)



The mortality rate in Figure 3 was calculated taking into consideration the year of establishment and that of the latest available financial statements. Specifically, if the closing year of the last available financial statements is less than 2021, the activity of the start-up is considered concluded (operational otherwise). The mortality rate is homogeneous among North regions (except for Valle d'Aosta), while South-central regions like Abruzzo, Molise, Sardegna, Umbria and Puglia have a low mortality rate compared to other regions in the same area. The national mean is equal to 4,21%.

Fig. 3 Mortality rate of Italian start-ups (04/21/2023)

Table 1 shows the frequency distribution of the Italian start-up sector. The mode sector is represented by the service industry, which is followed by industry and crafts, commerce and other sectors, such as tourism, agriculture and fishing.

Table 1 Frequency distribution of the Italian start-up sector

Sector	Absolute frequency	% frequency
Service industry	11213	79,79%
Industry and crafts	2115	15,05%
Commerce	456	3,24%
Other (tourism, agriculture and fishing)	212	1,91%
Total	13996	100%

3 Conclusions

Our findings are valuable for the following reasons. Startup founders, investors, and policy makers will find our results on startup class trends over time and space useful for their decisions about where to locate companies, where to invest, or who to invite to collaborate. In addition, our paper includes many empirical findings that can be linked to previous literature on entrepreneurship, sometimes corroborating previous outcomes with new evidence and sometimes challenging conclusions drawn previously. Finally, this work presents a case study that highlights the importance of a new business model.

References

1. Acs, Z.J. and Audretsch D.B: Innovation, market structure and firm size. The Review of Economics and Statistics, \textbf{69}, 4, 567-574 (1987)
2. Akcigit, U., Kerr W.R.: Growth through heterogeneous innovations, NBER Working Paper No. 16443 (2012)
3. Audretsch D.B., Falck O., Heblich S., Lederer A.: Handbook of Research on Innovation and Entrepreneurship. Edward Elgar Publishing Limited, Willinston, US (2011)
4. Baumol W.J.: Entrepreneurial enterprises, large established firms and other components of the free-market growth-machine. Small Business Economics, \textbf{23}, 9-21 (2004)
5. Cerved report PMI (2022). Available from https://research.cerved.com/rapporti/rapporto-cerved-pmi-2022/. Cited 13 March 2023
6. Decreto-legge 18 ottobre 2012, n. 179 https://www.mise.gov.it/images/stories/Art25-dl179-2012.pdf)
7. Florida, R. and Hathaway I.: Rise of the global startup city: The new map of entrepreneurship and venture capital (2018). https://startupsusa.org/global-startup-cities/report.pdf. Cited 1 April 2023
8. Global entrepreneurship network. The Global Startup Ecosystem Report GSER 2020. The New Normal for the Global Startup Economy and the Impact of COVID-19. (2020). Available from https://startupgenome.com/reports/gser2020. Cited 30 March 2023
9. Kerr, W.R. and Nanda, R.: Financing Innovation, NBER, Working Paper No. 20676 (2014)
10. Palazzo, L., Sabatino, P., Ievoli, R.: Determinants of social startups in Italy (2021) https://doi.org/10.36253/978-88-5518-304-8.18.
11. Registro Imprese.it https://www.registroimprese.it. Cited 13 March 2023)
12. Russo P. F. & Magri S. & Rampazzi C.: Innovative Start-Ups in Italy: Their Special Features and the Effects of the 2102 Law. Politica economica, Società editrice il Mulino, issue 2 (2016)
13. Savin, I., Chukavina, K. & Pushkarev, A. Topic-based classification and identification of global trends for startup companies. Small Bus Econ \textbf{60}, 659–689 (2023) https://doi.org/10.1007/s11187-022-00609-6

Permanent establishments and efficiency analysis with global enterprises

Branch estere ed analisi dell'efficienza con imprese globali

Antonio Frenda* and Enrica Sepe

Abstract The main objective of this work is to have a valid decisional support system that provides useful information for structural and economic intervention programs concerning global enterprises, by measuring the commercial presence through branches in foreign markets. The article summarises the results of the analysis of value added (VA) in global enterprises [1] through branches and the relation with intermediate consumptions. Through the aggregated and sectoral data provided by Outward Fats and regional production tax (2019), the presence of global resident enterprises and control over the branches without legal autonomy [2] was verified and ascertained for the year 2019: these data describe the activities of foreign affiliates abroad controlled by Italian legal units. The method and the results achieved already provide a useful interpretation of the efficiency frontier for the evaluation of value added.

Abstract *L'obiettivo principale di questo lavoro è quello di presentare un valido sistema di supporto decisionale che fornisca informazioni utili per i programmi di intervento strutturale ed economico riguardanti le imprese globali [1], misurando la presenza commerciale attraverso filiali nei mercati esteri. L'articolo riassume i risultati dell'analisi del valore aggiunto (VA) ottenuto attraverso le branch estere e la quota dei consumi intermedi. Attraverso i dati aggregati e settoriali forniti da Outward Fats e Irap (2019), è stata verificata e accertata per l'anno 2019 la presenza*

Antonio Frenda
Italian National Institute of Statistics; e-mail: frenda@istat.it

Enrica Sepe
Department of Economics, Management, Institutions, University of Naples Federico II; e-mail: enrica.sepe@unina.it

* The view expressed are those of the authors and do not necessarily reflect the views of Istat

di imprese residenti che controllano sedi secondarie estere prive di autonomia giuridica [2]: questi dati descrivono le attività di tali affiliate estere che sono quindi controllate da unità legali italiane. Il metodo ed i risultati raggiunti forniscono un'utile interpretazione della frontiera dell'efficienza per la valutazione del valore aggiunto.

Key words: Data envelopment analysis; Stochastic frontier analysis; Operations with correlated parties; Secondary seats; Branch exemption

1 Introduction

Foreign affiliate [3] in the framework of these work is a not resident branch over which a resident institutional unit (table 1) has direct or indirect control, since FATS data transmitted by the countries to Eurostat are compiled according to the Ultimate Controlling (UCI) that is an institutional unit (IU), proceeding up a foreign affiliate's chain of control, which is not controlled by another IU.

Table 1 Representing a Global Enterprise

Enterprise components				***Economic activities (NACE - Rev.2)***		
	Act. 1	Act. 2	…	Act. j	Act. n	Total
Offices	$LKAU_{11}$	…	…	$LKAU_{1j}$	$LKAU_{1n}$	Local $Unit_1$
Domestic secondary seats	…	…	…	…	…	Local $Unit_2$
Foreign permanent establishments	…	…	…	…	…	…
Foreign non permanent establishments						
Virtual Showroom	$LKAU_{m1}$	…	…	$LKAU_{mj}$	$LKAU_{mn}$	Local Unit $_m$

The transition from business accounting to national accounts must be carried out as required by European system of accounts [4] where IU that make up a country's economy and whose flows and stocks are recorded are those that are resident; the residence of a unit is defined as the economic territory in which its predominant centre of economic interest is situated therefore where a unit is active and carries out significant operations for at least one year: in this case a resident unit situated in the economic territory within which it carries out this activity should be identified for statistical purposes [5]. A combined use of statistical and administrative sources let to estimate businesses representing traditional and new boundaries of globalization.

2 Material, Methods and results

The method described in this work is suitable for the evaluation of efficiency. The efficiency estimates obtained have been utilized to rank the enterprises according to the common efficiency index. The comparison with the results obtained through the Stochastic Frontier Analysis [6,7] with same input and different outputs correlates very well with the results of the DEA. This result confirms the quality of the alternative method proposed in this paper. Therefore, we assess sectoral efficiency for permanent establishments abroad that are controlled by an Italian UCI: stochastic frontier analysis (SFA) and a nonparametric deterministic model structure (DEA) are the tools used to investigate the structural efficiency determinants in the paper [8,9,10]. We started from the model including all variables and interactions. The choice of the functional form has been taken under the hypothesis of a parsimonious model. The results show that construction and retail sector, where branches of clothing are clearly dominant, together with the Engineering and Enterprise services; scores tell us that this model identifies some efficient data making units (sectors) that reflect their distance from the frontier. The main input variables that are investigated (for each sector) are: intermediate consumptions, personnel costs, the ratio between the number of branches' employees over enterprises' employees; the sectoral rate of foreign production from direct controlled branches over enterprises' global production [11]. The coefficients of the variables are positive except in the case of personnel costs. Finally, we can affirm that intermediate consumptions and labour costs have a significant impact on the determination of the production frontier by global enterprises.

References

1. Frenda, A.: Approcci per la stima delle produzioni estere relative alle imprese nazionali. Corporate Governance and Research & Development studies, Open Access (2021)
2. Calzaroni, M., Pascarella C.: Units of the Production Process in the New National Accounts. Problems of Analysis and Measurement. XXXIX SIS Scientific Meeting, Sorrento, 14-17 April 1998
3. Amante, S., Ambroselli, S., Boselli, C., Faramondi, A., Nardecchia, R., Vicari, P.: Intensive Profiling. Istat working papers n. 4/2016. Roma: Istat. (2016) https://www.istat.it/it/archivio/181845
4. Eurostat: European System of National and Regional Accounts 2010 (ESA 2010). Published in the Official Journal on 26 June (2013)
5. Eurostat: European business profiling. Recommendations Manual (2020 Edition). Manuals and guidelines. Luxembourg: Publications Office of the European Union (2020)
6. Scippacercola, S., Sepe, E.: Ordinal principal component analysis for a common ranking of stochastic frontiers. Journal of Applied Statistics, 43.13: 2442-2451 (2016)
7. Scippacercola, S., Sepe, E.: Critical comparison of the main methods for the technical efficiency. Electronic Journal of Applied Statistical Analysis, 9.4: 760-780 (2016)
8. Kumbhakar S.C., Lovell C.K.: Stochastic frontier analysis, Cambridge University Press (2003)
9. Rao D.P, O'Donnell C.J., Battese G.E, Coelli T.J.: An introduction to efficiency and productivity analysis, Springer (2005)
10. Ray S.C.: Data envelopment analysis: theory and techniques for economics and operations research, Cambridge University Press (2004)
11. De Gregorio, C., Monducci, R.: Aspetti territoriali dei mercati di riferimento e delle relazioni fra unità produttive nelle strategie delle piccole e medie imprese manifatturiere. ISTAT (2011)

Techniques and constructs in some recent market and organizational research

Tecniche e costrutti in alcune recenti ricerche di mercato e sulle organizzazioni

Ivan Arcangelo Sciascia

Abstract The analysis of constructs applied in the study of market segmentation can be applied to other research fields such as the study of organizations. Through a study with a questionnaire survey, it is possible to detect constructs such as consumer satisfaction and loyalty to a given good or service by considering customer loyalty categories proposed in the literature. It is also possible to support the study of value systems of customers or workers of an organization to analyze the correlations through a latent class model.
Abstract *L'analisi di costrutti applicati nello studio di segmentazione di mercato può essere applicata ad altri campi di ricerca come lo studio delle organizzazioni. Attraverso uno studio con inchiesta tramite questionario è possibile rilevare costrutti quali la soddisfazione e fedeltà del consumatore ad un dato bene o servizio considerando categorie di fedeltà del cliente proposte in letteratura. E' possibile inoltre affiancare lo studio di sistemi di valori dei clienti o lavoratori di una organizzazione per analizzare le correlazioni attraverso un modello a classi latenti.*

Key words: customer loylaty, list of values, latent class analysis, organizational assets

1 Introduction

The list of values is a theme that has been studied in the literature of both market research [4] and organizational culture studies [11]. Market segmentation makes it possible to reduce the complexity of the market where many customers address

Ivan Arcangelo Sciascia
Università degli Studi di Torino, Dipartimento di Scienze della vita e biologia dei sistemi, viale Mattioli 25 10123 Torino, e-mail: ivan.sciascia@unito.it

different behaviors and attitudes to corporate brands such as a brand of clothing or a high-tech product.

Studies related to brand loyalty have considered antecedents that can be brand affect, brand performance [2]. As brand can be considered an organizational asset the concepts of attitude towards a brand can be linked through survey studies on consumer values [3]. Customer loyalty can be considered an attitude that precedes brand loyalty and in his study [8] Oliver proposes four categories into which customer loyalty can be divided: cognitive, affective, conative and action.

List of values were considered in the Rokeach 1973 survey work [9] where respondents could order a list of values. This survey has been re-proposed in some studies [11] in relation to research on organizations.

In this study, through a latent class model, we relate customer loyalty and the four categories [8] to a reworking of the ordering of the list of values [9, 10].

2 List of Values

Models of segmentation based on values have been developed [3] and resulting market segments can be considered influenced on the dynamics of values [1, 4, 7]. The values considered in our previous studies study [5, 6] are: Friendship, Love, Study or job, Family, Sport or hobby, Self-respect, Social recognition, A world at peace, A comfortable life, Self-realization, Forgiveness ability, Equality.

System values variables can be correlated to a loyalty index investigated on the basis of items proposed in the questionnaire which can include the categories considered in the [8] literature.

These are examples of item considered:
Would you switch brand for this service/product with the same performance and price? (cognitive)
Did you buy it regardless of price? (affective)
Would you repurchase a service/product of the same brand? (conative)
Would you repurchase it because you are oriented to this brand? (action)

On this responses we can calculate a loyalty index *LY* and relate it to classification with latent class analysis.

3 Latent class analysis

After calculating a loyalty index with a logistic regression model, it is possible to correlate the customer loyalty construct with the ranking given by respondents to a list of values. We can thus build a model that correlates the probability of inclusion in a class to the ranking of the values expressed by the respondents.

Thus, for example, one can describe whether the respondents belong to high or moderate or low fidelity classes in correlation with value systems, for example altruistic or self-realization. The variables considered are VL as value ranking and LY for loyalty.

Considering the loyalty LY The vector of characters possessed by individual i is

$$\mathbf{u}_i = (VL1_i, \ldots\ldots, VLn_i)$$

where each variable VLj_i is still subject i's rank of Value j among those we are considering and n are the values considered for the statistical analysis

We consider then a logit model to correlate VL and LY with M latent classes:

$$P(LY_i = s) = \Sigma_{m=1}^{M} p_s \Lambda(\mathbf{u}_i \beta_m) \tag{1}$$

where $P(LY_i = s)$ is the individual loyal level s, $\mathbf{u}_i$ is the vector of VLn variables considered explanatory, β_m is a vector of regression parameters for segment s, and $\Lambda(\cdot)$ is the cumulative logistic distribution function. For every segment, $\mathbf{u}_i\beta_m$ and $m = 1, \ldots, M$ equals

$$\mathbf{u}_i\beta_m = \beta_{m0} + \beta_{m1} \cdot VL1_i + \ldots.. + \beta_{mn} \cdot VLn_i.$$

4 Discussion

In this study we considered the possibility of correlating loyalty indices with a ranking of a value system. This study can be helpful in understanding both how market segments are dynamic based on changing value systems and how changing value systems can affect organizational assets.

System values can be evaluated based on a rielaboration of original Rokeach survey [9] and we can measure values through a questionnaire in which we detect the priorities of the values, such as altruistic and self-realization values.

This study can be applied to know the value systems of the consumer and workers in an organization as we can consider that values influence choices and attitudes influencing organizational assets.

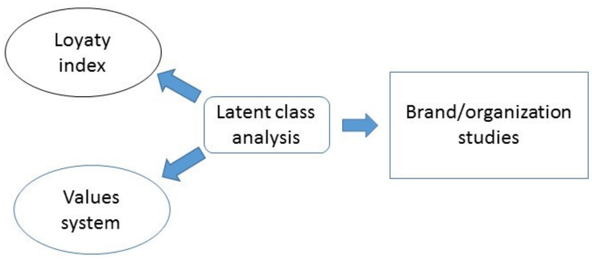


Fig. 1 A statistical model to evaluate attitudes towards organizational assets

Values and customer satisfaction and loyalty studies can be used to evaluate attitudes towards organizational assets such as brand equity or corporate culture. Through a statistical model of measurement as described in Figure 1 it is therefore possible to integrate the study of variables already proposed in different fields with questionnaires.

References

1. Brangule-Vlagsma, K., Pieters, R.G.M. and Wedel W.: The dynamics of value segments: modeling framework and empirical illustration. Int. J. Res. Mark. 19(3), 267-285 (2002)
2. Chaudhuri, A. and Holbrook, M.B.: The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. J. Mark. 65, 81-93 (2001)
3. Kamakura, W.A. and Mazzon, J.A.: Value segmentation: a model for the measurement of values and value-systems. J. Consum. Res. 18(3), 208-218 (1991)
4. Kamakura, W.A. and Novak, T.P.: Value-system segmentation: exploring the meaning of LOV. J. Consum. Res. 19(2), 119-132 (1992)
5. Montinaro, M., Dal Forno A., Lo Presti, A., Sciascia, I.A. (2009). From market segmentation to customer loyalty: an exploratory study. In: 57° ISI Proceedings. Durban, South Africa 163-166 (2009)
6. Montinaro, M. and Sciascia, I.: Market segmentation models to obtain different kinds of customer loyalty. J. Appl. Sc. 11(4), 655-662 (2011)
7. Novak, T.P. and MacEvoy, B.: On comparing alternative segmentation schemes: the list of values (LOV) and values and life styles (VALS). J. Consum. Res. 17(2), 105-109 (1990)
8. Oliver, R.L.: Whence consumer loyalty? J. Mark. 63, 33-44 (1999)
9. Rokeach, M.: The nature of human values, New York: Free Press (1973)
10. Rokeach, M. and Ball-Rokeach, S.J.: Stability and change in american value priorities, 1968 - 1981. Am. Psychol. 44(5), 775-784 (1989)
11. Tuulik, K., Õunapuu T., Kuimet, K., and Titov, E.:Rokeach's Instrumental and Terminal Values As Descriptors of Modern Organisation Values. Int. J. Organiz. Leadership 19(2), 151-161 (2016)

The value of buildings in the Italian general government balance sheet

Il valore degli immobili nei conti patrimoniali delle Amministrazioni pubbliche in Italia

Paola Santoro and Antonio Regano

Abstract In July 2021 Istat published new series of the value of non-financial stocks by institutional sector, introducing innovations and improvements in methods and sources. The paper describes the estimation approach for the real estate of general government, emphasizing the importance of treating data sources to be consistent with definitions, classifications and valuation principles of national accounts. The main results and international comparisons are presented.
Abstract *A luglio 2021 Istat ha pubblicato le nuove stime del valore delle attività non finanziarie dei settori istituzionali, introducendo innovazioni e miglioramenti nei metodi e nelle fonti. Il documento descrive la metodologia adottata per gli immobili delle Amministrazioni pubbliche ed il trattamento dei dati necessario per la coerenza con le definizioni, classificazioni e criteri di valutazione dei conti nazionali. Vengono presentati i principali risultati e confronti internazionali.*

Key words: Real estate, non-financial stocks, institutional sector, dwellings, non-residential buildings, general government, balance sheets

---

Paola Santoro
Istat - Italian National Institute of Statistics, National Accounts Department; e-mail: pasantor@istat.it

Antonio Regano
Istat - Italian National Institute of Statistics, National Accounts Department; e-mail: regano@istat.it

1 The new estimation method for the Italian general government real estate[1]

Istat estimates the value of the residential and non-residential buildings of Italy's General Government as part of the balance sheet of institutional sectors.

The national accounts framework is built around a sequence of interconnected accounts.[2] The full sequence of accounts for the institutional sectors (non-financial corporations, financial corporations, general government, households, non-profit institutions serving households) consists of current accounts, accumulation accounts and balance sheets. This way it is possible to describe, in a coherent and complete framework, all the flows generated by current economic activities and the accumulation process that stems from them. Balance sheets record the values of financial and non-financial assets owned by each institutional sector and the value of liabilities held to finance its activity.

The value of real assets, estimated by Istat since 2015 annually, together with the value of financial assets and liabilities, published by the Bank of Italy, represents the total wealth held by institutional sectors [4]. Non-financial wealth calculated by Istat expresses the value of the tangible assets (e.g., real estate; machinery and equipment including transport and ICT equipment) and intellectual property products (mainly research and development and software), inventories and land under cultivation.

In July 2021 Istat published new series of the value of non-financial stocks by institutional sector produced following an extraordinary revision, aimed at introducing innovations and improvements in methods and sources [5, 8]. One of the main revisions is the new method for estimating buildings owned by general government (hereafter also referred to as S13 according to ESA 2010 coding).

A new source of information is introduced, i.e., the "General government buildings census" (henceforth also indicated as DT Census) carried out by the Treasury Department (DT) of the Ministry of Economy and Finance (law 191/2009). Very detailed information is provided in this statistical source at micro level: for each surveyed property, the database indicates the type, time of construction, surface, type of use (currently used, unused, unusable), right of use (ownership, rental, free use), identification code (cadastral data), address, name and tax code of the owner. For each property, the information gathered from the census is then supplemented with an estimate of its current value. Every unit is evaluated according to a methodology (mass appraisal estimate), developed by the DT in collaboration with Sogei (*Area Modelli di Previsione e Analisi Statistiche*). The estimation procedure is conducted at micro level for all assets, divided into seventeen homogeneous clusters[3] by type, applying three alternative methods: the "comparative method", the "cost approach"

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Istat. The paragraphs were authored as follows: §1 and 3 were written by P. Santoro and §2 by A. Regano.

[2] The rules for preparing the national accounts are described at European level in the European System of accounts (ESA 2010) [1] and at global level in the System of National Accounts 2008 (SNA 2008) [10].

[3] Dwellings; offices; hotels; shops; collective parking lots; warehouses and storage rooms; buildings for production activities; scientific laboratories; garages; historic buildings; collective residential structures; cellar houses; barracks; schools; hospitals; prisons; sports facilities.

and the "residual value method". These three approaches are widely used in national and international accounting practice and standards [9]. They all provide valuation consistent with ESA 2010 rules, i.e., current market prices. The choice of method depends on the cluster in which the property is classified and on the type of use [6]. The census and national accounts classifications and domains do not overlap perfectly; therefore, it is important to analyze the correspondence of their definitions and then to treat the input data in order to obtain accurate estimates. On the one hand, some units that do not fit the ESA 2010 definitions of the S13 sector (e.g., IACP) or of the assets classified as buildings (some types of other structures) are excluded from the DT census database for estimating S13 buildings. On the other hand, the DT census does not impute values for non-respondents and excludes some types of buildings whose value is difficult to measure; moreover, some units classified in the S13 sector according to national accounts are not included in the domain of the DT census (out of the scope). Since ESA 2010 requires exhaustiveness for national accounts, completeness is obtained by integrating the DT estimates through an imputation procedure based on a "price-by-quantity approach": median surfaces and prices are calculated in reference strata on data provided by respondents. Two stratifications are carried out:

1. First stratification: median surfaces and prices are computed by type of institutional unit[1], cluster, geographical distribution[2] and size class of the municipality (in terms of population).
2. Second stratification:
   a) median surfaces are computed by type of institutional unit, cluster
   b) median prices are computed by cluster, geographical distribution and size class of the municipality.

Surfaces and prices for units that are non-respondents or out of the scope are imputed based on data from the first or second stratum according to the significance of information of the strata (number of observations[3]).

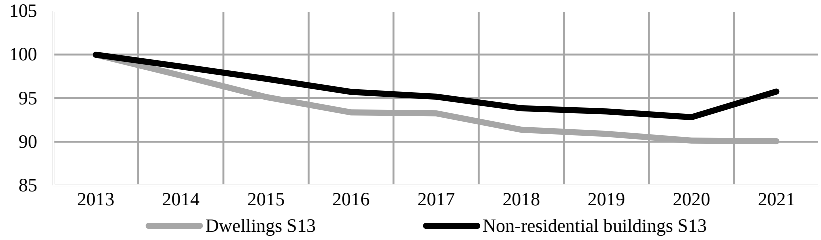
Finally, all properties are classified as "dwellings" or "non-residential buildings" as defined by ESA 2010, through a univocal correspondence. The only exception are garages and warehouses: they may be either residential units (appurtenances to dwellings as accessory areas) or non-residential buildings; in order to make the necessary reclassification to match the ESA 2010 coding, these building types are divided into dwellings or non-residential buildings based on the size of the unit.[4]

[1] All S13 units are divided into twenty-two homogenous groups.
[2] NUTS 1: Northeast Italy, Northwest Italy, Central Italy, South Italy, and Insular Italy.
[3] If the first stratum is empty or has fewer than 30 observations, data from the second stratum are used. The same imputation procedure is used to evaluate properties for which Census DT provides surfaces but not values. For some non-respondents units (social security funds other than INPS and INAIL), values are taken from their financial statements. Finally, an estimate of real estate assets from public-private partnership contracts (outside the scope of the census) is added. It is calculated by applying the perpetual inventory method, adjusted to take into account the value of the underlying land. For the perpetual inventory method, see [3], [7].
[4] Units less than 30 square meters (warehouses) or less than 50 square meters (garages and parking spaces) are classified as residential outbuildings.

2 International comparison on general government real estate

Figure 1 shows the evolution of the S13 real estate wealth, classified as "dwellings" or "non-residential buildings", from 2013 to 2021. The values of the two assets decrease significantly until 2020, especially in 2013-2018; this decline is mainly due to the falls in the real estate market prices for almost all the years of this period. In 2021 the value of dwellings is quite stable and compared with 2013 it decreased by 9.9%. In 2021 there is a rise for non-residential buildings and with respect to 2013 the value decreased by 4.4%.

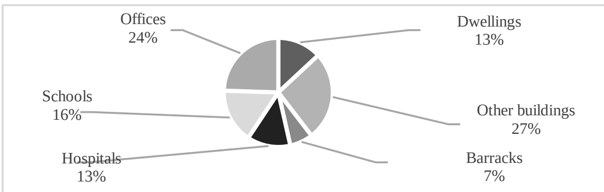
Fig. 1 S13 dwellings and non-residential buildings - Years 2013-2021, index 2013=100



Source: our elaboration on Istat and Ministry of Economy and Finance data.

Figure 2 shows the composition by type of buildings in 2018: 24% are offices, 16% schools, 13% dwellings and hospitals, 7% barracks and 27% are other buildings, mainly historic buildings (such as libraries, art galleries, theaters, cinemas) and warehouses and storage facilities.

Fig. 2 S13 buildings by type – Year 2018 - Percentages



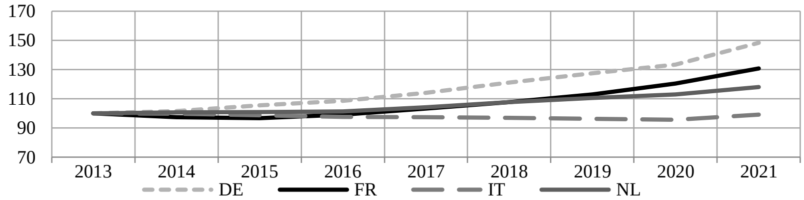
Source: our elaboration on Istat and Ministry of Economy and Finance data.

A comparison with other advanced European economies is now presented, although the data are not perfectly comparable between countries[1]. Therefore, it is

[1] The international comparison is based on non-financial asset statistics published on the Eurostat website (for Germany, France, the Netherlands) and on Istat website (for Italy). On Eurostat website, buildings do not include the value of underlying land, which is entered under non-produced non-financial asset "land" (ESA 2010 and SNA 2008), see also [2]. In the statistics published by Istat at national level, and included in this paper, the value of buildings (also referred to as "real estate") includes the value of the underlying land as a result of the adopted estimation method.

important to calculate homogeneous and comparable assets, henceforth defined as "constructions"[1] and "total non-financial assets"[2].

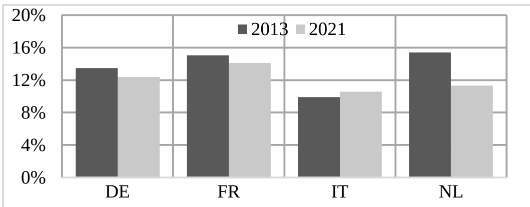
Fig. 3 S13 constructions in international comparison, Years 2013-2021, index 2013=100



Source: our elaboration on Eurostat and Istat data

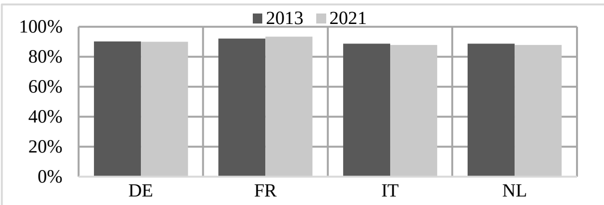
The value of S13 constructions (Figure 3) increases in Germany and the Netherlands for all years of the series and from 2016 in France; in particular, from this year, Germany and France are experiencing a rapid growth. Italy's S13 constructions remain stable from 2013 and 2021.

Fig. 4 Ratio of S13 constructions to S1 constructions (per cent) - Years 2013, 2021



Source: our elaboration on Eurostat and Istat data

Fig. 5 Ratio of constructions to total non- financial assets S13 (per cent) – Years 2013, 2021



Source: our elaboration on Eurostat and Istat data

Figure 4 shows the ratio of S13 constructions on the total economy (S1) constructions. During this period, the ratio has decreased in Germany (13.5% in 2013 and 12.4% in 2021), France (15.1% in 2013 and 14.1% in 2021) and the Netherlands (15.4% in 2013 and 11.3% in 2021); however, the ratio in Italy has increased: it was 9.9% in 2013 and became 10.6% in 2021. The weight of constructions on the total non-financial assets owned by the general government (figure 5) is quite the same in all countries (about 88%-90%) and it is stable in the period 2013-2021.

[1] It is the value of all types of construction (dwellings, non-residential buildings, other structures including land improvement) and land (land underlying buildings, under cultivation…).
[2] It is calculated as the sum of fixed capital and land; inventories are excluded, as these data are not available for some countries.

3 Conclusion

The estimation of real estate is a challenging issue for national accountants, and the results provide important insights for the analysis of wealth accumulation and distribution. If we consider the value of real estate as the combined value of the buildings and the underlying land, for the Italian total economy it accounts for about 75% of total non-financial wealth, and in particular 95% for households. Consequently, any improvement in estimates of real estate is crucial.

Buildings are also relevant for S13, accounting for one-third of the sector total non-financial wealth of the sector[1] and for 21% of its gross wealth (sum of real and financial assets). Italy has recently made a major effort to introduce innovations and improvements in methods and sources for estimating the S13 real estate. Some open issues still remain, especially the valuation of historic and artistic heritage: it may suffer from a systematic underestimation that penalizes countries like Italy even in the international comparison. International guidelines on appropriate and harmonized estimation methodologies would be necessary.

References

1. EUROSTAT: European System of Accounts, ESA 2010 (2010), Luxemburg
2. EUROSTAT, OECD: Compilation guide on land estimation (2015)
3. EUROSTAT, OECD: Survey of National practices in estimating net stock of structures (2015)
4. ISTAT, THE BANK OF ITALY: The wealth of Italy’s institutional sectors, years 2005-2021 (2023)
5. ISTAT: Non-financial wealth in Italy, years 2005-2019 (2021)
6. MEF: Modello di stima del valore del patrimonio immobiliare pubblico (2018)
7. OECD, Measuring Capital Second edition (2009), Paris
8. Santoro, P., Tartamella, F.: Estimate of Real Estate in Italian Balance Sheet, International Association for Research in Income and Wealth I.A.R.I.W. Conference, Oslo (2021)
9. TEGOVA, European Group of Valuers' Associations: European Valuation Standards (2020)
10. UNITED NATIONS STATISTICAL COMMISSION (UNSC), System of National Accounts (2008), New York

[1] This weight differs from the one included in the previous section on international comparison as it does not include the value of other structures such as road works and other civil engineering works.

Linear and nonlinear factors affecting default risk in the peer-to-peer lending market

Fattori lineari e non lineari che influenzano il rischio di default nel mercato dei prestiti peer-to-peer

Francesco Giordano, Sara Milito and Maria Lucia Parrella

Abstract In this article we apply a fully nonparametric variable selection approach to select the predictor variables (among macroeconomic, borrower and state-specific characteristics) affecting the default risk in the peer-to-peer lending market, also distinguishing linear predictors from nonlinear ones. We consider data from loan book of LendingClub, a FinTech lending company in US that provides loans through an on-line platform. We use a screening selection approach, the Derivative Empirical Likelihood Sure Independence Screening (DELSIS) proposed by [3], in combination with a subsampling technique. The results show that some macroeconomic variables, such as inflation and the interest rate, are relevant and have a nonlinear effect on default risk.

Abstract *In questo articolo applichiamo un approccio non parametrico per selezionare le variabili predittive (tra caratteristiche del mutuatario, specifiche dello stato e macroeconomiche) che influenzano il rischio di default nel mercato dei prestiti, distinguendo inoltre le variabili lineari da quelle non lineari. Consideriamo i dati del portafoglio prestiti di LendingClub, una società di prestito FinTech degli Stati Uniti che opera attraverso una piattaforma on line. Utilizziamo come approccio di selezione il metodo DELSIS (Derivative Empirical Likelihood Sure Independence Screening) proposto da [3], combinandolo con una tecnica di subsampling. I risultati mostrano che alcune variabili macroeconomiche, come inflazione e tasso di interesse, sono rilevanti e hanno un effetto non lineare sul rischio di default.*

Key words: Default, Peer-to-peer lending, Variable Selection, Nonparametric.

Francesco Giordano
University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy e-mail: giordano@unisa.it

Sara Milito
University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy e-mail: smilito@unisa.it

Maria Lucia Parrella
University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy e-mail: mparrella@unisa.it

1 Introduction

Since 2005, peer-to-peer (P2P) lending has been increasingly adopted as an alternative financing method. Over the past decade, many start-ups have started using P2P lending to secure the funds they need, through platforms that directly connect borrowers to lenders. The P2P service mediates between the parties by charging a service fee, eliminating the financial institution as the intermediary (investors act as suppliers and borrowers act as demanders). The risk is that a large portion of investors does not get back the borrowed money after the loan expires. In other words, a borrower no longer repays the debt. Therefore, while P2P lending platforms meet the demand and supply of funds, they must correctly estimate the probability of loan default.

Most of the works in the recent literature have focused on determining the factors related to default in this context, considering only the borrower characteristics and through the use of a parametric model. On one hand, even in P2P platforms, the default might depend on other factors, such as the economic environment, state-specific demographics and country-specific economic conditions, the impact of inflation and interest rate, as it has also been pointed out by [5]. On the other hand, the use of parametric models can lead to misspecification problems. The literature lacks in the use of a variable selection approach which, taking all the variables referring to the characteristics listed above, identifies those that are linked to the probability of default. Moreover, to improve robustness of the results, it is desirable to use a variable selection approach that works with nonparametric models, which, by not imposing assumptions on the relationship between the response variable and the explanatory variables, is much more flexible than its parametric counterparts.

Taking all these considerations into account, in this study we identify the risk factors that may influence P2P loan delinquencies in US areas, between 2018 and 2019, using a fully nonparametric variable selection procedure. We start by using a nonparametric screening selection approach proposed by [3], which is able not only to rank the main risk factors related to the occurrences of default, but also to discriminate between linear and nonlinear predictors. Furthermore, in this work this screening approach is also combined with a subsampling technique which allows to transform the screening procedure into a more precise selection procedure for the identification of the relevant variables.

The rest of the paper is organized as follows. In Section 2 we present the dataset and the variable selection approach used. The results of the real data analysis and some concluding remarks are reported in Section 3.

2 Material and methods

2.1 Preliminary analysis

In this paper, we retrieved data from [6], where the author combines data covering the period from 2008 to 2019 by the raw loan book dataset of the LendingClub (USA) (a FinTech lending company that provides loans through a technology-driven platform). The dataset includes loan specific, loan type and borrower specific variables, and other county-level data from a multitude of open source and government datasets. As a result, there are 2,703,430 observations with 195 features in total.

To identify linear and nonlinear risk factors, only the last two years of observations are considered (2018 and 2019), as they are the years with the highest number of loans. Although the United States uses federal risk-sharing policies to protect the creditworthiness of states, states have different economic and demographic characteristics. To account for these differences, we have aggregated the data by macro-area or zone (based on the first three digits of the zip code). Then, after eliminating the variables with a correlation greater than 0.9, we obtained a final dataset of 79 variables and 849 observations, one for each macro-area. An accurate description of the macroeconomic variables is given by [6], while in [4] all the characterizations of the other covariates are listed.

Furthermore, 200 noise variables, generated independently from a standard normal, are included into the dataset for control reasons (in order to check if the procedure is able to identify them as irrelevant variables). Finally, as a response variable, we consider a proxy for the default risk, dividing the total number of defaults by the number of loans, for each area.

Thus, to summarize, we apply the variable selection procedure, described in the next subsections, on two datasets for the two years (each of 200 + 79 potential predictors and 849 observations).

2.2 The proposed variable selection procedure

Suppose we have a random sample $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ from the data model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where ε_i is the error with zero conditional mean, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots, X_{ip})$ is the $p-$vector of candidate variables and Y_i is the response variable. The index i is used to denote a given unit in the sample while index j refers to a given variable. For simplicity of notation, we will generally drop the index i and will just keep the index j for variables. We assume that model (1) is sparse, *i.e.* only a small fraction of the candidate variables contributes to the response ($s \ll p$). Moreover, no specification of the unknown regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ is required, and, without loss of generality, we assume $E(Y) = 0$, so that $E(m(\mathbf{X})) = 0$.

Let us denote with

$$M_* = \{1 \leq j \leq p : \text{the } j\text{-th variable in } \mathbf{X} \text{ is relevant for explanation of } Y\}$$

the set of s true relevant covariates of model (1). This set can be partitioned in two parts: the set N_* of nonlinear covariates, having nonlinear effects on the response variable Y, and the set of linear covariates, denoted with L_*.

In order to estimate the three sets M_*, N_* and L_*, we use an independence model-free feature screening technique, called Derivative Empirical Likelihood Independence Screening (DELSIS), proposed by [3]. It is based on: *a)* the local polynomial regression, to estimate the marginal contribution of X_j locally at x, through the first derivatives of the marginal functions $f_j(x) = E(Y|X_j = x)$, for $j = 1, \ldots, p$; and *b)* the empirical likelihood technique, to test if these derivatives are uniformly zero for all x in the support of each variable X_j. The procedure involves two steps: the first one is devoted to screen the relevant covariates while the second to filter the nonlinear ones. We omit the details of the procedure to save space and refer to [3] for more information. Actually, the DELSIS procedure is a *screening* method, not a *variable selection* method. The substantial difference lies in the number of starting and final covariates. In fact, screening approaches are used in selection problems characterized by an excessively high number of variables, say p, in order to easily reach a smaller number of variables, say $p_* << p$. More precisely, the result of a screening procedure is a ranked list of variables which contains the true relevant ones in the top p_* positions, with high probability (ensured by the so called *sure screening property*). On the other side, any variable selection approach starts with a moderately high number of candidate variables (*i.e.*, p_*) and has, as final result, the (estimated) few relevant covariates (*i.e.*, $\widehat{s} < p_*$). A common approach in high dimensional variable selection problems is to implement a screening method as a first stage of the procedure, and then take its final result as input of a variable selection technique (second stage of the procedure, usually based on penalized regression). The difficulty in this process is generally in the need to set complex regularization parameters in the final penalization stage of the procedure. This approach could also be used here, but we prefer to propose an alternative procedure that avoids penalization.

Therefore, in this paper, we consider an hybrid procedure which is a combination of [3] and [1] procedures, in the following way.

In the first stage of the procedure, we create 100 subsamples from the entire data set, each containing $n = 200$ observations ($i = 1, \ldots, 200$). Then we apply the first step of DELSIS (briefly, we performs a local quadratic derivative estimation and test the "relevance degree" of this function by the empirical likelihood technique, see [3] for the details). In this way we obtain 100 rankings (each for every subsample) and collect only the first $p^* = \lfloor n/\log(n) \rfloor$ positions of these rankings. We calculate how often each covariate appears in the top p^* positions of the ranked lists. The estimated set of relevant variables, say $\widehat{M}_*$, is given by those predictors for which the relative frequency is greater than 0.7.

In the second stage, we use the same 100 subsamples to screen nonlinearities (shortly, we performs a local quadratic derivative estimation and test the "nonlinearity degree" of this function by the empirical likelihood technique, see [3] for the details), obtaining new rankings based on the "nonlinearity statistic". Then we calculate how often each covariate appears in the first $q^* < p^*$ positions of these new rankings. In practical situation we could choose $q^* = p^*/2$. Once again, the set of relevant nonlinear covariates, say $\widehat{N}_*$, is given by those covariates that occur with a frequency greater than 0.7.

Finally, the estimated set of linear covariates, say $\widehat{L}_*$, is given by the relevant variables that are in $\widehat{M}_*$ and not in $\widehat{N}_*$.

3 Results and Conclusion

The results of our analysis are reported in Table 1. We highlight that for all the scenarios considered, the proposed variable selection procedure identifies correctly all the 200 noise variables added to the model as not relevant, in both years considered.

There are three variables with nonlinear effect during the two years: the percentage of US inflation (INFLUSA), the interest rate (INTRATE) and the number of internet users (INTUSER). The last variable is certainly related to the probability of delinquency since the number of loan increases as the users of the on-line platform increases. The other two are macroeconomic variables. This finding is consistent with empirical evidence in existing work by [5], where the authors explored the impact of inflation and interest rate on loan defaults in the P2P lending market. The authors highlighted that inflation reduces the real value of debt, interest payments and the value of real income. All this leads to a reduction in the funds available to the borrowers and thus the repayment of the debt becomes more difficult.

The variables TOTAL_REC_LATE_FEE (the amount of late fees received) and COLLECTION_RECOVERY_FEE (post charge off collection fee) have a linear effect in 2019 and a nonlinear in 2018. The amount of both covariates are related to the delay in repayment of the loan, so they definitely affect the risk of delinquencies. The delay in payments, and therefore the payment of the fee, could effect the default since a person who is late will have a greater probability of going into insolvency.

In 2018, the variables COLLECTIONS_12_MTHS_EX_MED (number of collections in 12 months excluding medical collections) and FEDFUNDS (Federal Funds Effective Rate) have a nonlinear effect on delinquencies. The first predictor is relevant also in [2], where the authors state that the higher is the level of reported collections in the last year, the higher is the probability of default. Increasing the Federal Reserve rate makes borrowing more expensive, so it has an effect on the probability of default because the cost (due to the interest rate) increases. In 2018, this variable is significant as this rate increases throughout the year, while it is not relevant in 2019 because for the majority of the year it remained fixed.

Table 1 Relevant linear and nonlinear variables on default risk in 2018 and 2019

Year	Linear	Nonlinear
2018	–	COLLECTION_RECOVERY_FEE
		COLLECTIONS_12_MTHS_EX_MED
		INTUSER
		FEDFUNDS
		INFLUSA
		INTRATE
		TOTAL_REC_LATE_FEE
2019	COLLECTION_RECOVERY_FEE	INTUSER
	TOTAL_REC_LATE_FEE	INFLUSA
		INTRATE
		RECOVERIES

In 2019, the variable RECOVERIES (post charge-off gross recovery) has a nonlinear effect on delinquencies. This is a relevant variable since a lender can use a charge when the borrower has become substantially insolvent after a period of time. In fact, a charge-off is when a company writes off the debt as a loss, deems it can no longer collect, as the borrower has defaulted on payments.

As a final concluding remark, we want to underline the fact that our procedure, which is used for the identification of the relevant variables that affect default risk in the peer-to-peer lending market and for their classification between linear and nonlinear covariates, does not require the setting of crucial regularization parameters. This is an advantage for the data analyst, compared to the difficulty he usually encounters in setting the values of the complex regularization parameters of the more classic penalized methods.

References

1. Baranowski, R., Chen, Y., Fryzlewicz, P.: Ranking-based variable selection for high-dimensional data. Statistica Sinica. 30, 1485-1516 (2020)
2. Croux, C., Jagtiani, J., Korivi, T., Vulanovic, M.: Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform. Journal of Economic Behavior & Organization. 173, 270-296 (2020)
3. Giordano, F., Milito, S., Parrella, M.: A nonparametric procedure for linear and nonlinear variable screening. Journal of Nonparametric Statistics. 34, 859-894 (2022)
4. Ko, P., Lin, P., Do, H., Huang, Y. : P2P Lending Default Prediction Based on AI and Statistical Models. Entropy. 26, 801 (2022)
5. Nigmonov, A., Shams, S., Alam, K.: Macroeconomic determinants of loan defaults: evidence from the US peer-to-peer lending market. Research in International Business and Finance. 59 (2022)
6. Nigmonov, A.: Dataset from the US Peer-to-peer Lending Platform with Macroeconomic Variables. Mendeley Data. 4 (2021) doi: 10.17632/wb3ndt69gf.4

Session of free contributes SFC6 - *Methodological issues*

1. *Macroeconomic Time Series Classification by Nonparametric Trend Estimation* (Feo G., Giordano F., Niglio M. and Parrella M.L.)
2. *A Normalization Method for Space-time Analysis of Evaluation and Quality Indicators* (Mazziotta M. and Pareto A.)
3. *Unveiling Latent Structures: exploring Multidimensional IRT Models using Dirichlet Process Mixtures* (Valentini P., Fontanella S. and Fontanella L.)
4. *On a technique to detect accounting data manipulation* (Passamonti C.)

Macroeconomic Time Series Classification by Nonparametric Trend Estimation

Classificazione di Serie Storiche Macroeconomiche mediante Stima Nonparametrica del Trend

Giuseppe Feo, Francesco Giordano, Marcella Niglio and Maria Lucia Parrella

Abstract This paper considers the classification of macroeconomic nonstationary time series where the nonstationarity is given by the presence of a deterministic trend. The classification is achieved by means of a nonparametric estimator of the first derivative of the trend, which is then used in a two stage procedure: the first stage selects the time series with constant trend and the second stage classifies the remaining time series between series with linear and non-linear trend, in order to allow an effective time series analysis. An application to the Gross Domestic Product (GDP) of the world countries is conducted to show the validity and applicability of the procedure.
Abstract *Questo articolo considera la classificazione delle serie storiche macroeconomiche non stazionarie dove la non stazionarietà è data dalla presenza di un trend deterministico. La classificazione è ottenuta per mezzo di uno stimatore nonparametrico della derivata prima del trend, che viene utilizzato in una procedura a due stadi, dove al primo stadio si selezionano le serie storiche con trend costante e al secondo stadio si dividono le rimanenti serie tra quelle serie con trend lineare e quelle con trend non lineare, al fine di permettere un'efficace analisi di tali serie. Viene condotta un'applicazione ai dati del Prodotto Interno Lordo (PIL) dei paesi del mondo per mostrare la validità e l'applicabilità della procedura.*

Key words: nonparametric regression, testing procedure, mixing processes, GDP, nonstationary time series.

Giuseppe Feo
University of Salerno, Via Giovanni Paolo II 132, Fisciano (SA), Italy, e-mail: gfeo@unisa.it

Francesco Giordano
University of Salerno, Via Giovanni Paolo II 132, Fisciano (SA), Italy, e-mail: giordano@unisa.it

Marcella Niglio
University of Salerno, Via Giovanni Paolo II 132, Fisciano (SA), Italy, e-mail: mniglio@unisa.it

Maria Lucia Parrella
University of Salerno, Via Giovanni Paolo II 132, Fisciano (SA), Italy, e-mail: mparrella@unisa.it

1 Introduction

Trend composition is a very important issue in time series analysis. In the recent literature there are examples of how trend composition is becoming the subject of an automation process (see [7] and [1]). In fact, the interest in knowing the type of the trend (constant, linear or nonlinear) turns out to be of considerable importance when a time series needs to be decomposed and analysed. The de-trending procedure, based on differencing the series, tends to transform the series into one which is mean stationary but does not provide any information on the nature of the trend. For example, if a time series has a parabolic trend, the transformed series obtained from the first differences could appear without trend (i.e. with a constant trend). This could be erroneously confused with a time series having a linear trend, since only the first differences transformation is used. In this context, a procedure that distinguishes the linear or nonlinear nature of the trend is necessary and it represents the main aim of this contribution. In more detail, we classify the time series on the basis of the trend composition to allow an effective time series decomposition and analysis.

Formally, suppose to observe p time series of the form

$$Y_{it} = m_i(t/T) + \varepsilon_{it}, \quad i = 1, \dots, p;\ t = 1, \dots, T \tag{1}$$

where p may go to infinity as function of T, $m_i : [0,1] \to \mathbb{R}$ are unknown trend functions and $\{\varepsilon_{it}\}_{t=1}^{T}$ are zero mean, strongly mixing error processes [8]. Classical examples of strongly mixing processes are some *ARMA* processes (see [2]). In order to classify those time series according to their trend composition (constant, linear or nonlinear), the first derivative of the trend could be estimated by using a nonparametric estimator which gives at least three main advantages: (i) on the mathematical point of view, the use of the first derivative is quite intuitive to highlight the linearity of a function; (ii) it allows to assert if a trend is linear or not without imposing a predefined model a priori; (iii) knowing the classification of the trend allows a more accurate estimate of the trend itself.

The rest of the paper is organized as follows: in Section 2, the proposed method for classifying the time series is presented; in Section 3, an application of the method is presented considering the total annual growth of the GDP; finally, Section 4 collects some concluding remarks.

2 The proposed method

Considering model (1) and assuming that $m(\cdot)$ has fifth finite derivative, the proposed nonparametric estimator for the first derivative of the trend, at point $x \in [0,1]$, has the form

$$\hat{\beta}(x) = \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)Y_t, \tag{2}$$

where $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ with $K(\cdot)$ a symmetric Lipshitz continuous kernel function with bounded support, $h = h_T > 0$ is the bandwidth such that $Th^3 \to \infty$ and $h \to 0$ as $T \to \infty$. This estimator, based on the technique of the Local Polynomial estimation with fixed design (see [3] and [5], among others), is asymptotically normal distributed and its expected value is asymptotically proportional to the true first derivative, by a known quantity, as $T \to \infty$.

Assuming that $h = O\left(T^{-1/7}\right)$, the proposed procedure consists of two stages. In the first one, the estimator $\hat{\beta}(x)$ is tested to be zero by the following statistic

$$\hat{I}_\beta = \frac{\sqrt{T^{4/7}}\sup_{x\in S}|\hat{\beta}(x)|}{\sqrt{\mu_2^*\hat{c}(\varepsilon)}}, \tag{3}$$

where $\mu_2^* = \int_{-1}^{1} u^2 K(u)^2 du$, $S \subseteq (h, 1-h)$ and $|S| = O(T)$. In particular, $\hat{c}(\varepsilon)$ is a consistent estimator of $c(\varepsilon) = \gamma_\varepsilon(0) + 2\sum_{k=1}^{\infty}\gamma_\varepsilon(k)$, with $\gamma_\varepsilon(k) = Cov(\varepsilon_t, \varepsilon_{t-k})$. This statistic $\hat{I}_\beta$ allows to distinguish the time series with constant trends since, under the null hypothesis that the time series has a constant trend function, it can be proven that $\hat{I}_\beta$ converges to the absolute value of a Standard Normal distribution.

In the second stage, the statistic

$$\hat{I}_D = \frac{\sqrt{T^{4/7}}\sup_{x_1,x_2\in S}|\hat{D}(x_1,x_2)|}{\sqrt{2\mu_2^*\hat{c}(\varepsilon)}}, \tag{4}$$

with $\hat{D}(x_1,x_2) = \hat{\beta}(x_1) - \hat{\beta}(x_2)$, allows to test if the remaining time series (those that in the first stage have not been classified with constant trend) have a linear trend. In fact, under the null hypothesis that the time series has a linear trend function, it can be proven that $\hat{I}_D$ converges to the absolute value of a Standard Normal distribution. In both stages, the test has the structure of a multiple test, therefore we use the Bonferroni correction in order to calibrate the global size of the test. The procedure proposed in this work represents an important progress of the procedure used in [4]. In fact, while in [4] a simple screening procedure was considered to roughly order the time series based on their trends, here we propose the two statistics (3) and (4) (theoretical details are omitted) that allow to carry out a classification procedure for the trends.

3 Real data application

The time series data used in this paper was downloaded from Gapminder's website (www.gapminder.org). In particular, we choose the total annual GDP growth for all countries in the world. These values measure the changes in percentage of the value of everything produced in a country during a year with respect to the previous year. The different time series, for different countries, were made comparable by adjusting them for inflation and differences between countries in the cost of living.

The time period ranges from 1860 to 2013 for a total of 221 annual time series. Before applying the proposed procedure, we fill the missing values with the mean of the time series. After that, for each time series we have carried out the classification procedure, shortly described in Section 2, using the Epanechnikov kernel $K(u) = \frac{3}{4}\max(0, 1-u^2)$ and a FeedForward Neural Network estimator in order to obtain a plug-in estimator for the optimal bandwidth h (see [6]). Moreover, we set $\alpha = 0.05$ as the global size of the multiple test.

The results of the proposed procedure are shown in Table 1. Considering the period 1860-2013, the majority of countries (i.e. 175) have a GDP growth rate with a nonlinear trend, only 9 countries present a constant trend and 37 countries show a linear trend in their GDP growth. Some examples are given given in Figure 1, where the GDP time series are represented with the estimate of the corresponding trend function, according to the results of the proposed procedure. In fact, considering the first time series, since it has been classified as a time series with linear trend, we use the Least Squares (LS) method to estimate its linear trend function. In the same way, considering now the last two time series, since they have been classified as time series with a nonlinear trend, we use a Local Polynomial (LP) estimator to estimate their nonlinear trend functions. Furthermore, to evaluate our procedure, we have compared its performances in terms of values of Akaike Information Criterion (AIC) of the AutoRegressive (AR) representation for the stationary time series $Y_{it} - \widehat{m}_i(t/T)$. As mentioned before, to make the time series mean stationary, we estimate the trend using the LS method for the linear-trend time series, the LP estimator for the nonlinear-trend time series and a simple mean for the time series with no trend. On the other hand, to make the comparison, we use the alternative method of polynomial difference transformation of order $r = 1, 2, \ldots, k$ as a technique to remove the trend. More specifically, we estimate the variance of the differenced time series considering increasing values for r, until we obtain a value for the variance lower with respect to the previous one. Figure 2 highlights that using our procedure for the classification of the time series, we are able to obtain better results in terms of AIC values compared to the alternative method. In fact, the AIC distribution is flattened downwards and has lower variability. The results obtained with the alternative method of difference transformation are shown in Table 1. Specifically, this method required at most the use of the transformation into first order differences (i.e. 106 time series out of 221). Furthermore, for the 9 time series identified by our procedure as having a constant trend, the alternative procedure did not require transformation; for 37 series that we have classified with linear trend, the alternative procedure requires the first difference only for 15 of them (an example is the time series of Indonesia in Figure 1). Finally, only 91 of the 175 time series we classify as time series with nonlinear trend, have required transformations. Examples of time series that have not undergone transformations are the last two time series contained in Figure 1 that show a clear nonlinear behaviour.

Table 1 Classification sets obtained applying our procedure to the 221 GDP time series (first row), compared with the results by the de-trending procedure (last two rows): time series with constant trend (No Diff transformation) and time series with linear trend (First Diff transformation).

	Constant Trend	Linear Trend	Nonlinear Trend	Total
Our procedure:	9	37	175	221
No Diff transformation:	9	22	84	115
First Diff transformation:	0	15	91	106

4 Conclusions

In this paper, a new procedure for the classification of nonstationary time series is proposed. The nonstationarity is given by the presence of a deterministic trend, for which we do not assume a pre-specified functional form, in order to obtain an effective decomposition of the time series. This is achieved by using the first derivative trend estimator $\hat{\beta}(x)$, based on the Local Polynomial estimation technique for fixed design. The proposed procedure consists of two stages: in the first one, the estimator is tested to be zero, which allows to select the time series with constant trend; in the second stage, the difference between the estimator at different points is used to make the further linear/nonlinear partition of the remaining time series from the previous stage. An application to 221 total annual GDP growth time series is conducted to show the validity and applicability of the procedure. Furthermore, a comparison with the well known de-trending technique, based on the difference transformation, highlights the advantage of using the proposed procedure in terms of AIC values.

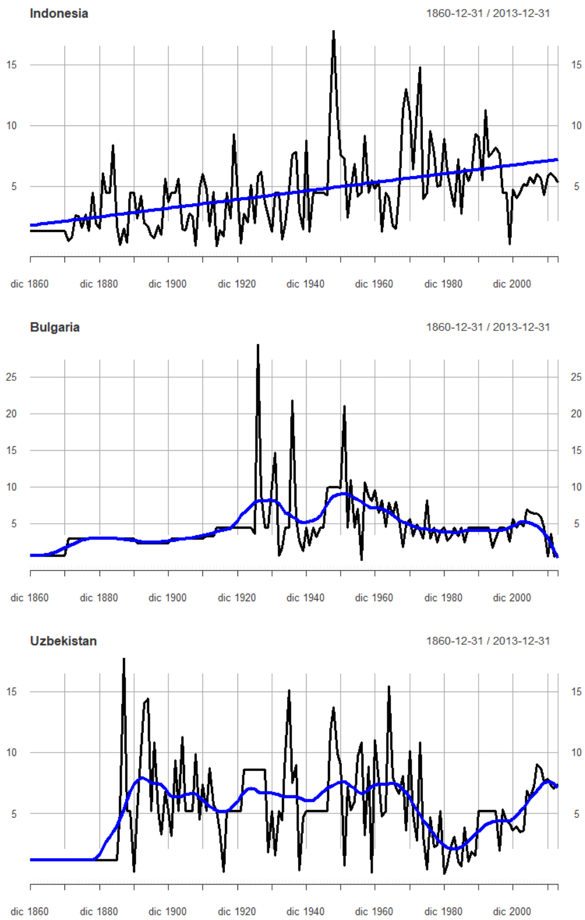
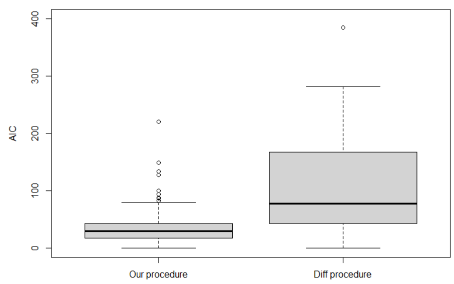


Fig. 1 Example of some GDP time series classified by the proposed procedure with the estimated trend shown in blue.

Fig. 2 Comparison of the Akaike information criterion (AIC) distributions between the AR representation of the detrended time series using the proposed procedure (Our) and using the procedure based on the difference transformation (Diff).



References

1. Chen, L. and Wu, W.B.: Testing for trends in high-dimensional time series. J. Am. Stat. Assoc., 114, 869–881 (2019)
2. Doukhan, P.: Mixing: Properties and Examples. Lecture Notes in Statistics 85. Springer-Verlag New York (1994)
3. Fan, J. and Gijbels, I.: Local polynomial modelling and its applications: monographs on statistics and applied probability. CRC Press (1996)
4. Feo, G. and Giordano, F. and Niglio, M. and Parrella, M.L.: Financial time series classification by nonparametric trend estimation. In: Mathematical and Statistical Methods for Actuarial Sciences and Finance: MAF 2022, pp. 241-246. Springer (2022)
5. Francisco-Fernández, M. and Vilar-Fernández, J.M.: Local polynomial regression estimation with correlated errors. Commun. Stat. Theory Methods, 30, 1271–1293 (2001)
6. Giordano, F. and Parrella, M.L.: Efficient nonparametric estimation and inference for the volatility function. Statistics, 53, 770–791 (2019).
7. Zhang, Ting: Clustering high-dimensional time series based on parallelism. J. Am. Stat. Assoc., 108, 577–588 (2013)
8. Zhengyan, L. and Chuanrong, L.: Limit theory for mixing dependent random variables. Springer Science & Business Media (1997)

A Normalization Method for Space-time Analysis of Evaluation and Quality Indicators

Un metodo di normalizzazione per l'analisi spazio-temporale di indicatori di valutazione e qualità

Matteo Mazziotta and Adriano Pareto

Abstract In recent years, there has been a significant increase of interest in composite evaluation and quality indicators. When a set of individual indicators with different units of measurement is to be aggregated into a composite index, the data must be normalized to make them comparable both between units and over time. The most commonly used methods are the standardization and the Min-Max method, but these have some limitations that can make reading the results difficult. This paper considers an alternative method that normalizes the range of indicators, similar to the Min-Max method, but uses a common reference that allows to 'centre' them, as in the standardization. An application to the "Government surplus/deficit" indicator of the Regional Competitiveness Index, for EU countries, is also presented.

Abstract *Negli anni recenti, c'è stato un significativo aumento di interesse negli indici sintetici di valutazione e qualità. Quando un insieme di indicatori elementari deve essere aggregato in un indice sintetico, i dati devono essere normalizzati per renderli comparabili sia tra unità che nel tempo. I metodi più comunemente usati sono la standardizzazione e il metodo Min-Max, ma tali metodi hanno alcune limitazioni che rendono difficile la lettura dei risultati. Questo articolo considera un metodo alternativo che normalizza il campo degli indicatori, similmente al metodo Min-Max, ma usa un riferimento comune che consente di 'centrarli', come nella standardizzazione. Infine, viene presentata un'applicazione all'indicatore di "Avanzo/disavanzo pubblico" dell'Indice Regionale di Competitività, per i paesi dell'Unione Europea.*

Key words: composite indicator, normalization, ranking

Matteo Mazziotta
Istat; e-mail: mazziott@istat.it

Adriano Pareto
Istat; e-mail: pareto@istat.it

1 Introduction

Composite indicators are a useful tool for performance comparisons, benchmarking and country rankings in various fields, and can also be used for the statistical evaluation of both private and public services.

A common problem in constructing composite indicators is how to normalize individual indicators with different measurement units and ranges in order to make them comparable. Unfortunately, most of the reports in the literature describe the main normalization methods [2], but do not explain how to choose the method that best fits the researcher's needs. However, normalization is a very important step, as the comparability of data between units and over time depends on the method used.

In this paper we consider a normalization method based on a change in the Min-Max method, where individual indicators are 'centred' with respect to a common reference. This solution allows the values of the units to be easily compared, both in space and time. An empirical comparison with the traditional normalization methods is also shown.

2 The constrained Min-Max method

As known, the Min-Max method (or re-scaling) converts indicators to a common scale with range [0, 1], but it does not 'centre' them with respect to a reference value and this leads to difficulties of interpretation. In particular, the normalized value 0.5 is the mean of the range, but not of the distributions, and then it cannot be used as a reference for reading results (e.g., if the normalized value of a given unit is 0.3., we cannot know if its original value is above or below the mean).

The constrained Min-Max method overcomes this limitation, as it 'normalizes' indicators, similar to the classical method, but uses a common reference that allows to 'centre' them, as the standardization does [1]. The formula is the following:

$$y_{ijt} = \frac{x_{ijt} - x_{j0}}{\max_{x_j} - \min_{x_j}}$$

where x_{ijt} is the value of indicator *j*, for unit *i*, at time *t*, $\min_{xj}$ and $\max_{xj}$ are, respectively, a minimum and a maximum that represent the possible range of indicator *j* for all units over time (*goalposts*) and x_{j0} is the reference (or base value) for indicator *j* (e.g., a benchmark). The goalposts and the reference can be calculated based on the data or set by the researcher.

This method combines the strengths of the standardization and Min-Max method, without having their weaknesses. More specifically, compared to the standardization, normalized indicators have a common reference that does not necessarily have to be the mean. Compared to the Min-Max method, normalized indicators are not forced to range between 0 and 1. Finally, normalized values are easier to interpret: if the normalized value of a given unit is greater than zero, then it is above the reference value, otherwise it is below it.

3 An application to real data

In this Section, an application of the different normalization methods to the "Government surplus/deficit" indicator of the Regional Competitiveness Index (RCI), for EU countries, in 2016 (average 2012-2014) and 2019 (average 2016-2018), is presented.

Original data and normalized data by the standardization, Min-Max method, and constrained Min-Max method are shown in Table 1.

In order to perform a space-time analysis, the parameters of the standardization are set to the mean and standard deviation of the indicator, in 2016 (-3.96 and 2.89, respectively), as suggested in Tarantola [3]. The goalposts of the Min-Max method (classic and constrained) are the minimum and the maximum of the indicator across the two times (-11.23 and 2.10, respectively); whereas the reference for the constrained Min-Max is set equal to 0, that is zero balance.

From the table we can observe that:

- standardized indicators have mean 0 and standard deviation 1 only in 2016;
- re-scaled indicators range between 0 and 1, from 2016 to 2018 (Greece has the value 0 in 2016 and Malta has the value 1 in 2018);
- normalized indicators by constrained Min-Max method have a unique reference value equal to zero (balance) for both 2016 and 2018.

Let us consider Belgium. All normalization methods show a score increase from 2016 to 2018. In particular, the standardized value changes from 0.08 to 0.92 (above the 2016 mean). The re-scaled value changes from 0.56 to 0.74, but these values are very difficult to interpret compared to goalposts. Finally, the normalized value with the constrained Min-Max method is -0.28 in 2016 and -0.10 in 2019. These values are more easily interpretable and show that Belgium had government deficit both in 2016 and in 2019, but it was reduced and almost cancelled in 2019.

Suppose now that a new year of data becomes available, where Belgium has a government surplus of +2.5. In this case, we will have a standardized value of 2.24 (above a 2016 mean of -3.96), a re-scaled value of 1.03 (out of the range), and a normalized value with the constrained Min-Max of 0.19. This latter value is more informative than the standardized value (as the reference value is zero) and does not require the recalculation of the goalposts, in contrast to the re-scaled value that is out of the range. Indeed, in the classic Min-Max method, if the goalposts are based on the existing minimum and maximum, future values may fall below 0 or above 1. To avoid this, they should be updated and the composite index recalculated for all series in order to ensure comparability between the existing and new data.

This simple application shows that to perform a clear space-time analysis, individual indicators must be normalized by using an easily understandable common reference that remains fixed over time.

Table 1 Comparing normalization methods

Country	Original value		Standardization		Min-Max method		Constrained Min-Max method	
	Average 2012/14	Average 2016/18	Average 2012/14	Average 2016/18	Average 2012/14	Average 2016/18	Average 2012/14	Average 2016/18
Austria	-1.97	-0.77	0.69	1.11	0.69	0.78	-0.15	-0.06
Belgium	-3.71	-1.30	0.08	0.92	0.56	0.74	-0.28	-0.10
Bulgaria	-0.88	1.10	1.07	1.75	0.78	0.92	-0.07	0.08
Croatia	-6.23	0.00	-0.79	1.37	0.37	0.84	-0.47	0.00
Cyprus	-5.71	-0.90	-0.61	1.06	0.41	0.77	-0.43	-0.07
Czech Rep.	-2.70	1.07	0.44	1.74	0.64	0.92	-0.20	0.08
Denmark	-2.18	0.60	0.62	1.58	0.68	0.89	-0.16	0.05
Estonia	0.22	-0.43	1.45	1.22	0.86	0.81	0.02	-0.03
Finland	-1.92	-1.07	0.71	1.00	0.70	0.76	-0.14	-0.08
France	-4.59	-2.93	-0.22	0.36	0.50	0.62	-0.34	-0.22
Germany	-0.39	1.20	1.24	1.79	0.81	0.93	-0.03	0.09
Greece	-11.23	0.77	-2.52	1.64	0.00	0.90	-0.84	0.06
Hungary	-3.45	-2.00	0.18	0.68	0.58	0.69	-0.26	-0.15
Ireland	-8.56	-0.33	-1.60	1.26	0.20	0.82	-0.64	-0.03
Italy	-3.14	-2.33	0.29	0.56	0.61	0.67	-0.24	-0.18
Latvia	-1.58	-0.50	0.82	1.20	0.72	0.80	-0.12	-0.04
Lithuania	-4.63	0.47	-0.23	1.53	0.49	0.88	-0.35	0.04
Luxembourg	0.50	1.90	1.55	2.03	0.88	0.98	0.04	0.14
Malta	-2.75	2.10	0.42	2.10	0.64	1.00	-0.21	0.16
Netherlands	-3.49	0.90	0.16	1.68	0.58	0.91	-0.26	0.07
Poland	-4.08	-1.37	-0.04	0.90	0.54	0.74	-0.31	-0.10
Portugal	-6.02	-1.83	-0.72	0.74	0.39	0.70	-0.45	-0.14
Romania	-3.61	-2.80	0.12	0.40	0.57	0.63	-0.27	-0.21
Slovakia	-3.61	-1.23	0.12	0.94	0.57	0.75	-0.27	-0.09
Slovenia	-8.46	-0.40	-1.56	1.23	0.21	0.81	-0.64	-0.03
Spain	-9.09	-3.37	-1.78	0.21	0.16	0.59	-0.68	-0.25
Sweden	-0.79	1.10	1.10	1.75	0.78	0.92	-0.06	0.08
U.K.	-6.79	-2.10	-0.98	0.64	0.33	0.68	-0.51	-0.16
Mean	**-3.96**	**-0.52**	**0.00**	**1.19**	**0.55**	**0.80**	**-0.30**	**-0.04**
Standard dev.	**2.89**	**1.47**	**1.00**	**0.51**	**0.22**	**0.11**	**0.22**	**0.11**
Min	**-11.23**	**-3.37**	**-2.52**	**0.21**	**0.00**	**0.59**	**-0.84**	**-0.25**
Max	**0.50**	**2.10**	**1.55**	**2.10**	**0.88**	**1.00**	**0.04**	**0.16**

References

1. Mazziotta, M., Pareto, A.: Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the Min–Max method. Environmetrics (2022) https://doi.org/10.1002/env.2730
2. OECD: Handbook on Constructing Composite Indicators. Methodology and user guide. OECD Publications, Paris (2008)
3. Tarantola, S.: European Innovation Scoreboard: strategies to measure country progress over time. JRC Scientific and Technical Reports, EUR 23526 EN, Luxembourg (2008)

Unveiling Latent Structures: exploring Multidimensional IRT Models using Dirichlet Process Mixtures

Svelare le Strutture Latenti: esplorazione di Modelli IRT Multidimensionali attraverso misture di Processi di Dirichlet

Pasquale Valentini, Sara Fontanella and Lara Fontanella

Abstract Statistical regularisation techniques have been recently proposed to investigate the structure of sociological measurement scales and psychometric tests. In this context, we propose a Bayesian estimation procedure able to simultaneously estimate the number of latent traits and induce sparsity in the factorial solution.
Abstract *Recentemente sono state proposte tecniche di regolarizzazione statistica per investigare la struttura delle scale di misurazione sociologiche e dei test psicometrici. In questo contesto, proponiamo una procedura di stima Bayesiana in grado di stimare simultaneamente il numero di tratti latenti e indurre sparsità nella soluzione fattoriale.*

Key words: MIRT models, Bayesian estimation, Dirichlet Processes

1 Introduction

Statistical regularisation is increasingly adopted in latent variable modelling to induce sparsity in the factorial solution in order to recover the underlying structure of psychometric scales or models [14, 12]. In a classical exploratory approach, when the number of factors is known, factor rotation is usually conducted conditional on an initial solution to achieve model simplicity. Comprehensive reviews of rotation methods, aimed at detecting simpler or sparse structures, can be found in Browne [3] Mulaik [10] and Trendafilov [15]. In practice, selecting rotation techniques can

Pasquale Valentini
Department of Economics, University of Chieti-Pescara, e-mail: pasquale.valentini@unich.it

Sara Fontanella
Department of Medicine, Imperial College London e-mail: s.fontanella@imperial.ac.uk

Lara Fontanella
Department of Legal and Social Sciences, University of Chieti-Pescara e-mail: lara.fontanella@unich.it

be challenging and subjective because these methods vary in their effectiveness in recovering accurate parameters across different population patterns. The use of regularised (or sparse) exploratory factor analysis (EFA) has been proposed as a substitute for the factor rotation stage. Rather than rotating factor loadings, regularised EFA attempts to obtain a more understandable solution, such as a simple structure, by penalising factor loadings and/or factor correlations directly during the estimation process. This is achieved by shrinking insignificant parameters towards zero. In the context of multidimensional IRT models, [6] proposed a Bayesian estimation procedure where soft constraints are imposed on the discrimination parameters by introducing a sparsity-inducing prior that favours shrinkage, enforcing the sparsity of the factorial solution.

When the number of factors is uncertain, in the classical approach to EFA an extra step becomes necessary to initially extract the factors. This adds to the complexity, as there are various methods available for factor extraction [1]. Bayesian EFA can address factor extraction and parameter estimation in one step [5]. Also regularised EFA, whether conducted through the frequentist or Bayesian approaches, offers a one-step solution [14, 11, 4, 7].

Focusing on Item Factor Analysis, we propose a Bayesian estimation procedure that allows us to simultaneously identify the number of latent traits and recover the underlying sparse structure. To identify the item dependence structure and group items based on fewer dimensions, a finite mixture model is used as a flexible approach.

2 Sparse Mixtures for MIRT models

Given a test consisting of K ordered categorical items and assuming the presence of M latent traits, the two-parameter normal ogive (2PNO) formulation of the multidimensional graded response model is presented by [2] as follows:

$$P(X_{i,k}=c|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_k,\boldsymbol{\gamma}_k)=\Phi\left(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i-\gamma_{k,c-1}\right)-\Phi\left(\boldsymbol{\alpha}_k'\boldsymbol{\theta}_i-\gamma_{k,c}\right), \qquad (1)$$

with $i=1,\ldots,N$, $k=1,\ldots,K$, and $c=1,\ldots,C$.

Equation 1 computes the probability, using the standard normal cumulative distribution function Φ, of a person i responding to item k in category c based on their unobserved latent trait scores $\boldsymbol{\theta}_i$, item discrimination parameters $\boldsymbol{\alpha}_k=(\alpha_{k,1}\ldots,\alpha_{k,M})'$ and ordered category thresholds $\boldsymbol{\gamma}_k=(\gamma_{k,1}\ldots\gamma_{k,C-1})'$. In the IRT literature, the latent traits are referred to as person parameters, while the discrimination parameters and the thresholds are known as item parameters. The probability of responding with a certain category c depends on the M-dimensional vector $\boldsymbol{\theta}_i=(\theta_{i,1},\ldots,\theta_{i,M})'\sim\mathcal{N}(\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_\theta)$. The discrimination parameters are contained in a $(K\times M)$ matrix $\boldsymbol{A}$, which represents the factorial structure of the model.

Within a Bayesian framework, we exploit the data augmentation technique [13] to obtain samples from the joint posterior distribution of the parameters. Our assumption is that the observed ordinal measure X_k is underpinned by a continuous variable Z_k, with a linear relationship between the item and person parameters and the underlying variable. Specifically, we define $Z_{i,k} = \boldsymbol{\alpha}'_k \boldsymbol{\theta}_i + \varepsilon_{i,k}$, where $\varepsilon_{i,k} \sim \mathcal{N}(0,1)$ for all i and k. The observed items are related to the underlying variables through the threshold model given by

$$X_{i,k} = c \quad if\ \gamma_{k,c-1} \leq Z_{i,k} \leq \gamma_{k,c},\ c = 1,\ldots,C;\ \gamma_{k,0} = -\infty, \gamma_{k,C} = \infty. \tag{2}$$

The full conditional of most parameters can be expressed in closed form, allowing for a Gibbs sampler. However, Metropolis-Hastings steps are necessary for sampling the ordered threshold parameters.

To identify the item dependence structure and group items based on fewer dimensions, a finite mixture model is utilised as a flexible approach. The model assumes that the K items originate from M^* hidden classes, and within each class, items can be characterised by a common data generating mechanism that is defined in terms of a probability distribution for the item, which depends on unknown class-specific parameters $\boldsymbol{\phi}_m$. Hence, if we let $\boldsymbol{Z}_k = \left(Z_{1,k}, Z_{2,k}, \ldots, Z_{N,k}\right)'$ denote the $(N \times 1)$ vector of the underlying variable related to the observed variable $\boldsymbol{X}_k$, the density of the mixture can be expressed as follows:

$$f(\boldsymbol{Z}_k|\boldsymbol{\phi}) = \sum_{m=1}^{M^*} \pi_m f(\boldsymbol{Z}_k \mid \boldsymbol{\phi}_m), \tag{3}$$

where we utilise a truncated Dirichlet process model to specify the prior over the mixing probabilities, π_m, based on an upper bound M^* [8].

3 Preliminary simulation results

In order to evaluate the performance of the proposed procedures, we perform a simulation study. We consider a multidimensional structure, which represents a generalisation of the unidimensional model since the data matrix contains more than one latent variables, but each item loads only onto a specific factor. In other terms, there is an independent-cluster (IC) latent structure [9]. Assuming $M = 6$ latent constructs, each measured by 7 four-point Likert items, such that $K = 42$ is the total number of observed categorical variables, we simulated 100 datasets, setting the sample size 500 and considering strongly correlated latent traits with the determinant of the simulation correlation matrices equal to 0.367. Figure 1 represents the simulation discrimination parameters and the means of the posterior estimates across the 100 simulated datasets, along with the 2-standard deviation intervals around the means. We highlight how the proposed method correctly retrieves the underlying factorial structure.

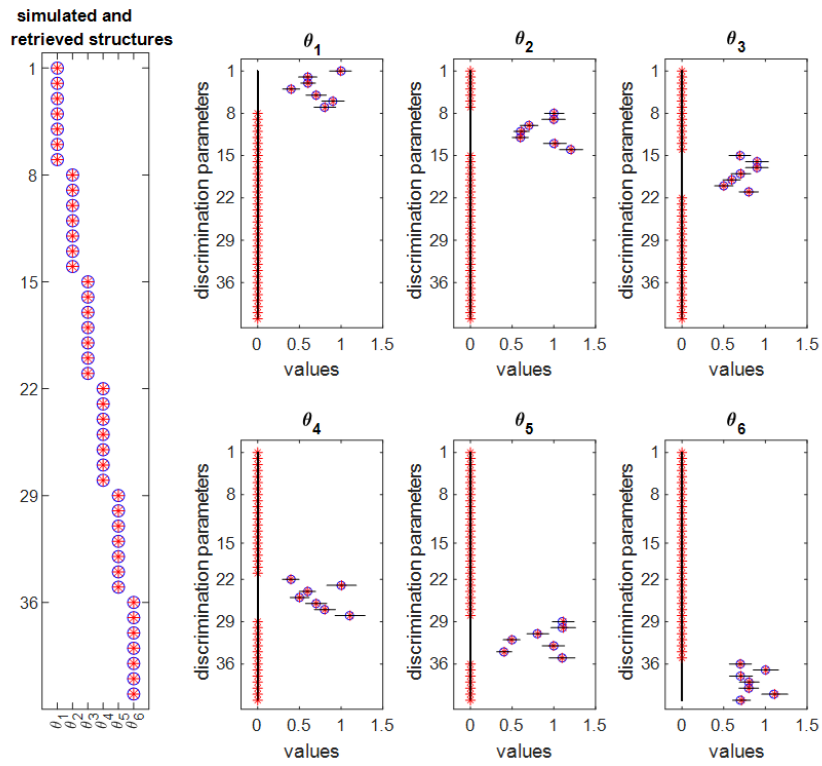


Fig. 1 Discrimination parameter estimates: simulation values ($*$); mean of the posterior estimates ($\circ$) over all the 100 simulated datasets; lines represents 2−standard deviation intervals around the means ($N = 500$).

4 Conclusions

Our proposal is a Bayesian estimation method that can simultaneously identify the number of latent traits and the sparse structure, using a Dirichlet process. A simulation study has demonstrated that the proposed model is capable of correctly retrieving the underlying factorial structure. Future improvements might include extending the model to take into consideration a bifactor model. This topic will be covered in greater detail in an expanded edition of the paper, along with a more thorough simulation analysis and a real-world application.

References

1. Auerswald, M., Moshagen, M.: How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. Psychological Methods 24, 468–491 (2019). DOI 10.1037/met0000200
2. Béguin, A.A., Glas, C.A.W.: MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66, 541–562 (2001). DOI 10.1007/BF02296195
3. Browne, M.W.: An Overview of Analytic Rotation in Exploratory Factor Analysis. Multivariate Behavioral Research 36, 111–150 (2001)
4. Chen, J.: A bayesian regularized approach to exploratory factor analysis in one step. Structural Equation Modeling: A Multidisciplinary Journal 28, 518–528 (2021). DOI 10.1080/10705511.2020.1854763
5. Conti, G., Frühwirth-Schnatter, S., Heckman, J.J., Piatek, r.: Bayesian exploratory factor analysis. Journal of Econometrics 183, 31–57 (2014). DOI 10.1016/j.jeconom.2014.06.008
6. Fontanella, L., Fontanella, S., Valentini, P., Trendafilov, N.: Simple Structure Detection Through Bayesian Exploratory Multidimensional IRT Models. Multivariate Behavioral Research 54, 100–112 (2019). DOI 10.1080/00273171.2018.1496317
7. Fŕuhwirth-Schnatter, S., Hosszejni, D., Lopes, H.: Sparse Bayesian factor analysis when the number of factors is unknown (2023)
8. Ishwaran, H., James, L.F.: Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. Journal of Computational and Graphical Statistics 11, 508–532 (2002)
9. McDonald, R.P.: A Basis for Multidimensional Item Response Theory. Applied Psychological Measurement 24, 99–114 (2000). DOI 10.1177/01466210022031552
10. Mulaik, S.A.: Foundations of factor analysis 2nd ed. CRC press, Boca Raton, FL (2010)
11. Papastamoulis, P.: Clustering multivariate data using factor analytic Bayesian mixtures with an unknown number of components. Statistics and Computing 30, 485–506 (2020). DOI 10.1007/s11222-019-09891-z
12. Scharf, F., Nestler, S.: Should regularization replace simple structure rotation in exploratory factor analysis? Structural Equation Modeling: A Multidisciplinary Journal 26, 576–590 (2019). DOI 10.1080/10705511.2018.1558060
13. Tanner, M.A., Wong, W.: The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540 (1987). DOI 10.2307/2289457
14. Trendafilov, N., Fontanella, S., Adachi, K.: Sparse Exploratory Factor Analysis. Psychometrika 82, 778–794 (2017). DOI 10.1007/s11336-017-9575-8
15. Trendafilov, N.T.: From simple structure to sparse components: a review. Computational Statistics 29, 431–454 (2014). DOI 10.1007/s00180-013-0434-5

On a technique to detect accounting data manipulation

Una tecnica per rilevare la manipolazione di dati contabili

Chiara Passamonti

Abstract Benford's Law is a mathematical model, very recurrent in practice for a wide variety of datasets, used to represent the frequencies of digits. Specifically, Benfordness statistical testing is employed within investigations aimed to ascertain if balance sheet and income statement data are genuine. However, a typical, frustrating problem of Benfordness statistical tests is that they often provide *p-values* smaller than expected, even if the Benfordness null hypothesis is true. In this paper we propose a technique to alleviate the issue through a deconvolution approach.
Abstract *La legge di Benford è un modello matematico, molto ricorrente nella pratica per un'ampia varietà di collezioni di dati numerici, usato per rappresentare le frequenze delle cifre. Nello specifico, il test statistico di Benford viene impiegato nell'ambito di indagini volte ad accertare se i dati di stato patrimoniale e di conto economico sono autentici. Tuttavia, un tipico, frustrante problema dei test statistici sulla legge di Benford è che questi spesso forniscono p-values più piccoli di quelli attesi, anche se l'ipotesi nulla è vera. In questo lavoro proponiamo una tecnica per ovviare il problema attraverso l'approccio della deconvoluzione.*

Key words: deconvolution, equivalence approach, mantissa, measurement error, Rayleigh test

1 Introduction

Benford [1] noticed that, in a large collection of numbers, the frequency distribution of the first significant digit (f.s.d.) was well described by the model

$$P(f.s.d. = i) = \log_{10}\left(\frac{i+1}{i}\right)$$

Chiara Passamonti
DSFPEQ, University of Chieti-Pescara, e-mail: chiara.passamonti@unich.it

where $i = 1, \ldots, 9$, and not by a uniform one, could be expected. An intuitive explanation of this law, based on arguments from circular statistics, follows.

Assume that W is an absolutely continuous random variable. Consider the random variable $D = cW$, where c is a real number. A corollary of the Riemann Lebesgue theorem implies that the distribution of the wrapped variable $\lim_{c\to\infty} D(mod1)$ tends to be $U[0,1)$.

This result is an old one, as [4] observed that the stopping position of a needle, which is free to rotate about the center of a disc, is uniformly distributed over different points on the circumference, if the total distance D covered by the rotations is described by a random variable having a large spread. Now, the random variable $Y = \log_{10} D$ can be represented on a circle of unitary circumference where each turn represents a unitary logarithm increment, and the stopping position $\theta = Y(mod1)$ is an angle that tends to be uniformly distributed because Y has a large spread.

Thanks to the uniformity of the stopping point θ, we are able to derive the formula of the first significant digit distribution:

$$\begin{aligned}
P(f.s.d. = i) &= P\left(i \times 10^k \leq D < (i+1) \times 10^k\right) \\
&= P\left(\log_{10} i + k \leq Y < \log_{10}(i+1) + k\right) \\
&= P\left(\log_{10} i \leq \theta < \log_{10}(i+1)\right) \\
&\approx \log_{10}(i+1) - \log_{10} i \\
&= \log_{10}\left(\frac{i+1}{i}\right)
\end{aligned} \tag{1}$$

where i ranges from 1 to 9 while k is an integer number.
Consolidated practical experience suggests that first digits of data taken from a big number of balance sheets follow a Benford distribution. A typical way to manipulate an accounting figure is to increase its first digit. Now, assume that this operation is performed several times within a collection of balance sheets. Obviously, the consequence on frequency distribution of the first digit is a smaller proportion of small digits than the expected, along with an higher than expected frequency of high digits. This describes a typical example of Benford's Law violation.
The fact that deviations from Benford's Law could indicate fraudulent behavior has attracted the interest of several authors. For example, [2] examined the second digit occurrence frequency of 220 New Zealand companies balance sheets and he noticed abnormalities regarding an excessive presence of the zero digit and an occurrence of the nine digit lower than expected. This denoted that profits were rounded up. Similarly, [5] detected the same model for american firms recording annual profits while an inverse model, i.e. fewer zeroes and more nine digits, characterized net losses.

In this paper we address the problem of reducing an additive measurement error in Benford data. In fact, we know that Benford datasets have the disappointing feature of producing very small *p-values* of conformity tests even if the null Benfordness hypothesis is widely recognized to hold. This phenomenon is commonly explained by both the presence of some noise in observed data and usually huge sample sizes, that make tests very powerful. Even for more practical sample sizes, Benford data could still be affected by measurement errors as any other ordinary data set. We will see that deconvolving Benford data presents some specific issues worth to be treated.

2 Statistical problem

2.1 The measurement error model

Assume that we can observe a random sample $\mathscr{Y}_1, \ldots, \mathscr{Y}_n$ from an unknown density $f_{\mathscr{Y}}$. Also assume that $\mathscr{Y}_i = \mathscr{X}_i + \varepsilon_i$, where ε is a random measurement error independent of $\mathscr{X}$. Our aim is to estimate the density $f_{\mathscr{X}}$ of unobservable random variable $\mathscr{X}$. A classical assumption is that the distribution of the error, say f_ε, is a known density, symmetric around zero. Observe that $f_{\mathscr{Y}}$ is a convolution between $f_{\mathscr{X}}$ and f_ε, therefore a sensitive strategy is to deconvolve a density estimate of $f_{\mathscr{Y}}$ from the known density f_ε, in order to obtain an estimate of $f_{\mathscr{X}}$.

2.2 The deconvolution approach

A typical Benford dataset exhibits the mode near zero, which usually is the left boundary of the support. Such scenario tremendously complicates the standard density estimation task in the boundary region, and this, in turn, makes the deconvolution quite difficult. So, usually, it could appear beneficial to deconvolve in the mantissa or in the base-10 logarithm space because these transformations typically lead to a shape easier to estimate. More specifically, we would have these models

$$\log_{10} \mathscr{Y}_i = \log_{10} \mathscr{X}_i + \log_{10}\left(1 + \frac{\varepsilon_i}{\mathscr{X}_i}\right),$$

or

$$\log_{10} \mathscr{Y}_i \approx \log_{10} \mathscr{X}_i + \frac{\varepsilon_i}{\mathscr{X}_i}$$

Here the main problem, provided that $\mathscr{X}_i$ and ε_i are independent, is to ascertain independence of $\mathscr{X}_i$ from $\frac{\varepsilon_i}{\mathscr{X}_i}$. For this reason we propose to use the equivalence approach, introduced by [3], as detailed in Section 2.3.

Another important feature of most Benford observations is that they range along various magnitude orders. This would suggest a multiplicative error model, but unfortunately, such an error model does not change Benfordness of data for obvious mathematical reasons. The alternative is to consider an additive error model, but this has a serious issue when the magnitude orders of data are at least two. In fact, the choice of error variance becomes not obvious because we would need a different value for each magnitude. However, such a request makes un-feasible deconvolving the dataset as a whole, because we would have the problem of selecting the variance of the common error model. In our opinion, when we have different magnitude orders, we should separately deconvolve each subgroup of data using the appropriated error variance. In this case, the final estimate will amount to the mixture of the deconvoluted densities. At first glance, this approach could appear a bit artificial, but we recall that Benford data are often sets of several and disconnected datasets.

2.3 The equivalence approach

Let $\mathscr{X}_1, \ldots, \mathscr{X}_n$ be a random sample from a Benford distribution. We have seen that a sequence of real numbers is Benford if and only if the decimal logarithm of its absolute value is uniformly distributed modulo one.

Let $\mathscr{Y}_i = \mathscr{X}_i + \varepsilon_i$ and $\mathscr{W}_i = \mathscr{Y}_i + \varepsilon_i^*$ with the ε_i and ε_i^*s being two samples drawn from the error density, which is assumed to be known. We consider that the variance of the error is not constant, such that, for each observation, it is proportional to the magnitude of data.

We hypothesize that the link between the estimate based on the base-10 logarithm of the $\mathscr{X}_i$s and the estimate based on the base-10 logarithm of the corrupted data $\mathscr{Y}_i$s is the same as the link between this latter and the estimate based on the base-10 logarithm of the sample data corrupted by an additional (simulated) level of error, that is

$$\hat{f}_X(x) : \hat{f}_Y(x) = \hat{f}_Y(x) : \hat{f}_W(x),$$

where

$$X = \log_{10} \mathscr{X}, Y = \log_{10} \mathscr{Y}, W = \log_{10} \mathscr{W}.$$

Considering the symbol ":" either as a difference or a ratio, one can, respectively, define estimators like the following ones

$$\hat{f}_X(x) = 2\hat{f}_Y(x) - \hat{f}_W(x) \tag{2}$$

$$\hat{f}_X(x) = \frac{(\hat{f}_Y(x))^2}{\hat{f}_W(x)}.$$

Remark 1. We notice that, although the deconvolution happens in the log space, data are artificially corrupted in the original space. So we do not need to know how the error model is structured in the log space, where a simple proportion among estimated densities makes us able to obtain the target.

In conclusion, a variable change technique leads to the estimate of original data at $v = 10^x$, for x being an element of the support of f_X,

$$\hat{f}_{\mathscr{X}}(v) = \frac{1}{v \ln 10} \hat{f}_X(v).$$

3 Simulation experiments

3.1 On generating Benford data

We first present a simple algorithm to create Benford datasets. The point appears worth to be treated because standard methods are not directly employable due to the fact that "Benford distribution"does not indicate any kind of parametric family. While, if Benford data are regarded in the mantissa space, they come from a population without parameters, i.e. continuous uniform density in [0,1).

1. Consider K densities whose mixture is continuous Uniform in [0,1) (for example consider an equal mixture between two Beta densities and a Uniform one: $\frac{1}{3}Be(1,2)+\frac{1}{3}U(0,1)+\frac{1}{3}Be(2,1)$). Draw from the *ith* density $n_i = n \times \omega_i$ observations, where ω_i is the weight of each component of the mixture, and $\sum_{i=1}^{k} n_i = n$.
2. Add the same integer number to all of the elements of the Kth sub-sample. In order to variegate the magnitude of data we need to choose K distinct integers.
3. Let $\mathscr{X}$ be the generic obtained observation, then the corresponding Benford datum is $10^{\mathscr{X}}$.

Notice that point 2. is reminiscent of line two in formula (1).

3.1.1 Results

We consider a very basic simulative scenario, where raw data range within a single magnitude order, expressed as a power of ten, and the random noise is Gaussian (with mean 0 and standard deviation 0.5). We extract $1,000$ samples of size n from a Uniform density $U(0,1)$. For each sample we estimate $\hat{f}_X$ both by performing the equivalence approach as in formula (2) and naive estimation ignoring the presence of the error, in both cases using kernel density estimation.

A direct way to evaluate performances is to compare *p-values* of a Benfordness test, in our case the Rayleigh one, conducted on the mantissas of $\mathscr{X}_1,\ldots,\mathscr{X}_n$, $\mathscr{Y}_1,\ldots,\mathscr{Y}_n$ and a sample of size n drawn from the equivalence estimate. In Table 1 we report average *p-values* calculated using the samples of previous experiment. Clearly, the effect of measurement error is a strong decrease of *p-values*, while samples from the estimate exhibit an evident recover towards what obtained from uncorrupted samples. From a more general perspective, these simulative results say that, in presence of measurement errors, *p-values* decrease as n increase. This is in accordance with the very small *p-values* observed in the literature in presence of big samples of Benford data with noise. Clearly, our method alleviate significantly this phenomenon, but does not completely solve it.

Table 1 Average (median) *p-values* of Benford test over 1000 samples of various size.

n	Benford test on original data	Benford test on corrupted data	Benford test on samples from deconvoluted density
100	0.50 (0.49)	0.36 (0.27)	0.33 (0.24)
400	0.49 (0.50)	0.12 (0.04)	0.26 (0.14)
700	0.49 (0.47)	0.05 (0.006)	0.23 (0.09)
1000	0.48 (0.48)	0.02 (0.0009)	0.19 (0.06)

References

1. Benford, F.: The Law of Anomalous Numbers. American Philosophical Society. Appl. 78, 551–572 (1938)
2. Carslaw, C. A. P. N.: Anomalies in Income Numbers: Evidence of Goal Oriented Behavior. The Accounting Review. Appl. 63, 321-327 (1890)
3. Di Marzio, M., Fensore. S., Panzera, A., Taylor, C.: Density estimation for circular data observed with errors. Biometrics. Appl. 78, 248-260 (2021)
4. Poincaré, H.: Chance. The Monist. Appl. 22, 31-52 (1912)
5. Thomas, J.K.: Unusual Patterns in Reported Earnings. The Accounting Review. Appl. 64, 773-787 (1890)

Session of free contributes SFC7 - *Economics and environment*

1. *Determinants of Water Conservation Behaviour and Spatial Heterogeneity in their Coefficients* (Mammadli R. and Gigliarano C.)
2. *Modeling the economic burden of grass pollen allergoid immunotherapy* (Bilancia M. and Di Bona D.)
3. *Swine fewer in Liguria: who does pay for economic losses? A causal analysis* (Baggetta C., Cavalletti B. and Corsi M.)
4. *Measuring Multidimensional Poverty of the Italian Regions in the era of COVID-19* (Chelli F.M.C., Ciommi M., Gigliarano C. and Polinesi G.)
5. *Drivers of inflation: relationships changing over time* (Sokolenko O., Palumbo A., Fortuna F., Naccarato A. and Marie J.)
6. *Environmental accounting and sustainable cities: an explorative bibliometric–based literature analysis* (Kaur A., Javed A., D'Andreamatteo A. and Rapposelli A.)

Determinants of Water Conservation Behaviour and Spatial Heterogeneity in their Coefficients

Determinanti del Comportamento di Conservazione dell'Acqua e l'Eterogeneità Spaziale nei loro Coefficienti

Rashad Mammadli and Chiara Gigliarano

Abstract This study investigates the determinants of water conservation behaviour through a comprehensive empirical analysis and explores spatial heterogeneity in their coefficients at a regional unit scale, with a specific focus on Italy. The results of the ordinal logistic regression indicate that various socio-demographic, infrastructural, behavioural, and social factors, including trust in public institutions, trust in people, and charitable donations, have a significant impact on water conservation. Furthermore, the study reveals statistically significant spatial variations in the relationship between water conservation and its five factors, including education, energy saving, and trust in public institutions, using geographically weighted regression.

Abstract *Questo studio indaga i determinanti del comportamento di conservazione dell'acqua attraverso un'analisi empirica completa ed esplora l'eterogeneità spaziale dei loro coefficienti a livello di unità regionale, con un focus specifico sull'Italia. I risultati della regressione logistica ordinale indicano che vari fattori sociodemografici, infrastrutturali, comportamentali e sociali, tra cui la fiducia nelle istituzioni pubbliche, la fiducia nelle persone e le donazioni caritatevoli, hanno un impatto significativo sulla conservazione dell'acqua. Inoltre, lo studio rivela variazioni spaziali statisticamente significative nella relazione tra la conservazione dell'acqua e i suoi cinque fattori, tra cui l'educazione, il risparmio energetico e la fiducia nelle istituzioni pubbliche, utilizzando regressione ponderata geograficamente.*

Key words: water conservation behaviour, pro-environmental behaviour, spatial non-stationarity, geographically weighted regression

---

Rashad Mammadli
Department of Economics, University of Insubria, Via Monte Generoso 71, 21100 Varese (VA), Italy; e-mail: rmammadli@uninsubria.it

Chiara Gigliarano
School of Economics and Management, University Carlo Cattaneo, CorsonMatteotti 22, 21053 Castellanza (VA), Italy; e-mail: cgigliarano@liuc.it

1 Introduction

Water conservation is a critical global issue, due to population growth, climate change, and overexploitation of natural resources. Developing effective policies and programs to reduce water consumption requires a thorough understanding of the factors that influence water saving behaviour. Prior research has found that water consumption or conservation is affected by multiple parameters, such as socio-demographic characteristics including age, household size, income and education [4,5], and psychosocial factors, such as attitudes, subjective norms, perceived behavioural control and environmental concerns [1,9]. Furthermore, physical building characteristics, including housing typology and household appliances, have been shown to affect water consumption behaviour in several ways [4]. Various other factors like water prices [8], ownership structure [6], and churchgoing [1] are also significant in determining water consumption.

It is worth noting that the impact of socioeconomic and infrastructural variables on water consumption behaviour varies across geographic locations [6], and therefore, localized data should be used to assess the relevance of previous studies' results to the target region's local conditions.

This study first investigates the determinants of water conservation behaviour through a comprehensive analysis, and later explores spatial heterogeneity in the relationship between this behaviour and its factors. In addition to exploring the impact of the previously known factors, we test the association between water conservation and various domains of social capital, such as volunteering, charitable donations, trust in others and trust in public institutions, and some self-reported well-being and water-related variables to go beyond the existing empirical research and to determine new implications for policy measures.

2 Materials and Methods

The present research utilizes data from the 2021 wave of Aspects of Daily Life survey conducted by the Italian National Institute of Statistics (ISTAT) since 1993. We restrict the sample to individuals aged 16 years and older (43,541 observations) and exploit the kNN imputation for addressing the missing data. To assess water conservation behaviour, a question asking respondents to report the frequency with which they attempt to avoid wasting water, using a 4-point scale (recoded) where higher values indicate more frequent water-saving behaviour, is utilized.

Two models are constructed. First, we identify the variables representing the determinants of water conservation behaviour that have been supported by existing research and use them to build Model 1. Second, we expand our analysis by including possible explanatory factors that have not yet been tested in existing literature including trust in public institutions. All these factors are listed in Table 1.

Table 1 Variables used to predict water conservation behaviour

Classification	Variables
Socio-Demographic Factors and Well-being	age, gender, civil status, education, number of family members, income, perceived health; ***new possible explanatory variables:*** satisfaction with life, satisfaction with the environment.
Housing Typology and Household Appliances	presence of: terrace or balcony, garden, and heating in the house; type of occupation, possession of: dishwasher, washing machine, air conditioning, and car.
Pro-environmental Behaviour and Perceptions	reading labels during shopping, organic food, local food, throwing paper in the streets, saving energy, more sustainable transportation means instead of private car, disposable products, waste sorting,[1] perception of pollution in the streets, perception of air pollution.
Social Capital	frequency of going to church, attendance to the meetings of: political parties, voluntary associations, ecological associations, free activities for voluntary organizations; ***new possible explanatory variables:*** charity in the last 12 months, trust in people, trust in public institutions.[2]
Water-related	***new possible explanatory variables:*** satisfaction with water services, drinking tap water, judgment on the cost of water.

3 Results

Prior to the analysis, we assess multicollinearity among independent variables using the Variance Inflation Factor (VIF). Results show that the age variable has the highest VIF value of 2.352, below than the recommended threshold of 5, indicating moderate collinearity with other independent variables. Additionally, all other independent variables have VIF values less than 1.7 suggesting no significant multicollinearity issues. Therefore, the model is valid for further investigation.

Table 2 presents the results of the ordinal logistic regression analysis, which accounts for 55% of the variability in water conservation behaviour in the first model, where only variables representing factors found in relevant literature are used. The second model, which adds new features as described in Section 2, shows almost the same goodness of fit. Only variables with a statistically significant effect ($p<0.05$) in at least one of the models are included in the table. The coefficients for the independent variables represent the log odds ratio of a higher category of the dependent variable occurring with a one-unit increase in the independent variable.

Considering Model 1, contrary to some prior research [9], education is positively associated with water conservation, while the results on the effects of perceived health [1] and household size [4] are in line with prior findings reporting positive relationship. Regarding pro-environmental behaviour (PEB), individuals who engage more in various domains of sustainable behaviours report a higher frequency of water

[1] A composite indicator constructed as an arithmetic mean of the eight variables which are sorting habits for paper, glass, medicine, battery, metals, plastic, organic, and textile. Cronbach's alpha is 0.81.

[2] A composite indicator constructed as an as an arithmetic mean of the seven variables which represent trust in Italian Parliament, European Parliament, regional government, municipalities, political parties, justice system and law enforcement. Cronbach's alpha is 0.91.

conservation, which is similar to existing evidence [3]. On the other hand, more organic food consumers are less likely to save water which may be the result of the motivation behind consuming naturally. Consistent with the findings by Aprile and Fiorillo [1], church attendance is correlated with more frequent water-saving, but the effect is minimal. Furthermore, individuals conserve water more when the judgment on the cost of water increases, as also revealed by Romano et al. [8].

Table 2 Determinants of Water Conservation Behaviour

Variables	Model 1		Model 2	
	Odds Ratio	S.E.	Odds Ratio	S.E.
Education	0.081***	0.016	0.081***	0.016
Income	-0.015	0.024	-0.054*	0.025
Household Size	0.030*	0.013	0.018	0.013
Health (perceived)	0.062***	0.013	0.041**	0.014
Heating	0.138**	0.052	0.140**	0.052
Dishwasher	-0.084**	0.029	-0.071*	0.029
Washing Machine	0.275*	0.119	0.281*	0.120
Air Conditioner	-0.068*	0.027	-0.064*	0.027
Reading Labels	0.221***	0.015	0.219***	0.015
Organic Food	-0.075***	0.018	-0.067***	0.018
Local Food	0.178***	0.016	0.178***	0.016
Throwing paper in the streets	0.131***	0.019	0.126***	0.019
Energy Saving	2.414***	0.020	2.400***	0.020
More sustainable transports	0.051***	0.012	0.056***	0.012
Waste Sorting	0.061**	0.020	0.057**	0.020
Churchgoing	0.049***	0.011	0.042***	0.011
Cost of water: adequate	0.211*	0.091	0.192*	0.091
Cost of water: high	0.215*	0.091	0.231*	0.092
Satisfaction with Life			0.037***	0.009
Satisfaction with Environment			0.056**	0.020
Charitable Donations			-0.197***	0.044
Trust in People			-0.108***	0.031
Trust in Public Institutions			0.051***	0.007
Pseudo R^2 (Nagelkerke R^2)	*0.5545*		*0.5568*	

***p < 0.001, **p < 0.01, *p < 0.05; the table includes only the variables with statistically significant estimates.

In Model 2, with the addition of potential explanatory variables, that had not been previously tested, the marginal effect of household size reduced to statistically insignificance, while income becomes as a significant factor with a negative effect as presented in prior research [5]. Regarding the well-being variables, those who report higher satisfaction with life and the environment are also more likely to report a higher frequency of water conservation behaviour. Surprisingly, individuals who donate to charity and who trust others emerged as negatively associated with water conservation, contradicting previous findings as Brekke et al. [2] argue that people who act more prosocially make higher contributions to public goods including the environment. In contrast, individuals become more likely to save water as their trust in public institutions increases.

3.1 Regional Disparities in the GWR Coefficients

Geographically weighted regression (GWR) is exploited to explore spatial variation in the relationship between water conservation and its factors which had statistically significant coefficients in OLR models. To make GWR estimation feasible for both time and memory complexity of $\boldsymbol{O}(\boldsymbol{kn^2})$, we use 50% of the total sample.
Table 3 presents the output of the and GWR and global OLS models. F2 test suggested by Leung et al. (2000) indicates that the GWR outperforms the global OLS, while F3 test, which verifies the significance of the spatial variation for each coefficient [7], indicates that the estimates of education, energy saving, more sustainable transportation means compared to private cars, satisfaction with the environment and trust in public institutions vary significantly across the regions of Italy (Figure 1).

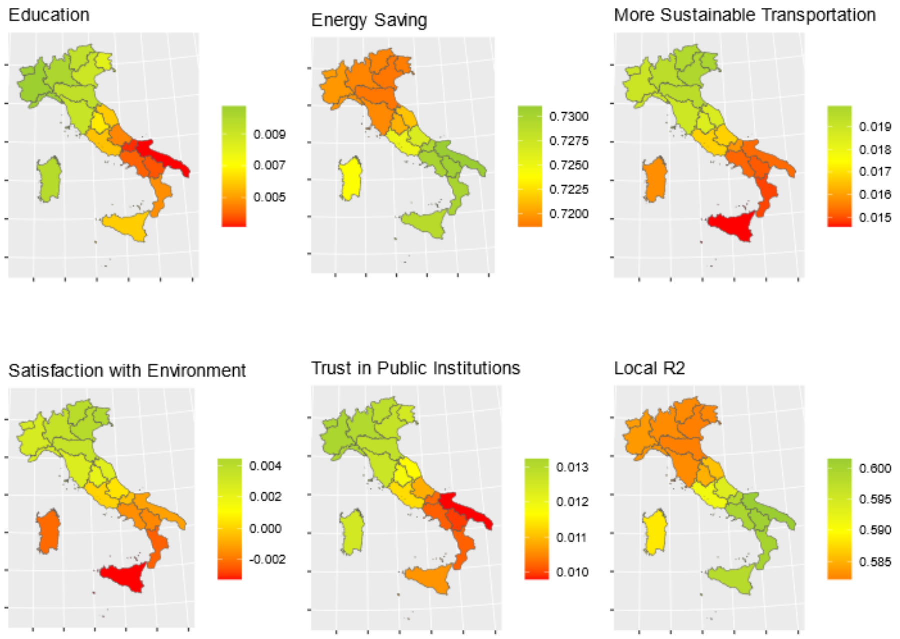
Table 3 Results of the geographically weighted regression (GWR) model

Variable	Min	Median	Max	F3 Test	Global (OLS)
Intercept	0.369	0.422	0.480	0.000	0.436
Education	0.003	0.006	0.011	0.000	0.008
Income	-0.025	-0.023	-0.022	1.000	-0.024
Health (perceived)	-0.008	-0.007	-0.006	1.000	-0.006
Heating	-0.013	-0.005	0.007	1.000	0.000
Dishwasher	-0.014	-0.013	-0.009	1.000	-0.015
Washing Machine	0.134	0.155	0.205	0.157	0.173
Air Conditioner	-0.013	-0.008	0.003	1.000	-0.001
Reading Labels	0.047	0.051	0.054	0.062	0.050
Organic Food	0.011	0.011	0.012	1.000	0.011
Local Food	0.041	0.042	0.045	1.000	0.044
Throwing Paper in Street	0.023	0.025	0.029	1.000	0.026
Energy Saving	0.719	0.721	0.731	0.000	0.725
More Sustainable Trans.	0.015	0.019	0.020	0.000	0.017
Waste Sorting	-0.019	-0.016	-0.013	1.000	-0.018
Churchgoing	0.007	0.008	0.010	1.000	0.009
Cost of Water: adequate	-0.055	-0.046	-0.037	1.000	-0.052
Cost of Water: high	-0.040	-0.022	-0.013	1.000	-0.033
Satisfaction with Life	0.002	0.003	0.005	0.961	0.002
Satisfaction with Envr.	-0.003	0.002	0.004	0.000	-0.000
Charitable Donations	-0.057	-0.051	-0.049	1.000	-0.055
Trust in People	-0.002	-0.000	0.005	1.000	0.003
Trust in Public Inst.	0.010	0.012	0.013	0.000	0.012
F2 test	1.3829	p<0.05			

4 Conclusion

The aim of the study is to explore the determinants of water conservation behaviour and investigate how the impact of these factors varies spatially at a regional unit scale in Italy. We find that in addition to the factors with existing evidence, trust in public institutions (positively), and trust in others (negatively), charitable donations (negatively), and two domains of subjective well-being (positively) contribute to water-saving behaviour. Furthermore, the study reveals significant spatial non-stationarity of the relationship between this behaviour and its five determinants.

Fig. 1 Spatial Heterogeneity in the GWR coefficients which vary significantly over the space



References

1. Aprile, M.C., Fiorillo, D.: Water conservation behavior and environmental concerns: Evidence from a representative sample of Italian individuals. J. Clean. Prod. 159, 119-129 (2017)
2. Brekke, K.A., Hauge, K.E., Lind, J. T., Nyborg, K.: Playing with the good guys. A public good game with endogenous group formation. J. Public Econ., 95(9-10), 1111-1118 (2011)
3. Dolnicar, S., Hurlimann, A., Grün, B.: Water conservation behavior in Australia. J. Environ. 105, (2012)
4. Fielding, K.S., Russell, S., Spinks, A., Mankad, A.: Determinants of household water conservation: The role of demographic, infrastructure, behavior, and psychosocial variables. Water Resour. Res. 48(10), (2012)
5. Gregory, G.D., Leo, M.D.: Repeated behavior and environmental psychology: role of personal involvement and habit formation in explaining water consumption. J. Appl. Soc. Psychol. 33, (2003)
6. Kontokosta, C.E., Jain, R.K.: Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy. Sustain. Cities Soc., 18, 44-55 (2015)
7. Leung, Y., Mei, C.L., Zhang, W.X.: Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environ. Plan A. 32(1), 9-32 (2000)
8. Romano, G., Salvati, N., Guerrini, A.: An empirical analysis of the determinants of water demand in Italy. J. Clean. Prod. 130, 74-81 (2016)
9. Russell, S.V., Knoeri, C.: Exploring the psychosocial and behavioural determinants of household water conservation and intention. Int. J. Water Resour. Dev. 36(6), 940-955 (2020)

Modeling the economic burden of grass pollen allergoid immunotherapy

Una valutazione dell'impatto economico dell'immunoterapia specifica per l'allergia al polline

Massimo Bilancia and Danilo Di Bona

Abstract Allergic rhinoconjunctivitis (ARC) is an IgE-mediated disease that occurs after exposure to indoor or outdoor allergens or nonspecific triggers. There are effective treatments for seasonal ARC, but the economic aspects of these therapies are not of secondary importance, even considering that the prevalence of ARC in Europe is estimated at 23%. To this end, we propose a novel multistate model for cost-effectiveness analysis (CEA) intended to provide healthcare professionals and policy makers with useful information for deciding on the cost-effectiveness of interventions for grass pollen-induced allergic rhinoconjunctivitis.

Abstract *La rinocongiuntivite allergica (ARC) è una malattia IgE-mediata che si verifica dopo l'esposizione ad allergeni interni o esterni o a fattori scatenanti non specifici. Esistono trattamenti efficaci per l'ARC stagionale, ma gli aspetti economici di queste terapie non sono di secondaria importanza, anche considerando che la prevalenza dell'ARC in Europa è stimata al 23%. A tal fine, proponiamo un nuovo modello multistato per l'analisi costo-efficacia (CEA), volto a fornire agli operatori sanitari e ai decisori politici uno strumento utile a valutare il rapporto costo-efficacia degli interventi terapeutici disponibili per la rinocongiuntivite allergica indotta dal polline delle graminacee.*

Key words: Cost-effectiveness analysis, non-stationary multistate Markov models, allergoid grass pollen immunotherapy.

Massimo Bilancia
University of Bari Aldo Moro, Department of Precision and Regenerative Medicine and Ionian Area (DiMePRe-J), e-mail: massimo.bilancia@uniba.it

Danilo Di Bona
University of Bari Aldo Moro, Department of Precision and Regenerative Medicine and Ionian Area (DiMePRe-J), School and Chair of Allergology and Clinical Immunology, e-mail: danilo.dibona@uniba.it

1 Introduction

Allergic rhinoconjunctivitis (ARC) is an IgE-mediated disease that occurs after exposure to indoor or outdoor allergens or nonspecific triggers such as smoke and viral infections [6]. Symptoms include rhinorrhea, nasal obstruction, itchy nose, and repeated sneezing, accompanied by eye symptoms such as itchy, red, watery, and swollen eyes. ARC severity may be classified as mild or more severe if symptoms are present but may or may not interfere with overall quality of life, exacerbate co-existing asthma, cause sleep disturbances, or interfere with daily activities and school/work performance. In addition, ARC and asthma often co-occur, and rhinitis usually precedes the onset of asthma. Allergen immunotherapy (AIT) should be considered for patients suffering from moderate or severe persistent symptoms which interfere with usual daily activities or sleep, despite compliant and appropriate drug therapies, or in patients experiencing unacceptable side-effects associated with symptomatic first-line treatments .

As for AIT, subcutaneous injection (SCIT) has been the predominant delivery method. However, in the last two decades, sublingual administration of allergens (SLIT) has increased and is now the predominant method in several European countries. The SCIT regimen involves an initial loading phase consisting of weekly administration of allergen extracts for 1-3 months, followed by monthly maintenance injections. In contrast, the SLIT regimen eliminates the loading phase and patients receive a once-daily fixed dose administered continuously throughout the year or pre-seasonally or co-seasonally, depending on the allergen causing the symptoms. Maintenance doses for both SCIT and SLIT are traditionally recommended for at least 3 years.

The economic aspects of these therapies are not of secondary importance. The prevalence of ARC in Europe is estimated at 23% [4], and the costs charged to national health services have increased exponentially. These data suggest that it would be inappropriate to ignore the weight of the economic burden resulting from the use of allergoid immunotherapy for seasonal ARC. Cost-effectiveness analysis (CEA) summarizes the problem of valuing health-related outcomes by estimating the incremental cost per unit change in outcome due to treatment [1]. CEA combines information on morbidity and quality of life to produce a value for a quality-adjusted life year (QALY). A year in perfect health is associated with 1 QALY and death with 0 QALYs, while other possible states are classified between these two extremes.

With this goal in mind, we developed a multistate, time-inhomogenous Markov model for CEA of allergoid immunotherapy that is flexible enough to account for treatment-related problems that are common in practice but cannot be adequately represented in randomized control trials [2, 3]. The therapeutic strategy in which the patient does not take any form of immunotherapy and symptomatic drug therapy is administered is used as the baseline. The transition probabilities between states can be described by a square stochastic matrix T (whose rows sum to 1). Utility values for health states are based on preferences for the various health states in the sense that the more desirable (i.e., less severe) health states are given greater

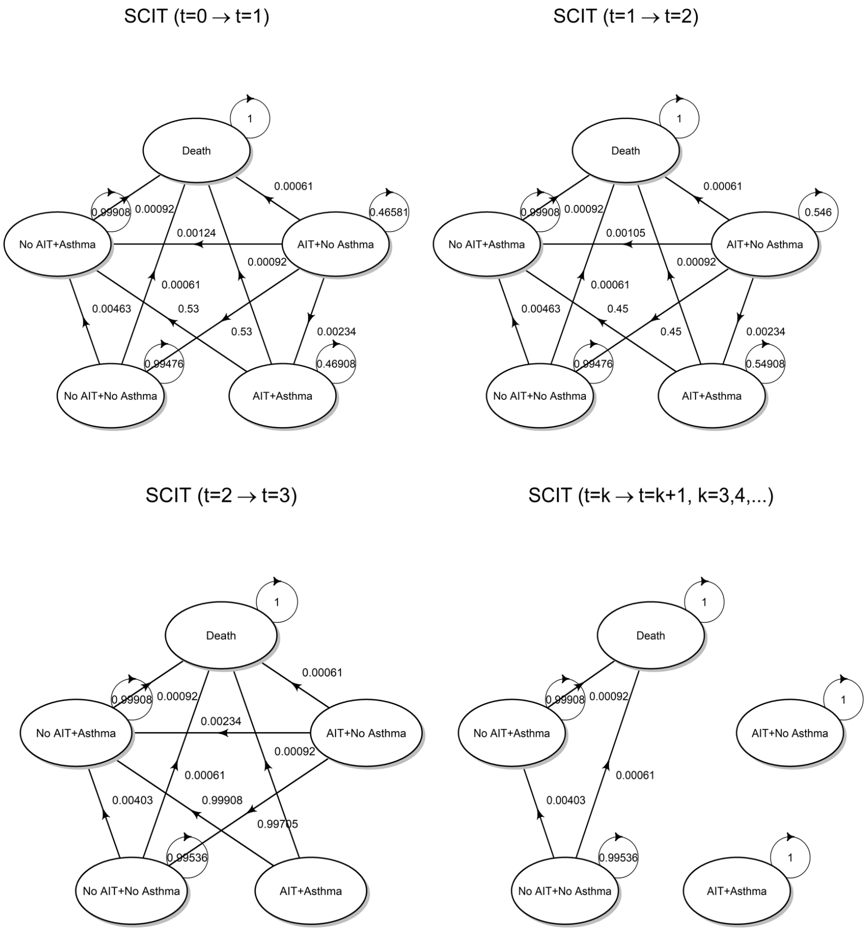


Fig. 1 State transition diagram for the SCIT strategy with superimposed transition probabilities.

weight. Utility is measured on a cardinal scale 0-1, with 0 representing death and 1 representing full health [7].

As an example, Figure 1 shows the state transition diagram for the SCIT strategy with transition probabilities superimposed, taking into account the fact that all patients complete their immunological treatment in three years. Therefore, any transition at time $t = 3$, starting from the state at $t = 2$, must must take into account the fact that the probability of being in a state where immunotherapy is administered must be set to zero, and the corresponding Markov states become isolated.

2 Discussion and conclusions

Our model is intended to provide effective information about cost-effective interventions that can help policy makers about which option provides the best value. The focus is also on simulating selected scenarios, each based on a specification with alternative assumptions to match relevant clinical situations. We have also explored other techniques for quantifying uncertainty in parameter estimates, such as probabilistic sensitivity analysis (PSA), a Monte Carlo method for simulating the sampling the joint distribution of costs and effectiveness [5]. In our model, each clinical aspect can be associated with a specific input parameter. For example, it is straightforward to introduce a prior distribution for the proportion of the pollen season in which immunotherapy is effective, thereby unifying and expanding many scenarios with meaningful clinical implications.

References

1. Baio, G.: Statistical Modeling for Health Economic Evaluations. Annual Review of Statistics and Its Application 5, 289–309 (2018). DOI 10.1146/annurev-statistics-031017-100404
2. Carta, A., Conversano, C.: On the Use of Markov Models in Pharmacoeconomics: Pros and Cons and Implications for Policy Makers. Frontiers in Public Health . DOI 10.3389/fpubh.2020.569500
3. Di Bona, D., Bilancia, M., Albanesi, M., Caiaffa, M.F., Macchia, L.: Cost-effectiveness of grass pollen allergen immunotherapy in adults. Allergy 75, 2319–2329 (2020). DOI 10.1111/all.14246
4. Linneberg, A., Petersen, K.D., Hahn-Pedersen, J., Hammerby, E., Serup-Hansen, N., Boxall, N.: Burden of allergic respiratory disease: a systematic review. Clinical and Molecular Allergy 14, 12 (2016). DOI 10.1186/s12948-016-0049-9
5. Neine, M., Curran, D.: An algorithm to generate correlated input-parameters to be used in probabilistic sensitivity analyses. Journal of Market Access & Health Policy 9 (2021). DOI 10.1080/20016689.2020.1857052
6. Roberts, G., Pfaar, O., Akdis, C.A., Ansotegui, I.J., Durham, S.R., van Wijk, R.G., Halken, S., Larenas-Linnemann, D., Pawankar, R., Pitsios, C., Sheikh, A., Worm, M., Arasi, S., Calderon, M.A., Cingi, C., Dhami, S., Fauquert, J.L., Hamelmann, E., Hellings, P., Jacobsen, L., Knol, E., Lin, S.Y., Maggina, P., Mösges, R., Elberink, J.N.G.O., Pajno, G., Pastorello, E.A., Penagos, M., Rotiroti, G., Schmidt-Weber, C.B., Timmermans, F., Tsilochristou, O., Varga, E.M., Wilkinson, J.N., Williams, A., Zhang, L., Agache, I., Angier, E., Fernandez-Rivas, M., Jutel, M., Lau, S., van Ree, R., Ryan, D., Sturm, G.J., Muraro, A.: EAACI Guidelines on Allergen Immunotherapy: Allergic rhinoconjunctivitis. Allergy 73, 765–798 (2018). DOI 10.1111/all.13317
7. Whitehead, S.J., Ali, S.: Health outcomes in economic evaluation: the QALY and utilities. British Medical Bulletin 96, 5–21 (2010). DOI 10.1093/bmb/ldq033

Swine fewer in Liguria: who does pay for economic losses? A causal analysis

Peste suina in Liguria: chi paga per le perdite economiche? Un'analisi causale

Chiara Baggetta, Barbara Cavalletti and Matteo Corsi

Abstract This paper aims to investigate the amount of economic loss of Ligurian enterprises located in the red zone following the bans imposed by the Liguria Region to stem the spread of swine fever at the beginning of 2022. The methodological approach is based on counterfactual methods that establish a causal relationship between outcome (business income) and treatment (restrictions).

Abstract *Questo paper si pone l'obiettivo di indagare l'ammontare della perdita economica delle imprese liguri situate nella zona rossa a seguito dei divieti imposti dalla Regione Liguria per arginare il diffondersi della peste suina all'inizio del 2022. L'approccio metodologico si basa su metodi controfattuali che stabiliscono una relazione causale tra l'outcome d'interesse (il reddito d'impresa) e il trattamento (le restrizioni).*

Key words: Ecosystem services, causality, swine fever, Liguria, local economy

---

Chiara Baggetta
University of Genoa, Department of Economics, Via Vivaldi 5, 16126, e-mail: chiara.baggetta@edu.unige.it

Barbara Cavalletti
University of Genoa, Department of Economics Via Vivaldi 5, 16126, e-mail: barbara.cavalletti@unige.it

Matteo Corsi
University of Genoa, Department of Economics Via Vivaldi 5, 16126, e-mail: matteo.corsi@edu.unige.it

1 Introduction

Swine fever is currently a cause for concern in Italy, particularly in Liguria. Since 2022, this disease has been causing a great deal of disruption to land management and the local economy. The spread of the virus has direct and indirect effects. On the one hand, the businesses confined to the red zone have seen a drastic drop in their income; on the other hand, the restrictions have prevented outdoor recreational activities, which are also indirect sources of income for the Ligurian territory's operators.
In line with the objectives of the European Green Deal and the New European Bauhaus, cultural ecosystem services are a tool for the preservation of biodiversity, restoration of ecosystems and the bottom-up involvement of European citizens. Millennium Ecosystem Assessment [11] defines cultural ecosystem services (ES) in terms of the "nonmaterial benefits people obtain from ecosystems," and specifically lists "cultural diversity, spiritual and religious values, knowledge systems, educational values, inspiration, aesthetic values, social elations, sense of place, cultural heritage values, recreation and ecotourism". However, some authors [7] noted that cultural ESs, characterised by being 'intangible', are difficult to quantify in biophysical or monetary terms. Although recently both European institutions (e.g., INCA platform) and the scientific community [9,10,11,13] are trying to converge in one direction regarding the accounting of cultural ecosystem services, the debate in the literature is still heated.
Starting from the specific case considered in this study, our intention is to quantify the damage caused by the imposition of the bans on businesses located within the red zone. The damage is due to the difference in income of businesses before and after the ban and includes also the loss caused by the non-use of outdoor cultural ecosystem services. Understanding how to quantify the damage is useful not only for operators claiming a form of compensation but also for regional authorities who would have an estimate of the impact of outdoor activities on the local economy.
From a methodological point of view, the situation described above is an ideal setting for the application of different types of so-called counterfactual models. Counterfactual models for the evaluation of the impact of a policy (more generally, of a treatment) allow the observed outcomes to be compared with what would have happened without the intervention, thus providing a robust and rigorous causal interpretation. Here, three models are proposed based on two different methods: spatial RDD and Difference-in-Differences. In this analysis, the aim is to exploit both the spatial dimension (the boundary between municipalities) and the temporal dimension (the pre- and post-imposition periods).

2 Background

Swine fever is a highly contagious disease that can cause serious damage to the pig industry and to local economy. Since the beginning of 2022, several cases of swine fever were reported in Italian regions of Piemonte and Liguria, causing concern among pig producers and health authorities.

Swine fever is caused by a virus that affects pigs, both domestic and wild. The disease can be transmitted through direct contact with infected animals, or through contaminated food or water. Symptoms of the disease include fever, diarrhoea, vomiting and difficulty breathing. If left untreated, swine fever can lead to the death of affected pigs.

In an attempt to prevent the spread of the disease, the Italian National and the Ligurian health authorities have taken a number of preventive measures. These restrictions include surveillance of the borders between involved areas, the abatement of affected animals, and the quarantine of suspected animals. In particular, region Liguria has banned for six months (from 13/01/2022 to 13/07/2022) all outdoor recreational activities (e.g., trekking, mountain biking, etc.) in the identified areas because human beings could spread the transmission of the virus through, for example, the soles of shoes. Below, Table 1 lists the decrees emanated by both Italian Health Minister and Health Management of region Liguria and Figure 1 shows the area under restriction.

Table 1 *Decrees by Italian authorities*

Decree	*Date*	*Promulgator*
DGSAF n. 583	11/1/2022	Italian Ministry of Health
Ordinanza del Ministro (O.M)	13/1/2022	Italian Ministry of Health in accordance with Italian Ministry of Agriculture
Ordinanza 04/22	19/1/2022	Ligurian region
Ordinanza 05/23	25/2/2022	Ligurian region

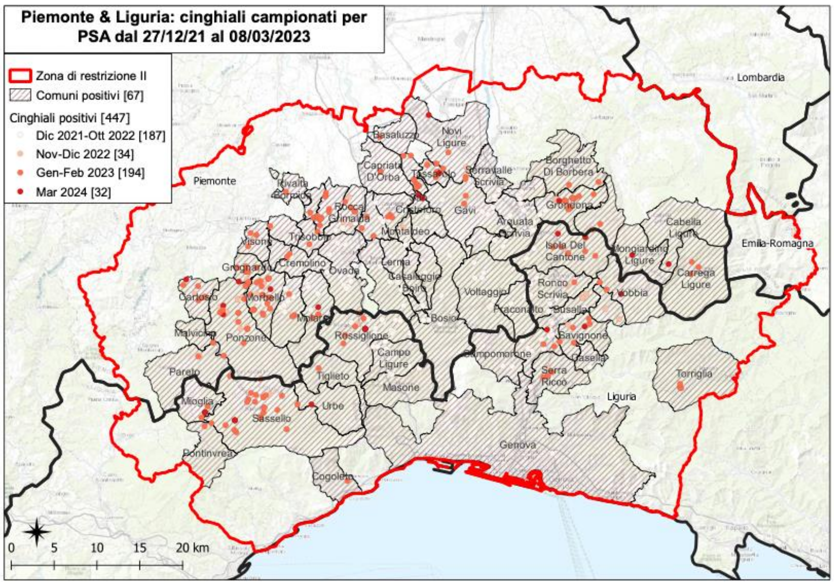


Fig. 1 *Red zone boundaries. Source: Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta*

Despite the measures taken, swine fever still continues to spread and pose a threat to both the food industry and the local economy. The spread of the virus and the resulting bans have direct and indirect effects. First of all, businesses located in the red zone have experienced a decrease in business income as access to their premises has been restricted. Secondly, the suspension of outdoor recreational activities implied a reduction of so-called cultural ecosystem services. The damage due to the closures exacerbated the problem, already known in the literature, of accounting for and of quantifying the contribution that outdoor services make to local GDP. As far as accounting for outdoor ecosystems is concerned, Highfill and Franks [8] propose the BEA project which is an attempt to measure the impact of outdoor recreation activities in the U.S.

In summary, swine fever is a serious threat to Liguria and requires rapid and effective action to prevent its spread. Health authorities must continue to monitor the region to detect any new outbreaks of the disease and take the necessary measures to prevent its spread.

3 Data

This analysis exploits several data sources. The first source is data provided by the MEF (Ministry of Finance) and concerns average income at municipal level broken down by type of taxpayer (self-employed, business, etc.), by age group and by gender. The second source of data is the Chamber of Commerce, Agriculture and Crafts of the Metropolitan City of Genoa, which provides data at the individual enterprise level, giving specific information on type, turnover and number of employees.
Finally, with regard to tourism data that allow us to investigate the extent of cultural ecosystem services activity, we exploit the Liguria Region's Tourism Satellite Account. In 2021, Liguria recorded +24.7% overnight stays compared to the previous year, just slightly below the national average (+29.5%)[1]. These data are analysed with particular attention as tourism is one of the most profitable revenues for the region.
The time span taken into consideration is the period 2019-2022. 2019 was taken as the pre-treatment year because 2020 and 2021 could have suffered from a bias due to Covid-19 pandemic.

4 Identification strategy

In this study, different techniques for inferring causality are tested and compared. Chiefly, two counterfactual methods are exploited: spatial Regression Discontinuity Design and Difference-in-Differences. Both methods can be applied as the former exploits the spatial dimension of the intervention, while the latter exploits the temporal dimension. Looking at the broad applications of the method [1,2,5,3], the DiD compares the development of a variable of interest (in our case, business income) between the treatment group (i.e., the group within the banned area) and the control group (i.e., the group that did not undergo the restrictions) before and after the intervention. The effect of the intervention is estimated by calculating the difference in the change in the variable of interest between the treatment group and the control group after the intervention, compared to the difference in the change in the same variable between the two groups before the intervention. Technically speaking:

$$Y_{ict} = \alpha + \beta(Treat_c \times Post_t) + \gamma X_{ict} + \mu_c + \tau_t + \varepsilon_t \qquad (1)$$

where Y_{ict} is the outcome of interest (firm income), β is the coefficient associated with the interaction between the presence of the ban (*Treat*) and the post-treatment dummy (*Post*). This coefficient provides the effect on the outcome that results from the activities interruption. Therefore, looking at β we can see the differential impact of the ban. X_{ict} is the set of covariates that could influence the outcome such as rate of inflation, tourist flows, broadband diffusion, type of firm and atmospheric agents. μ_c

[1] Osservatorio Turistico Regionale della Regione Liguria "Report sui Big data: il turismo in Liguria nel 2021". Dicembre, 2021

and t_t are municipalities and time fixed effects, respectively. Standard errors are clustered at the municipality level.
The second possible method is the Spatial RDD. The SRDD is based on the idea that in a continuous geographical area (in our case the boundary between municipalities), the effect of a policy or intervention may vary gradually along a boundary line, called a 'discontinuity'. This boundary line divides the geographical area into two parts, one 'treated' (municipality subject to the ban) and one 'untreated' and is used to identify the causal effects of the intervention. Following the wide literature about spatial RDD [6,4], we adopt this technique for our setting. More formally:

$$Y_{ict} = \alpha + f(distance_i) + Treat_i(\beta_1 \times f(distance_i)) + \gamma X_{it} + \varepsilon_{it} \quad (2)$$

where *distance* is the forcing variable, that is, the centred distance between the centroid of each firm and the nearest point of the policy-change boundary and *f* is a polynomial function. In order to avoid any restrictions on the underlying conditional mean functions, the polynomial function of centred distance is interacted with the treatment dummy (having the ban). X_{it} is the set of covariates that could influence the outcome such as rate of inflation, tourist flows, broadband diffusion, type of firm and atmospheric agents.

References

1. Abadie, A.: Semiparametric difference-in-differences estimators. The review of economic studies, 72(1), 1-19 (2005)
2. Angrist, J, and Pischke, J.F.: Mostly Harmless Econometrics. Princeton, NJ: Princeton University Press (2009)
3. Cerruti, G., Mazzarella, G., Migliavacca, M.: Employment protection legislation and household formation: evidence from Italy. Review of Economics of the Household, 1-27 (2022)
4. Crescenzi, R., Di Cataldo, M., & Giua, M.: It's not about the money. EU funds, local opportunities, and Euroscepticism. Regional Science and Urban Economics, 84, 103556 (2020)
5. Draca, M., Machin, S., & Van Reenen, J.: Minimum wages and firm profitability. American economic journal: applied economics, 3(1), 129-151 (2011)
6. Giua, M.: Spatial discontinuity for the impact assessment of the EU regional policy: The case of Italian objective 1 regions. Journal of Regional Science, 57(1), 109-131 (2017)
7. Daniel, T. C., Muhar, A., Arnberger, A., Aznar, O., Boyd, J. W., Chan, K. M., ... & von der Dunk, A.: Contributions of cultural services to the ecosystem services agenda. Proceedings of the National Academy of Sciences, 109(23), 8812-8819 (2012)
8. Highfill, Tina, and Connor Franks.: Measuring the US outdoor recreation economy, 2012–2016. Journal of Outdoor Recreation and Tourism 27 100233 (2019)
9. Maes, J., Egoh, B., Willemen, L., Liquete, C., Vihervaara, P., Schägner, J. P., ... & Bidoglio, G.: Mapping ecosystem services for policy support and decision making in the European Union. Ecosystem services, 1(1), 31-39 (2012)
10. Mayer M., Müller M., Woltering M., Arnegger J., Job, H.: The economic impact of tourism in six German national parks. Landsc Urban Plan 97:73–82 (2010)
11. Mendoza, G.A., Martins, H.: Multi-criteria decision analysis in natural resource management: A critical review of methods and new modelling paradigms. For Ecol Manage 230:1–22 (2006)
12. Millennium Ecosystem Assessment: Ecosystems and Human Well-Being. A Framework for Assessment (Island Press, Washington, DC) (2003)
13. TEEB: The Economics of Ecosystems and Biodiversity: Ecological and Economic Foundations, ed Kumar P (Earthscan, Oxford, UK) (2010)

Measuring Multidimensional Poverty of the Italian Regions in the era of COVID-19

Misurare la povertà multidimensionale nelle regioni italiane nell'era del COVID-19

Francesco M. Chelli, Mariateresa Ciommi, Chiara Gigliarano and Gloria Polinesi

Abstract We analyse the effect of COVID-19 on multidimensional poverty in three Italian Regions, namely Trentino Alto Adige, Marche and Sicily by measuring changes in households poverty in 2019 and 2020. Multidimensional poverty is defined in terms of deprivation in five dimensions: economic, health, education, neighbourhood quality, subjective well-being. The index at regional level is computed according to the Alkire and Foster methodology. The analysis is based on data from the 'Aspects of daily life' survey provided by Istat.
Abstract *In questo lavoro si analizza l'effetto del COVID-19 sulla povertà multidimensionale in tre regioni italiane (Trentino Alto Adige, Marche e Sicily) misurando le variazioni tra il 2019 e il 2020, a livello familiare. La povertà multidimensionale è definita in termini di deprivazione in cinque dimensioni: economica, educativa, di salute, di qualità del vicinato, benessere soggettivo. L'indice a livello regionale è calcolato secondo la metodologia Alkire e Foster. I dati provengono dall'indagine "Aspetti della vita quotidiana" condotta dall'Istat.*

Key words: Multidimensional Poverty, Italian Regions, Dominance, Indicators.

Francesco M. Chelli
Università Politecnica delle Marche, Piazzale Martelli, 8 - 60121 Ancona (Italy), e-mail: f.chelli@staff.univpm.it

Mariateresa Ciommi
Università Politecnica delle Marche, Piazzale Martelli, 8 - 60121 Ancona (Italy), e-mail: m.ciommi@staff.univpm.it

Chiara Gigliarano
LIUC - Università C. Cattaneo, C.so Matteotti, 22 - 21053 Castellanza (VA, Italy) e-mail: cgigliarano@liuc.it

Gloria Polinesi
Università Politecnica delle Marche, Piazzale Martelli, 8 - 60121 Ancona (Italy), e-mail: g.polinesi@staff.univpm.it

1 Introduction

It is well-known that poverty is a multidimensional phenomenon that cannot be adequately captured by a single indicator, such as income. Literature accounts for several indices meanly dealing with quantitative data. However, deprivation is often measured by mean of ordinal or categorical variables. Thus, here we apply a method that can be applied to all kind of variables. Moreover, using the so-called deprivation curve, we obtain robust rankings. More in details, the paper aims at measuring poverty at local level, namely Italian regions, in the era of COVID-19, by means indices and curves. On the one hand, we adopt the double cut-off approach proposed by Alkire and Foster [1]. To add robustness to the analysis, we apply the dominance criteria introduced by Lasso de la Vega [5] and we construct First order (FD) and Second order Deprivation (SD) curves. The multidimensional poverty index that we propose accounts for 15 elementary indicators grouped into five dimensions, namely health, education, economic well-being, neighbourhood quality and subjective well-being. We use data from 'Aspects of daily life' (hereafter AVQ, using the Italian acronyms), a survey conducted annually by ISTAT . Here we focus on the years 2019 and 2020 in order to capture the effect of COVID pandemic. Since AVQ does not provide information on equivalized household income, we impute it from IT-SILC data by applying a statistical matching technique based on the Coarsening Exact Matching (CEM).

The rest of paper is organized as follows. Section 2 introduces the methodology. Section 3 describes the data and illustrates the empirical results. Finally, Section 4 draws some conclusions and further research.

2 Methodology

Let $\mathbf{x}_i$ $(i = 1,\dots,n)$ be a vector representing the multidimensional distribution of a population of n units[1]. Since poverty is a multidimensional concept that encompasses several dimensions, we denote by K the number of dimensions $k = 1,\dots,K$ and, for each dimension, we denote by D_k the number of indicators in each dimensions. We also assume that indicators as well as dimensions can be weighted, and we denote by v_d the weight of each indicator in dimension d such that $\sum_{d=1}^{D_k} v_d = v_k$ and v_k represents the weight of dimension k such that $\sum_{k=1}^{K} v_k = 1$. To identify an household as multidimensional poor, we adopt the *dual cut-off approach*[2] developed by Alkire and Foster [1] combined with the First order (FD) and Second order Deprivation (SD) curves developed by Lasso de la Vega [5]. Thus, following the procedure as described in Ciommi et al. [3] and Castellano et al. [2], we compute the percentage of households deprived in at least m dimensions, the so-called *multidimensional*

[1] Here, the units of analysis are the households.

[2] The first cutoff is a threshold set for each dimension to identify whether the unit is deprived. The second cut-off is the number of dimensions in which a person must be deprived to be labelled poor.

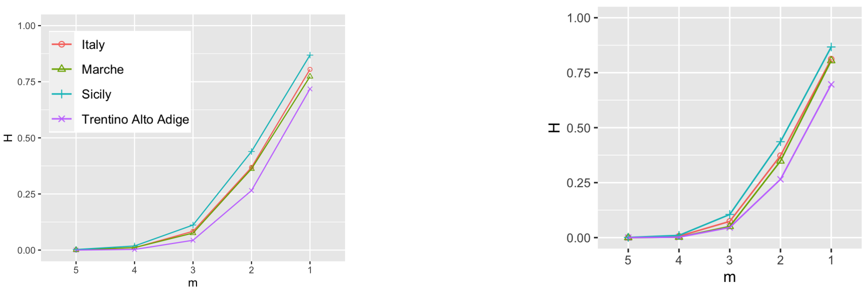
Table 1: Index framework; domains, indicators, deprivation cut-off and weight.

Domain	Indicator	Deprivation cut-off	Weight
Health	Self-reported health	Deprived if general health is bad or very bad	1/10
	Chronic illness	Deprived if suffers from chronic illness.	1/10
Education	Edu depriv.	Secondary school not completed.	1/10
	Cultural deprivation	Persons that have joined less than 2 activities: 1) at least once to cinema, theatre, exhibitions and museums, archaeological sites, monuments, concerts of classical music, opera, concerts of other kind of music; 2) read the newspaper at least once a week; 3) read at least one book in the last 12 months.	1/10
Economic well-being	Material deprivation	Person deprived if possessing less than 4 out of 6 items (washing machine, color tv, scooter/moto or car, phone, personal computer.	1/35
	Housing deprivation	Deprived if experience 3 or more deprivations related to the house (overcrowding; distance from basic services; overall poor condition of the floors and/or walls; expenses too high; house not owned).	1/35
	Gas	Deprived if their house is not served by methane gas.	1/35
	Water	Deprived if irregularities in water supply.	1/35
	Unemployment	Deprived if unemployed.	1/35
	Financial distress	Deprived if economic sources are not sufficient to make ends meet.	1/35
	Income	Poor if his/her income falls short the poverty line.	1/35
Neighb. quality	Noise	Deprived if living in a area declared to be very noisy.	1/15
	Crime	Deprived if living in a area declared at risk of crime	1/15
	Pollution	Deprived if living in a area declared polluted	1/15
Subjective well-being	Life satisfaction and future expectations	Person deprived if experience 3 or more deprivations related life, economic situation, health, familiar and friends relationship, leisure and future expectations.	1/5

headcount ratio, $H_m = \frac{q_m}{n}$ (where q_m is the number of poor household in m dimensions), and the average number of deprivations suffered by the poor households, namely the *adjusted headcount ratio* $M_m = \sum_{i=1}^{n} \frac{c_i(m)}{nK}$, measured as the ratio of the average number of deprivation among those in poverty to the maximum number of deprivation of the overall population, where $c_i(m) = 1$ if household i is deprived respect to dimension m and $c_i(m) = 0$ otherwise. M_m can be written as the product of headcount ratio and mean among the poor of the number of deprivations suffered by the poor $\left(A_m = \sum_{i=1}^{n} \frac{c_i(m)}{q_m K}\right)$ and the intensity of poverty, that is $M_m = H_m \cdot A_m$. The FD curves are obtained by plotting the number of dimension of deprivation m ranked in decreasing order, against the multidimensional headcount ratio H_m, for $m \leq d$. If the FD curve associated to country A is for every value of m above the FD curve of country B, then B has lower deprivation than A for any identification cut-off and for any multidimensional deprivation measure satisfying the following properties: Focus, Monotonicity, Symmetry and Replication invariance.[3] If FD curves intersects, we need to restrict the set of measure or reduce the range for the cut-offs. The SD curves are obtained by plotting H_m against the adjusted headcount ratio M_m . If SD curve of country A is everywhere above the SD curve of country B, than the latter exhibits a lower deprivation for any multidimensional deprivation index satisfying all the previous properties plus Distribution Sensitivity.[4]

3 Empirical Results

We define multidimensional poverty by means of 15 indicators grouped into 5 dimensions as listed in Table 1. All indicators, with the only exception of *Income*, are measured using data from 'Aspects of daily life' survey for the years 2019 and 2020.[5] For setting deprivation cut-offs, we base on metadata available for the BES indicators, explicitly revised to account for poverty. For dimensions *health*, *education* and *subjective well-being* we define an household as deprived in that dimension if it is deprived in at least one indicator. We refer to Table 1 for a detailed description of deprivation in each indicator. For the dimensions of *economic well-being* and *neighborhood quality* we define as poor an household that is deprived in 4 and 2 indicators, respectively. Moreover, even if the approach allows for different weights, here, we assume equal weights among dimensions, so that $v_k = 1/5$, for $k = 1,\dots,5$, whereas the weight of each indicator depends on the number of indicators in each dimension. Deprivation cut-offs as well as weights are reported in Table 1. AVQ



(a) 2019 (b) 2020

Fig. 1: FD curves in 2019 (panel a) and 2020 (panel b).

[3] See Lasso de la Vega [5] for more details.

[4] See Lasso de la Vega [5] for more details.

[5] Data description is available here https://www.istat.it/it/archivio/129916. The data are provided by ISTAT, upon request, for research purposes.

does not provide information on equivalized family income. For this reason we impute it from IT-SILC data by applying the Coarsening Exact Matching (CEM), a statistical matching method aimed at controlling the maximum imbalance between treated and control groups ex ante developed by Iacus et al. [4].[6] Table (2) reports the estimates of the multidimensional headcount ratio H_m with confidence intervals for all the admissible cut-offs. Figures (1a) and (1b) display the FD curves, revealing that in both year Trentino Alto Adige is the least deprived region while Sicily is the most deprived one. This ranking of deprivation is also confirmed by SD curves shown in the Figures (2a) and (2b. Furthermore, going from 2019 to 2020, regions face slight improvements in the poverty index by dimension with few exceptions (Table 3), meaning that negative aspects of COVID-19 are not captured by the elementary indicators included in each domain. For example, a worsening in the economic dimension may have been prevented by social protection policies.

Table 2: The multidimensional headcount ratio for all the admissible cut-offs (H_m). Index and 95 % bootstrap confidence interval.

		m=1			m=2			m=3			m=4			m=5		
Year		Index	95% CI		Index	95% CI		Index	95% CI		Index	95% CI		Index	95% CI	
2019	Italy	0.805	0.801	0.809	0.369	0.363	0.374	0.084	0.081	0.087	0.011	0.010	0.012	0.001	0.001	0.001
	Trentino A.A.	0.718	0.702	0.737	0.265	0.248	0.282	0.044	0.036	0.052	0.003	0.001	0.005	0.000	0.000	0.000
	Marche	0.773	0.754	0.795	0.362	0.341	0.385	0.076	0.062	0.089	0.011	0.007	0.016	0.001	0.000	0.003
	Sicily	0.869	0.857	0.880	0.439	0.422	0.457	0.112	0.102	0.123	0.018	0.013	0.024	0.003	0.001	0.005
2020	Italy	0.812	0.807	0.816	0.373	0.367	0.378	0.074	0.071	0.077	0.006	0.005	0.007	0.000	0.000	0.000
	Trentino A.A.	0.697	0.679	0.714	0.265	0.249	0.282	0.046	0.036	0.055	0.002	0.000	0.004	0.000	0.000	0.000
	Marche	0.805	0.785	0.827	0.348	0.326	0.374	0.051	0.042	0.061	0.002	0.001	0.004	0.000	0.000	0.000
	Sicily	0.867	0.855	0.882	0.436	0.417	0.457	0.105	0.094	0.117	0.010	0.006	0.014	0.001	0.000	0.002

4 Conclusion

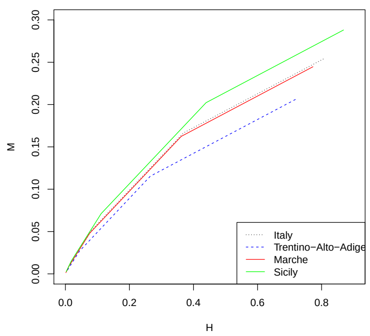
Looking at the multidimensional headcount ratio, Italy and the considered regions show a worsening only in the case of $m = 1$. Differently, analysis by dimension highlights slight improvements in the regional poverty index. Further researches will focus on the whole country to find if there are more vulnerable regions to the COVID-19 with respect to those considered and inter-regional similarities.

Acknowledgements Authors acknowledge financial support from Fondazione Cariplo (Project: *POverty and vulnerability Scenarios in The era of COVID-19: how the pandemic is affecting the well-being of the Italians*).

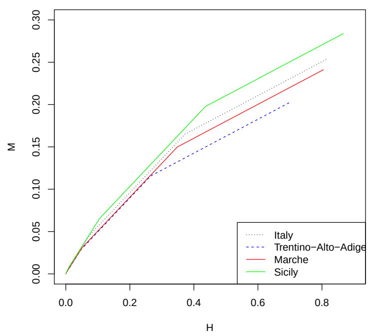
[6] For a robustness check, we also consider the Propensity Score Matching (PSM) introduced by [6] and we find that CEM estimates are closer to IT-SILC data than PSM estimates.

Table 3: Percentage of household deprived by year and by dimension

Year	Region	Health	Education	Economic	Neigh. quality	Subjective
2019	Trentino A.A.	0.3166	0.5727	0.0027	0.1099	0.0280
	Marche	0.3668	0.6512	0.0076	0.1613	0.0365
	Sicily	0.3714	0.7570	0.0897	0.1750	0,0474
2020	Trentino A.A.	0.3323	0.5537	0.0017	0.1003	0.0219
	Marche	0.3618	0.6815	0.0016	0.1307	0.0304
	Sicily	0.3583	0.7753	0.0744	0.1763	0.0353



(a) 2019



(b) 2020

Fig. 2: SD curves in 2019 (panel a) and 2020 (panel b).

References

1. Alkire, S. and Foster, J.: Counting and multidimensional poverty measurement. Journal of public economics. 95(7-8), 476–487 (2011)
2. Castellano, R., Chelli, F. M., Ciommi, M., Musella, G., Punzo, G. and Salvati, L.: Trahit sua quemque voluptas. The multidimensional satisfaction of foreign tourists visiting Italy. Socio-Economic Planning Sciences, 70, 100722 (2020)
3. Ciommi, M., de la Vega, C. L. and Chelli, F. M.: Evaluating deprivation in Italy using a multidimensional counting approach. RIEDS -The Italian Journal of Economic, Demographic and Statistical Studies, 68(1), 103-110 (2014)
4. Iacus, S.M, King, G. and Porro, G.: Matching for Causal Inference Without Balance Checking (June 26, 2008). Available at SSRN: https://ssrn.com/abstract=1152391
5. Lasso de la Vega M.C.: Counting poverty orderings and deprivation curves (ch.7), Bishop, J.A. (Ed.), Research on Economic Inequality, 18, 153-172 (2010)
6. Rosenbaum, P. R. and Rubin, D. B.: The central role of the propensity score in observational studies for causal effects. Biometrika. 70(1), 41-55 (1983)

Drivers of inflation: relationships changing over time

I driver dell'inflazione: relazioni che cambiano nel tempo

Oleksandra Sokolenko, Antonella Palumbo, Francesca Fortuna, Alessia Naccarato and Jonathan Marie

Abstract This paper adopts the cost-push and conflict interpretation of inflation and aims to analyse the relationship between inflation and other economic factors specifying how this relation changes over time. Both inflation and its possible drivers are analysed with the functional data analysis approach, where functions are the basic units of analysis. In particular, we consider a function-on-function regression model, where both the responses and predictors are curves. This approach allows us to identify the different causes of inflation and to understand how their effect varies over time, based on the historical and socio-institutional context.
Abstract *Il lavoro considera l'interpretazione dell'inflazione da costi e da conflitto e si propone di analizzare la relazione tra l'inflazione e altri fattori economici, specificando come questa cambi nel tempo. Sia l'inflazione che i suoi possibili driver sono analizzati attraverso l'approccio dell'analisi funzionale, dove le funzioni sono l'oggetto di analisi. In particolare, consideriamo un modello totalmente funzionale, in cui sia la variabile risposta che i predittori sono curve, permettendoci di identificare le diverse cause dell'inflazione e di comprendere come il loro effetto vari nel tempo, in base al contesto storico e socio-istituzionale.*

Key words: Functional data analysis, function on function regression model, conflict inflation

---

Oleksandra Sokolenko
'Roma Tre' University, Rome e-mail: oleksandra.sokolenko@uniroma3.it

Antonella Palumbo
'Roma Tre' University, Rome, e-mail: antonella.palumbo@uniroma3.it

Francesca Fortuna
'Roma Tre' University, Rome, e-mail: francesca.fortuna@uniroma3.it

Alessia Naccarato
'Roma Tre' University, Rome, e-mail: alessia.naccarato@uniroma3.it

Jonathan Marie
'Sorbonne Paris North' University, Paris, e-mail: jonathan.marie@univ-paris13.fr

1 Introduction

The cause of inflation is one of the most debated topics in Economics. This is partly due to the central role of inflation, and the fear of inflation, in policy making. Many institutions and policymakers use the threat of inflation as a motivation to not adopt expansionary fiscal policies, and when inflation does occur it is often addressed with the adoption of more restrictive fiscal measures. These responses are based on a certain understanding of the nature of inflation and the economic phenomenon which contribute to it. The mainstream interpretation explains inflation mostly as an excess demand phenomenon, accompanied by an excess supply of money and a too low level of the interest rate. In this approach, excess demand expresses itself through the Phillips curve, a model that represents the inverse relationship between the rate of inflation and the unemployment rate. However, this explanation has a long record of theoretical difficulties and empirical failures well documented in the literature [8, 9], something that has forced the economists to expand the list of possible causes and incorporate corrections to their models [5, 13]. An alternative explanation is provided by the tradition of political economy. This heterodox approach explores the sources of inflation from a different theoretical perspective, mainly focusing on the distributive conflict over income distribution and on the supply forces, as for example wage and inputs increases [17, 7]. According to this heterodox tradition, to understand inflation, it is necessary to take into account the socio-institutional context that impacts workers' bargaining power [16], consider open economy factors such as import prices and exchange rate [11], and historical factors like war, pandemic and climate shocks [3]. Another important contribution in this field is the idea that inflation can not be explained by the same stable relationship between two variables, because its causes may vary over time depending on the social and institutional context [1, 15].

This paper starts from the cost-push and conflict interpretation of inflation and aims to analyse the relationship between inflation and other economic factors specifying how this relation changes over time. To this purpose, the selected method is the functional data analysis (FDA) approach [12, 4], that focuses its attention on data seen as functions rather than a collection of data points. This perspective allows us to examine a phenomenon in its entirety and to capture its temporal variability. Specifically, we consider a function-on-function regression model, where both the responses and predictors are curves [12, 4, 6]. The peculiarity of this model is that it gives the opportunity not only to identify the main drivers of inflation, but also to understand how their effect varies over time, based on the historical and socio-institutional context.

The rest of the article is organized as follows: the methodology is presented in Sect. 2. Section 3 briefly introduces the data we use in our analysis. Section 4 presents some concluding remarks.

2 Methodology

Let I_{il} be the observed value of inflation for the i-th country, $i = 1, 2, ..., n$, at the time point t_l, $l = 1, ..., L$. We assume that these time measurements represent discrete and noisy observations of an underlying smooth function, $y_i(t)$, which belongs to a temporal domain T, with $t \in T$. We also assume that the smooth functions belong to the Hilbert space of square integrable functions, $L^2(T)$, with the usual inner product $< f, g >= \int_T f(t)g(t)\, dt$, $\forall f, g \in L^2(T)$ and the L^2-norm $||f|| =< f, f >^{1/2}< \infty$. The smooth function $y_i(t)$ is reconstructed via a linear expansion as follows [12]:

$$y_i(t) = \sum_{b=1}^{B} c_{ib}\phi_b(t), \tag{1}$$

where c_{ib} is the i-th basis coefficient, $\phi_b(t)$ represents the b-th basis function, and B is the total number of basis functions. To control the amount of smoothness, the basis coefficients are obtained by minimizing, for each i, the following penalized residual sum of squares:

$$PENSSE_i = \sum_{l=1}^{L} \Big(I_{il} - \sum_{b=1}^{B} c_{ib}\phi_b(t) \Big)^2 + \lambda \int [y''(t)]^2\, dt, \tag{2}$$

where λ is a smoothing parameter, that reflects the trade-off between smoothness and data fit, and $\int [y''(t)]^2\, dt$ is the integrated squared second derivative, representing a penalty term. It has been demonstrated that the function that minimizes the penalized residual sum of squares in Eq. (2) is a cubic spline with knots at the data points t_l [14, 2].

To study the relationship between inflation and other economic factors over time, a function on function regression model is considered:

$$y_i(t) = \beta_0(t) + \int_T \beta(s,t)x_i(s)\, ds + \varepsilon(t) \tag{3}$$

where $y_i(t)$ and $x_i(s)$ are a functional response and a functional covariate, respectively; $\beta_0(t)$ is the functional intercept, which captures the variation in the response that does not depend on the covariate function; $\beta(s,t)$ is a bivariate regression coefficient function that defines the impact of x_i at time s on y_i at time t; $\varepsilon(t)$ is the error term. The model in Eq. (3) assumes that the errors are zero-mean random processes and are uncorrelated with the functional predictors and the functional covariate. Moreover, the model considers a linear relationship between the functional response and the functional predictor, so that the effect of the predictor is expressed as the integral of the corresponding covariate weighted by a smooth bivariate coefficient function. In our case, $y_i(t)$ represents the functional inflation in Eq. (1) and $x_i(s)$ the functional unemployment, whose smooth version is reconstructed starting from the raw data in a similar manner as for the inflation;

$$x_i(s) = \sum_{k=1}^{K} a_{ik}\phi_k(s). \tag{4}$$

Without loss of generality, we consider the case where $t = s$, that is, the response and the predictor are measured on a common grid of time point. The coefficient function $\beta(s,t)$ is estimated by minimizing the following criterion:

$$\sum \Big[\int (y_i(t) - \int_T \beta(s,t) x_i(s)\, dt)^2\, dt \Big] \tag{5}$$

which results in the following basis function expansion:

$$\beta(s,t) = \sum_{k=1}^{K} \sum_{b=1}^{B} d_{kb}\phi_k(s)\phi_b(t) \tag{6}$$

3 Data description

In this paper, we study different explanations of inflation in 11 European countries, for the years 1993-2022. The inflation rate, our response variable, corresponds to the quarterly change in the core consumer price index (CPI). Among our predictors we have the rate of unemployment, which measures the number of unemployed people as a percentage of the labour force, and is used in the literature to measure the slope of the standard Phillips curve [16, 15]. For the functional representation of both inflation and unemployment, cubic B-spline basis were considered in Eq.s (1) and (4), respectively, with knots at every data point. Then, the basis coefficients were estimated by adding to the least square criterion a roughness penalty, which involves the curvature of the function as in Eq. (2) and choosing $\lambda = 0.56$ by generalized cross-validation. Figure 1 shows the functional representation of inflation (left panel) and unemployment (right panel). The function-on-function regression model in Eq. (3) is then applied to the reconstructed smooth functions and the resulting functional intercept and the bivariate regression function are plotted in Figure 2, where it is evident the strong effect of the first years on almost the whole temporal domain. On the other hand, the last few years seems to have a limited effect on inflation.

4 Conclusion and further developments

The application of the functional data analysis approach to the study of inflation and its drivers allows us to extract additional information contained in the functions that would not be available with traditional methods. Functional data analysis techniques have been increasingly applied to investigate many different economic questions. Nevertheless, few works have approached the topic of inflation, as in Zafar et al. [18] to forecast the inflation in case of Pakistan and in Meeks and Monti

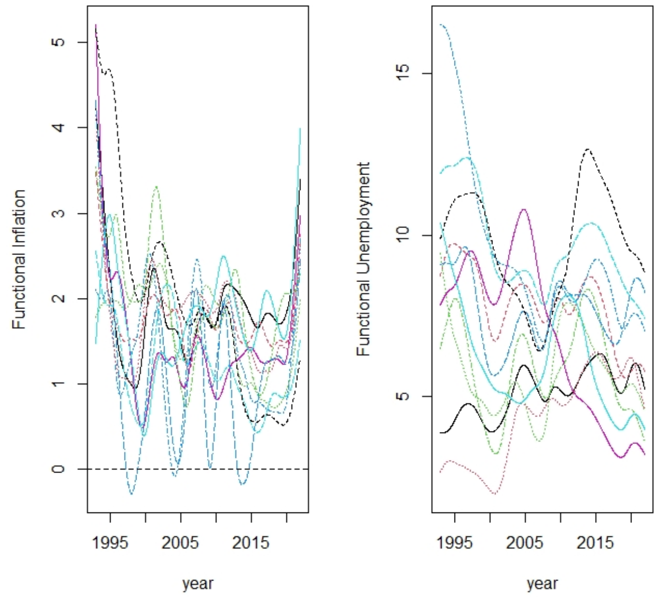


Fig. 1 Functional representation of inflation (left panel) and unemployment (right panel).

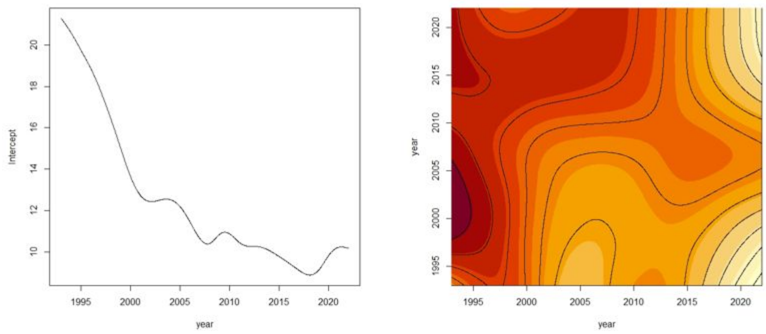


Fig. 2 Functional intercept (left panel) and bivariate regression coefficient function (right panel).

[10] to estimate the effects of expectations on inflation for the US and the UK. To the best of our knowledge, we are the first to apply a function-on-function regression analysis to this topic and to consider the time variability of the effects. Our objective is to contribute to the application of this method and to expand the work on models with multiple regressors.

References

1. Charles, S., Bastian, E.F., Marie, J.: Inflation Regimes and Hyperinflation. A Post-Keynesian/Structuralist typology. CEPN Working Papers hal-03363240, HAL (2021)
2. De Boor, C.: A Practical Guide to Splines. Springer, New York (2001)
3. Ferguson, T., Storm, S.: Myth and Reality in the Great Inflation Debate: Supply Shocks and Wealth Effects in a Multipolar World Economy. Working Papers Series inetwp196, Institute for New Economic Thinking (2023)
4. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer, New York (2006)
5. Hazell, J., Herreño, J., Nakamura, E., Steinsson, J.: The Slope of the Phillips Curve: Evidence from U.S. States. The Quarterly Journal of Economics, Oxford University Press, vol. 137(3), pages 1299-1344 (2022)
6. Ivanescu, A.E., Staicu, A., Scheipl, F., Greven, S.: Penalized function-on-function regression. Comput Stat **30**, 539–568 (2015)
7. Lavoie, M.: Post-Keynesian Economics. Edward Elgar Publishing (2014)
8. Lucas, R.E., Sargent, T.J.: After Keynesian macroeconomics. In After the Phillips Curve: Persistence of High Inflation and High Unemployment, Boston Federal Reserve Bank, Conference Series No 19, pp. 49-72 (1978)
9. Mankiw, N.G.: The Inexorable and Mysterious Tradeoff between Inflation and Unemployment. The Economic Journal, 111(471), C45–C61 (2001)
10. Meeks, R., Monti, F.: Heterogeneous Beliefs and the Phillips Curve. CAMA Working Papers 2022-51, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University (2022)
11. Perry, N., Cline, N.: Wages, Exchange Rates, and the Great Inflation Moderation: A Post-Keynesian View. Levy Economics Institute, Working Papers Series No. 759 (2013)
12. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)
13. Ratner, D., Sim, J.W.: Who Killed the Phillips Curve? A Murder Mystery. Finance and Economics Discussion Series 2022-028, Board of Governors of the Federal Reserve System (2022)
14. Reinsch, C.: Smoothing by spline functions. Numerische Mathematik **10**, 177–183 (1967)
15. Setterfield, M., Blecker, R.A.: Structural change in the US Phillips curve, 1948-2021: the role of power and institutions. Working Papers PKWP2208, Post Keynesian Economics Society (2022)
16. Summa, R., Braga, J.: Two routes back to the old Phillips curve: the amended mainstream model and the conflict augmented alternative. Bulletin of Political Economy, vol. 14(1), pages 81-115, June (2020)
17. Vernengo, M.: Money and inflation: A taxonomy. In Arestis and Sawyer (eds) A Handbook of Alternative Monetary Economics, Edward Elgar (2005)
18. Zafar, R.F., Qayyum, A., Ghouri, S.P.: Forecasting Inflation using Functional Time Series Analysis. MPRA Paper 67208, University Library of Munich, Germany (2015)

Environmental accounting and sustainable cities: an explorative bibliometric–based literature analysis

Environmental accounting e città sostenibili: un'analisi bibliometrica esplorativa della letteratura

Amandeep Kaur, Aamir Javed, Antonio D'Andreamatteo and Agnese Rapposelli

Abstract The evaluation of the impact of the use of natural resources in urban areas has gained more and more relevance in recent decades. Environmental accounting is a consolidated related framework that balances both economic and environmental dimensions. This bibliometric study explores the past and present publication analysis on urban environmental accounting research. Using the Scopus database, the study provides insights on two different datasets. The investigation highlights the primary research fields that address the sustainable assessment of cities using environmental accounting, as well as the highly cited studies in the field and the most active journals. Furthermore, utilizing most recurring keywords and keyword co-occurrence graph, the major topic explored and discussed are addressed.

Abstract *La valutazione dell'impatto dell'uso delle risorse naturali nelle aree urbane ha acquisito una rilevanza sempre maggiore negli ultimi decenni, tanto che la contabilità ambientale è un quadro di riferimento consolidato che bilancia sia la dimensione economica che quella ambientale. Il presente studio bibliometrico esplora l'analisi delle pubblicazioni passate e presenti sulla ricerca in materia di contabilità ambientale urbana. Utilizzando il database Scopus, lo studio fornisce*

Amandeep Kaur
"G. D'Annunzio" University of Chieti-Pescara; e-mail: amandeep.kaur@studenti.unich.it

Aamir Javed
"G. D'Annunzio" University of Chieti-Pescara; e-mail: aamir.javed@studenti.unich.it

Antonio D'Andreamatteo
University of Chieti-Pescara; e-mail: antonio.d'andreamatteo@ unich.it

Agnese Rapposelli
"G. D'Annunzio" University of Chieti-Pescara; e-mail: agnese.rapposelli@unich.it

approfondimenti su due diverse serie di dati. L'indagine evidenzia i principali campi di ricerca che affrontano la valutazione sostenibile delle città utilizzando la contabilità ambientale, nonché gli studi più citati nel campo e le riviste più attive. Inoltre, utilizzando le parole chiave più ricorrenti e il grafico di co-occorrenza delle parole chiave, vengono affrontati i principali argomenti esplorati e discussi.

Key words: City, Environmental accounting, Sustainable assessment, Bibliometric analysis.

1 Introduction

Due to the high concentration of people, infrastructures, housing and economic activities, cities are particularly vulnerable to climate change and natural disasters impacts on high scale. Mushrooming urban population has resulted in the introduction of urban sustainability evaluation tools that can assist practitioners in developing policy and municipal planning solutions. These can help prioritize environmental features, regions, or sectors for action, as well as proposed policy actions at various levels of government. Since 1970 United Nations has taken several steps to beat climate change, such as Kyoto Protocol and System of Environmental Accounting (SEEA) Central Framework, an international statistical standard for measuring the environment and its relationship with the economy. Environmental accounting is a framework used to quantify the state of ecosystem. There are some standards and norms introduced for environmental accounting on international and national basis. However, it is not very common to practice environmental accounting on urban scale. If applied on urban level, this computation of ecosystem can help planning practices in justifying the environmental impact [1]. Moreover, due to large and growing percentage of the global population living in cities, it has become increasingly important to calculate ecosystem assets and services in urban areas. As discussed in literature, environmental accounting has different forms [2] and urban environmental accounting is still in an emerging phase. The aim of this paper is to both correlate and explore the existing literature using bibliometric and social network analysis, to answer the following research question: how the literature field of environmental accounting in urban areas has flourished till the date?

The paper is organized as follows. Section 2 introduces the methodology used, Section 3 presents the data analysis and Section 4 concludes.

2 Research protocol

The objective of this study is to find the applicable literature related to the topic we have chosen and to investigate it with the use of bibliometric analysis. To this end, we adopt the research protocol suggested by Za and Braccini [3]. For the collection of a relevant dataset, we chose Scopus database. With the use of two relevant labels related to our study (Environmental accounting and sustainable cities), we formed our initial query by limiting the search only to English literature. The query obtained 229 papers. Together with the papers, we also downloaded the total citation for each paper. Then, we focused on high impact papers. The selection criterion was based on total number of citations: we therefore included only papers with more than 30 total citations, restricting the dataset from 229 to 35 high-impact papers. After deleting the false positives papers (which were not extensively related to our study), we reduced our high impact papers to 20.

These 20 publications show a collection of studies that are not only well-focused on the urban environmental accounting, but are also high-quality contributions with many theoretical views, methodological options, and levels of research available. These exceptional studies on urban environmental accounting frequently cross over into other important fields, pointing forth other intriguing new keywords. The resulting additional keywords are the following: urban system, energy efficiency, energy analysis, ecological footprint, corporate sustainability, carbon footprints, greenhouse gas emission, energy, environmental accounting, sustainable city, urban sustainability, sustainability. Using the chosen keywords from the 20 high impact articles, we have formed a new query, restricting our search only on the peer reviewed articles and articles written in English. The new query helped us to extract 1357 paper.

3 Analysis of results

This section presents the findings of the bibliometric analysis conducted on the selected dataset of 1357 paper. First, we performed a descriptive analysis (publication trend over time, most productive journals); secondly, the content analysis was done by highlighting the keywords defined by authors for the publication and the main cited references on both datasets. Finally, we created a co-occurrence graph for the keywords and the co–citation graph for the most cited references. The analysis was performed using R software.

3.1 Descriptive analysis

From 1991 to 1998 the concept of urban Environmental Accounting was almost not existing. The publication on this field started to increase in 1999 but it remained low until 2006, and gradually started to increase in 2007. With some fluctuations in augmenting trend from 2015 the publications reached the number of 205 in the year 2021. The number of the academic journals most active in the field of environmental accounting (number of paper higher than 13) are 20, out of which the most productive ones are Journal of Cleaner Production (150 papers), Science of Total Environment (64) and Sustainability (55). The least productive are International Journal of Environmental Research and Public Health, and Waste Management.

3.2 Topic analysis

Keyword analysis. To explore the most relevant topics discussed in the field we have identified the most recurring keywords used by authors in the papers of both datasets (Table 1).

Table 1 Most recurring keywords

Keywords	No. of papers	Keywords	No. of papers
First dataset			
Sustainable development	152	City	28
China	97	Urban development	28
Sustainability	48	Economics	27
Environmental protection	43	Ecology	26
Urban area	41	Decision making	24
Environmental impact	38	Urban growth	23
Urban planning	36	Land use	21
Article	35	Planning	21
Climate change	33	Waste management	21
Urbanization	32	Cities	20
Second dataset			
Sustainability	101	Sustainable development	30
Life Cycle Assessment	63	Environmental impact	26
China	60	Circular Economy	23
Energy	54	Energy analysis	23
Carbon footprint	51	Greenhouse gas emission	23
Climate change	51	Carbon emission	21
Environmental accounting	50	Energy accounting	21
Ecological footprint	37	Urban metabolism	21
Industrial ecology	32	CO_2 emission	20
Ecosystem service	30	Input – Output analysis	20

By observing the co-occurrence graph for the keywords of the first dataset (Figure 1), we may note that clusters containing keywords such as sustainability, sustainable development and environmental accounting are most connected. By focusing on the size of the keyword sustainability, we observe that it is the most highly co-occurred keyword in the dataset, followed by sustainable development, urban metabolism, environmental accounting and energy analysis.

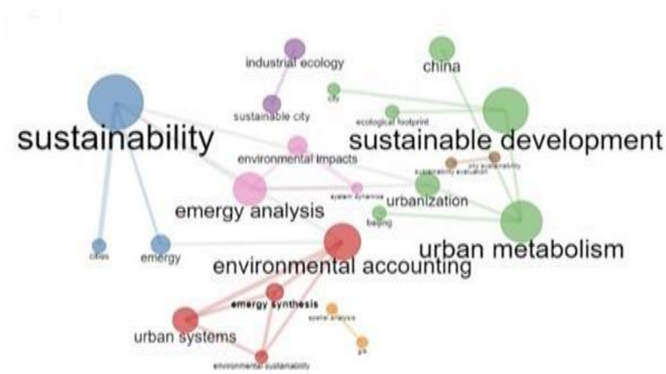


Fig. 1 Keywords co-occurrence graph (first dataset)

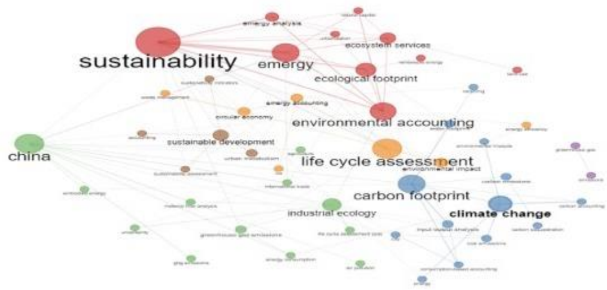


Fig. 2 Keywords co-occurrence graph (second dataset)

By focusing on Figure 2, we note that the biggest node for the second dataset is represented by the keyword sustainability, followed by life cycle assessment and China, which are also inter-connected from their individual clusters. Environmental accounting, climate change and energy register middle co-occurrence, while energy accounting, carbon emissions, and input and output analysis are the smallest sized in the cluster.

Co-citation analysis. To further investigate the highly impactful studies in the field of environmental accounting, we performed a network analysis based on the co–citations of the most cited references for the second dataset. Table 2 lists the first ten most cited references from the second dataset.

Table 2 Most cited references in second dataset

References	No. of citations
Demirbas, 2007	744
Bazilian et al., 2011	681
Burney et al, 2010	630
Marshall et al., 2009	611
Hanjra and Qureshi, 2010	604
Bixler and Porse, 2011	491
Singh et al., 2011	438
Wei et al., 2009	419
Melero et al., 2009	417
Wortley et al., 2004	377

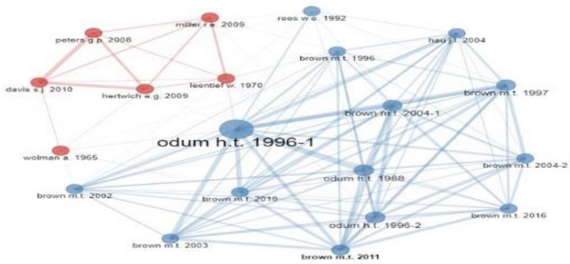


Fig. 3 Co–citation graph (second dataset)

The co–citation graph is divided in two main networks. The widest network contains the thickest node with the reference of Odum (1996) where all the other nodes meet and connect. The second cluster does not have any thickest node. The node with the reference Leontief (1970) connects the network with wide networks and other nodes within the second network (Figure 3).

4 Conclusion

Motivation behind our study was to explore the existing strand of literature on a fresh topic such as environmental accounting for cities and urban areas. For this purpose, we decided to do preliminary investigation on Scopus (Elsevier) with merely two keywords, environmental accounting and sustainable cities. As covering the ecological issue like environmental accounting refers to the sustainability of cities, this investigation returned us 229 papers, from which we chose highly impactful papers. These twenty high impact papers gave us further insights of more related keywords such as ecological footprint and energy efficiency and allow us to create a more focused query string to obtain more relevant article for urban environmental accounting. Subsequently, we extracted 1357 articles and analyzed the two-dataset using bibliometric analysis techniques. Our findings indicate that the literature started to grow since 2005 onwards. Sustainability, sustainable development, ecological footprint, energy analysis, life cycle assessment are the most investigated sub-topics. There are some limits to this study. Our bibliographic review is thorough, but it is not all inclusive: for examples it does not include grey literature or literature not written in English. Although we conducted our Scopus search using several relevant keywords, other related phrases could be utilized and may get different results. Despite these flaws, this study gives a broad overview of previous and ongoing research, resulting in a database of academic literature on urban environmental accounting.

References

1. Heris, M., Bagstad, K.J., Rhodes, C., Troy, A., Middel, A., Hopkins, K.G., Matuszak, J.: Piloting urban ecosystem accounting for the United States. *Ecosystem Services*, *48*, 101226 (2021)
2. Schaltegger, S., Burritt, R.: Contemporary environmental accounting: issues, concepts and practice. Routledge (2017)
3. Za, S., Braccini, A.M.: Tracing the Roots of the Organizational Benefits of IT Services BT. In: Za, S., Drăgoicea, M., Cavallari, C. (eds.) Exploring Services Science, pp. 3–11). Springer International Publishing (2017)

This book collects the papers presented at the 11th Scientific Meeting of the SIS Group "Statistics for the Evaluation and Quality in Services" *Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)*, which took place at the University of Chieti-Pescara, on 30th August-1st September, 2023.
The papers, which had been selected through a refereeing process, contain topics on statistical approaches and methodologies for the evaluation of public services in different contexts, and cover the areas of digital transition, e-commerce and digital marketing, enterprises, environment and territory, healthcare and wellness, finance, bank and FinTech, justice system, labour market, official statistics, public administration, food and wine, school, education and training, social, sports, sustainability, tourism, transport, university and research, well-being and welfare.

ISBN 9791280333698

ISBN: 979-12-803-3369-8
DOI: 10.60984/978-88-94593-36-5-IES2023

9 791280 333698 >